State-of-the-Art Survey

LNAI 4343

Christian Müller (Ed.)

Speaker Classification I

Fundamentals, Features, and Methods

Springer

Lecture Notes in Artificial Intelligence 4343

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Christian Müller (Ed.)

Speaker Classification I

Fundamentals, Features, and Methods

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editor

Christian Müller
International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA
E-mail: cmueller@icsi.berkeley.edu

Library of Congress Control Number: 2007932293

CR Subject Classification (1998): I.2.7, I.2.6, H.5.2, H.5, I.4-5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-540-74186-0 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-74186-2 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12107810 06/3180 5 4 3 2 1 0

Preface

> "As well as conveying a message in words and sounds, the speech signal carries information about the speaker's own anatomy, physiology, linguistic experience and mental state. These speaker characteristics are found in speech at all levels of description: from the spectral information in the sounds to the choice of words and utterances themselves."

The best way to introduce this textbook is by using the words Volker Dellwo and his colleagues had chosen to begin their chapter "How Is Individuality Expressed in Voice?" While they use this statement to motivate the introductory chapter on speech production and the phonetic description of speech, it constitutes a framework of the entire book as well: What characteristics of the speaker become manifest in his or her voice and speaking behavior? Which of them can be inferred from analyzing the acoustic realizations? What can this information be used for? Which methods are the most suitable for diversified problems in this area of research? How should the quality of the results be evaluated?

Within the scope of this book the term *speaker classification* is defined as assigning a given speech sample to a particular class of speakers. These classes could be Women vs. Men, Children vs. Adults, Natives vs. Foreigners, etc. *Speaker recognition* is considered as being a sub-field of speaker classification in which the respective class has only one member (Speaker vs. Non-Speaker). Since in the engineering community this sub-field is explored in more depth than others covered by the book, many of the articles focus on speaker recognition. Nevertheless, the findings are discussed in the context of the broader notion of speaker classification where feasible.

The book is organized in two volumes. Volume I encompasses more general and overview-like articles which contribute to answering a subset of the questions above: Besides Dellwo and coworkers' introductory chapter, the "Fundamentals" part also includes a survey by David Hill, who addresses past and present speaker classification issues and outlines a potential future progression of the field.

The subsequent part is concerned with the multitude of candidate speaker "Characteristics." Tanja Schulz describes "why it is desirable to automatically derive particular speaker characteristics from speech" and focuses on language, accent, dialect, ideolect, and sociolect. Ulrike Gut investigates "how speakers can be classified into native and non-native speakers of a language on the basis of acoustic and perceptually relevant features in their speech" and compiles a list of the most salient acoustic properties of foreign accent. Susanne Schötz provides a survey about speaker age, covering the effects of ageing on the speech production mechanism, the human ability of perceiving speaker age, as well as its automatic recognition. John Hansen and Sanjay Patil "consider a range of issues associated with analysis, modeling, and recognition of speech under stress." Anton Batliner and Richard Huber address the problem of emotion classification focusing on the specific phenomenon of irregular phonation or laryngealization and thereby point out the inherent problem of speaker-dependency, which relates the problems of speaker identification and emotion recognition with each other. The juristic implications of acquiring knowledge about the speaker on the basis of his or her speech in the context of emotion recognition is addressed by Erik Eriksson and his co-authors, discussing, "inter alia, assessment of emotion in others, witness credibility, forensic investigation, and training of law enforcement officers."

The "Applications" of speaker classification are addressed in the following part: Felix Burckhardt et al. outline scenarios from the area of telephone-based dialog systems. Michael Jessen provides an overview of practical tasks of speaker classification in forensic phonetics and acoustics covering dialect, foreign accent, sociolect, age, gender, and medical conditions. Joaquin Gonzalez-Rodriguez and Daniel Ramos point out an upcoming paradigm shift in the forensic field where the need for objective and standardized procedures is pushing forward the use of automatic speaker recognition methods. Finally, Judith Markowitz sheds some light on the role of speaker classification in the context of the deeper explored sub-fields of speaker recognition and speaker verification.

The next part is concerned with "Methods and Features" for speaker classification beginning with an introduction of the use of frame-based features by Stefan Schacht et al. Higher-level features, i.e., features that rely on either linguistic or long-range prosodic information for characterizing individual speakers are subsequently addressed by Liz Shriberg. Jacques Koreman and his co-authors introduce an approach for enhancing the between-speaker differences at the feature level by projecting the original frame-based feature space into a new feature space using multilayer perceptron networks. An overview of "the features, models, and classifiers derived from [...] the areas of speech science for speaker characterization, pattern recognition and engineering" is provided by Douglas Sturim et al., focusing on the example of modern automatic speaker recognition systems. Izhak Shafran addresses the problem of fusing multiple sources of information, examining in particular how acoustic and lexical information can be combined for affect recognition.

The final part of this volume covers contributions on the "Evaluation" of speaker classification systems. Alvin Martin reports on the last 10 years of speaker recognition evaluations organized by the National Institute for Standards and Technology (nist), discussing how this internationally recognized series of performance evaluations has developed over time as the technology itself has been improved, thereby pointing out the "key factors that have been studied for their effect on performance, including training and test durations, channel variability, and speaker variability." Finally, an evaluation measure which averages the detection performance over various application types is introduced by David van Leeuwen and Niko Brümmer, focusing on its practical applications.

Volume II compiles a number of selected self-contained papers on research projects in the field of speaker classification. The highlights include: Nobuaki Minematsu and Kyoko Sakuraba's report on applying a gender recognition system to estimate the "feminity" of a client's voice in the context of a voice therapy of a “gender identity disorder”; a paper about the effort of studying emotion recognition on the basis of a “real-life” corpus from medical emergency call centers by Laurence Devillers and Laurence Vidrascu; Charl van Heerden and Etienne Barnard’s presentation of a text-dependent speaker verification using features based on the temporal duration of context-dependent phonemes; Jerome Bellegarda’s description of his approach on speaker classification which leverages the analysis of both speaker and verbal content information – as well as studies on accent identification by Emmanuel Ferragne and François Pellegrino, by Mark Huckvale and others.

February 2007 Christian Müller

Table of Contents

I Fundamentals

II Characteristics

III Applications

IV Methods and Features

V Evaluation

How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification

Volker Dellwo, Mark Huckvale, and Michael Ashby

Department of Phonetics and Linguistics
University College London
Gower Street, London, WC1E 6BT
United Kingdom
v.dellwo@ucl.ac.uk, m.huckvale@ucl.ac.uk, m.ashby@ucl.ac.uk

Abstract. As well as conveying a message in words and sounds, the speech signal carries information about the speaker's own anatomy, physiology, linguistic experience and mental state. These speaker characteristics are found in speech at all levels of description: from the spectral information in the sounds to the choice of words and utterances themselves. This chapter presents an introduction to speech production and to the phonetic description of speech to facilitate discussion of how speech can be a carrier for speaker characteristics as well as a carrier for messages. The chapter presents an overview of the physical structures of the human vocal tract used in speech, it introduces the standard phonetic classification system for the description of spoken gestures and it presents a catalogue of the different ways in which individuality can be expressed through speech. The chapter ends with a brief description of some applications which require access to information about speaker characteristics in speech.

Keywords: Speech production, Phonetics, Taxonomy, IPA, Individuality, Speaker characteristics.

1 Introduction

Whenever someone speaks an utterance, they communicate not only a message made up of words and sentences which carry meaning, but also information about themselves as a person. Recordings of two people saying the same utterance will sound different because the process of speaking engages the neural, physiological, anatomical and physical systems of a specific individual in a particular circumstance. Since no two people are identical, differences in these systems lead to differences in their speech, even for the same message. The speaker-specific characteristics in the signal can provide information about the speaker's anatomy, physiology, linguistic experience and mental state. This information can sometimes be exploited by listeners and technological applications to describe and classify speakers, possibly allowing speakers to be categorised by age, gender, accent, language, emotion or

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 1–20, 2007.

health. In circumstances where the speaker is known to the listener, speaker characteristics may be sufficient to select or verify the speaker's identity. This leads to applications in security or forensics. The aim of this chapter is to provide a framework to facilitate discussion of these speaker characteristics: to describe ways in which the individuality of speakers can be expressed through their voices.

Always in the discussion of speaker characteristics, it must be borne in mind that a spoken utterance exists primarily for its communicative value – as an expression of a desire in the mind of the speaker to make changes in the mind of the listener. The study of the communicative value of utterances is the domain of Linguistics, which we take to include knowledge of articulation, phonology, grammar, meaning and language use. The study of speaker characteristics is in a sense parallel to this, where we concentrate on what a particular implementation of an utterance within the linguistic system tells us about the person speaking.

At first glance, it may appear that we should be able to separate speaker characteristics from message characteristics in a speech signal quite easily. There is a view that speaker characteristics are predominantly low level – related to the implementation in a particular physical system of a given set of phonetic gestures, while message characteristics operate at a more abstract level – related to the choice of phonetic gestures: the syllables, words and phrases that are used to communicate the meaning of a message. However this is to oversimplify the situation. Speakers are actually different at all levels, because speakers also differ in the way in which they realise the phonetic gestures, they vary in the inventory of gestures used, in the way in which gestures are modified by context, and in their frequency of use of gestures, words and message structure. A speaker's preferred means of morning greeting may help identify them just as much as their preferred average pitch.

To build a framework in which the many potential influences of an individual on his or her speech can be discussed, we have divided this chapter into three sections: section 2 provides an overview of vocal structures in humans, section 3 introduces the conventional principles of phonetic classification of speech sounds, while section 4 provides a discussion on how and on what levels speaker characteristics find their way into the speech signal and briefly discusses possible applications of this knowledge.

2 Vocal Apparatus

In this section we will give an overview of the physical structures in the human that are used in the physical generation of speech sounds. We will look at the anatomy of the structures, their movements and their function in speech. The first three sections look at the structures below the larynx, above the larynx and the larynx itself. The last section briefly introduces the standard signals and systems model of speech acoustics.

2.1 Sub-laryngeal Vocal Tract

Figure 1 shows the main anatomical structures that are involved in speaking. Looking below the larynx we see the lungs lying inside a sealed cavity inside the rib cage. The

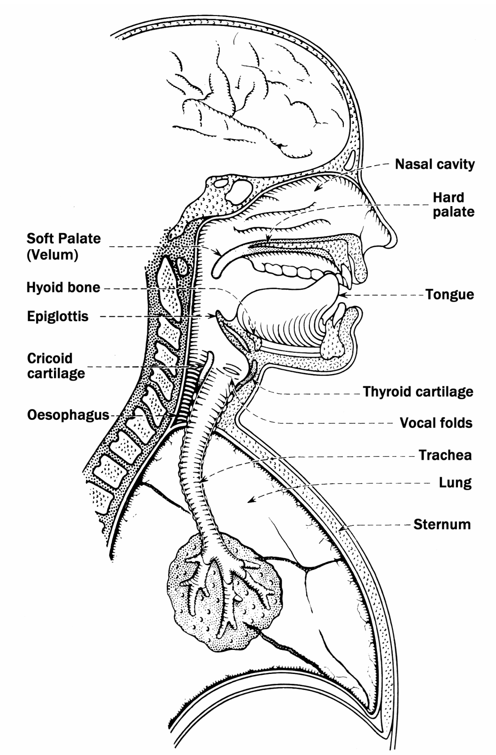


Fig. 1. Schematic diagram of the human organs of speech (Adapted from [1])

volume of the air spaces in the lungs can be varied from about 2 litres to about 6 litres in adults. The volume of the chest cavity and hence the volume of the lungs themselves is increased by lowering the diaphragm or raising the rib cage; the volume is decreased by raising the diaphragm or lowering the rib cage. The diaphragm is a dome of muscle, rising into the lower surface of the lungs, and tensing it causes it to flatten out and increase the size of the chest cavity; conversely relaxation of the diaphragm or action of the abdominal wall muscles makes the diaphragm more domed, reducing the size of the cavity. The external intercostal muscles bring the ribs closer together, but since they are pivoted on the vertebrae and are floating at the lower end of the rib cage, contraction of these muscles raises the rib cage and increases the volume of the chest cavity. The internal intercostal muscles can be used to depress the rib cage, and in combination with muscles of the abdominal wall, these can act to forcibly reduce the size of the chest cavity.

Changes in the size of the chest cavity affect the size of the lungs and hence the pressure of the air in the lung cavities. A reduction in pressure draws in air through the mouth or nose, through the pharynx, larynx and trachea into the lungs. A typical inspiratory breath for speech has a volume of about 1.5 litres, and is expended during speech at about 0.15 litres/sec [2]. One breath may be used to produce up to 30 seconds of speech. An increase in the pressure of air in the lungs forces air out through the trachea, larynx, pharynx, mouth and nose. To produce phonation in the larynx, the lung pressure has to rise by at least 300Pa to achieve sufficient flow for vocal fold vibration. A more typical value is 1000Pa, that is 1% of atmospheric pressure.

Pressure is maintained during speech by a control mechanism that connects stretch receptors in the trachea, bronchioles and lung cavities to the muscles that control chest cavity volume. The stretch receptors provide information about the physical extension of the lung tissues which indirectly measures lung pressure. At large volumes the natural elasticity of the lungs would cause too high a pressure for speaking, so nerve activation on the diaphragm and external intercostal muscles is required to maintain a lower pressure, while at low volumes the elasticity is insufficient to maintain the pressure required for speaking, so nerve activation on the internal intercostal muscles and abdominal wall muscles is required to maintain a higher pressure.

2.2 The Larynx

The larynx is the major sound generation structure in speech. It sits in the air pathway between lungs and mouth, and divides the trachea from the pharynx. It is suspended from the hyoid bone which in turn is connected by muscles to the jaw, skull and sternum. This arrangement allows the larynx to change in vertical position. The larynx is structured around a number of cartilages: the cricoid cartilage is a ring that sits at the top of the trachea at the base of the larynx; the thyroid cartilage is a V shape with the rear legs articulating against the back of the cricoid cartilage and the pointed front sticking out at the front of the larynx and forming the "Adam's apple" in the neck; the two arytenoid cartilages sit on the cricoid cartilage at the back of the larynx.

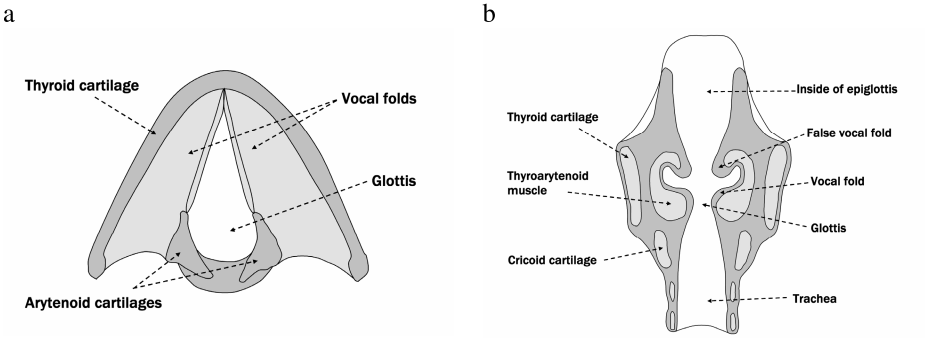


Fig. 2. Schematic diagrams of the larynx: (a) superior view, showing vocal folds, (b) vertical section, showing air passage

The vocal folds are paired muscular structures that run horizontally across the larynx, attached close together on the thyroid cartilage at the front, but connected at the rear to the moveable arytenoid cartilages, and forming an adjustable valve. For breathing the folds are held apart (abducted) at their rear ends and form a triangular opening known as the glottis. Alternatively, the arytenoids can be brought together (adducted), pressing the folds into contact along their length. This closes the glottis and prevents the flow of air. If the folds are gently adducted, air under pressure from the lungs can cause the folds to vibrate as it escapes between them in a regular series of pulses, producing the regular tone called "voice". Abduction movements of the vocal folds are controlled by contraction of the posterior cricoarytenoid muscles, which cause the arytenoids to tilt and hence draw the rear of the vocal folds apart. Adduction movements are controlled by the transverse interarytenoid muscles and the oblique interarytenoid muscles which draw the arytenoids together, also the lateral cricoarytenoid muscles which cause the arytenoids themselves to swivel in such a way as to draw the rear of the folds together.

The open glottis position gives voiceless sounds, such as those symbolised [s] or [f]; closure produces a glottal stop, symbolised [ʔ], while voice is used for all ordinary vowels, and for many consonants. Commonly, consonants are in voiced-voiceless pairs; for example, [z] is the voiced counterpart of [s], and [v] the voiced counterpart of [f].

As well as adduction/abduction, the vocal folds can change in length and tension owing to movements of the thyroid and arytenoid cartilages and of changes to the muscles inside the vocal folds. These changes primarily affect the rate of vocal fold vibration when air is forced through a closed glottis. The cricothyroid muscles rock the thyroid cartilage down and hence stretch and lengthen the vocal folds. Swivelling of the arytenoid cartilages with the posterior and lateral interarytenoid muscles also moves the rear of the folds relative to the thyroid, and changes their length. Within the vocal folds themselves, the thyroarytenoid muscle can contract in opposition to the other muscles, and so increase the tension in the folds independently from their length.

Generally, changes in length, tension and degree of adduction of the vocal folds in combination with changes in sub-glottal pressure cause changes in the loudness, pitch and quality of the sound generated by phonation. Normal (modal) voice produces a clear, regular tone and is the default in all languages. In breathy voice (also called murmur), vibration is accompanied by audible breath noise. Other glottal adjustments include narrowing without vibration, which produces whisper, and strong adduction but low tension which produces an irregular, creaky phonation.

2.3 Supra-laryngeal Vocal Tract

Immediately above the larynx is the pharynx, which is bounded at the front by the epiglottis and the root of the tongue. Above the pharynx, the vocal tract branches into the oral and nasal cavities, see Fig. 1. The entrance to the nasal cavity is controlled by the soft palate (or velum) which can either be raised, to form a closure against the rear wall of the pharynx, or lowered, allowing flow into the nasal cavity and thus out of the nostrils. The raising of the soft palate is controlled by two sets of muscles: the tensor veli palatini and the levator veli palatini which enter the soft palate from above. Lowering of the soft palate is controlled by another two sets of muscles: the palatopharyngeus muscle and the palatoglossus muscle which connect the palate to the pharynx and to the back of the tongue respectively.

Air flowing into the oral cavity can eventually leave via the lip orifice, though its path can be controlled or stopped by suitable manoeuvres of the tongue and lips. The main articulators which change the shape and configuration of the supra-laryngeal vocal tract are the soft palate, the tongue, lips and jaw.

The upper surface of the oral cavity is formed by the hard palate, which is domed transversely and longitudinally, and is bordered by a ridge holding the teeth. In a mid-sagittal view, the portion of this behind the upper incisors is seen in section, and generally referred to as the alveolar ridge. The lower surface of the oral cavity consists of the tongue, a large muscular organ which fills most of the mouth volume when at rest. Various parts of the tongue can be made to approach or touch the upper surface of the mouth, and complete airtight closures are possible at a range of locations, the closure being made not only on the mid-line where it is usually visualised, but extending across the width of the cavity and back along the tongue rims. The position and shape of the tongue are controlled by two sets of muscles: the extrinsic muscle group lie outside the tongue itself and are involved in the protrusion of the tongue, the depression of the tip of the tongue, the forward-backward movement of the tongue and the raising and lowering of the lateral borders of the tongue. The intrinsic muscles lie within the body of the tongue and are involved in flattening and widening the tongue, lengthening and narrowing the tongue, and also raising and lowering the tongue tip. Together the many sets of muscles can move the bulk of the tongue within the oral cavity and change the shape of the remaining cavity, which in turn affects its acoustic properties.

The available space in the oral cavity and the distance between the upper and lower teeth can be altered by adjusting the jaw opening. Raising the jaw is performed mainly by the masseter muscle which connects the jaw to the skull, while lowering the jaw is performed by muscles that connect the jaw to the hyoid bone.

At the exit of the oral cavity, the lips have many adjustments that can affect the shape of the oral opening and even perform a complete closure. Lip movements fall into two broad categories: retrusive/protrusive movements largely performed by the orbicularis oris muscles that circle the lips, and lateral/vertical movements performed by a range of muscles in the cheeks that attach into the lips, called the muscles of facial expression.

2.4 Sound Generation

To a very good approximation, we can describe the generation of speech sounds in the vocal tract as consisting of two separate and independent processes. In the first process, a constriction of some kind in the larynx or oral cavity causes vibration and/or turbulence which gives rise to rapid pressure variations which propagate rapidly through the air as sound. In the second process, sound passing through the air cavities of the pharynx, nasal and oral cavities is modified in terms of its relative frequency content depending on the shape and size of those cavities. Thus the sound radiated from the lips and nostrils has properties arising from both the sound source and the subsequent filtering by the vocal tract tube. This approach is called the source-filter model of speech production.

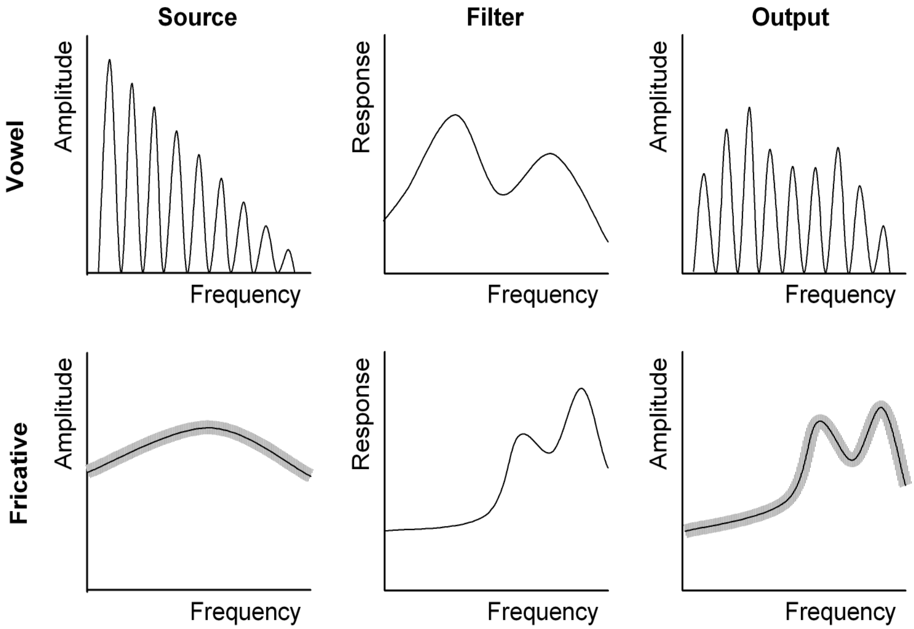


Fig. 3. Frequency domain diagram of the source-filter explanation of the acoustics of a voiced vowel (upper) and a voiceless fricative (lower). Left: the source spectrum, middle: the vocal tract transfer function, right: the output spectrum.

Phonation is periodic vibration in the larynx which starts when sub-glottal pressure rises sufficiently to push adducted folds apart. The resulting flow through the glottis causes a fall in pressure between the folds due to the Bernoulli effect, which in turn draws the folds together and ultimately causes them to snap shut, cutting off the flow and creating a momentary pressure drop immediately above the glottis. The cycle then repeats in a quasi-periodic manner at frequencies between about 50 and 500Hz depending on larynx size and larynx settings. The spectrum of this sound is rich in harmonics, extending up to about 5000Hz, and falling off at about -12dB/octave. See Fig. 3.

Apart from phonation, other sound sources are created by air-flow from the lungs becoming turbulent at constrictions in the larynx and oral cavity or at obstacles to the air-flow. Noise sources caused by the turbulence have broad continuous spectra which vary in envelope depending on the exact place and shape of constriction. Typically, noise sources have a single broad frequency peak varying from about 2 to 6kHz, rolling off at lower and high frequencies.

The frequency response of an unobstructed vocal tract closed at the glottis and with a raised soft-palate can be well described by a series of poles (resonances) called the formants of the tract, see Fig. 3. The formant frequencies and bandwidths are commonly used to parameterise the vocal tract frequency response. However, when the soft-palate is lowered, when there are constrictions to the air-flow through the tract, or when the glottis is open, additional zeros (anti-resonances) are present.

When sound is radiated from the lips and nostrils, it undergoes another frequency shaping which effectively differentiates the signal, providing a gain of +6dB/octave to the speech signal.

3 Phonetic Classification

Phonetic classification is the system of categories and descriptive labels which underlies the Phonetic Alphabet of the International Phonetic Association [3]. It regards speech as a succession of sounds (segments), and characterises the production of each such segment by specifying a relatively static target configuration. This section introduces the standard principles used by phoneticians to categorise the phonetic gestures used in speech.

3.1 Place and Manner of Articulation

Vowels are sounds produced with a relatively open vocal tract through which air flows with little resistance, while consonants involve some degree of obstruction to the airflow. Place of articulation refers to the location along the vocal tract where a consonantal obstruction is formed.

The terminology for place of articulation is summarised in Fig. 4, around a mid-sagittal schematic of the vocal tract. Words shown without a leading or trailing hyphen are complete place terms. So alveolar refers to a type of articulation in which the tip and blade of the tongue approach the ridge behind the upper teeth, velar to one made by the back of the tongue against the velum, and so on. More precise terminology consists of hyphenated terms on the left, which refer to 'active' articulators, paired with terms from the shaded box (which refer to 'passive' articulators).

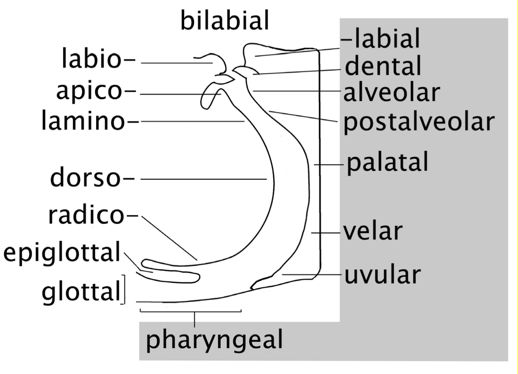


Fig. 4. Schematic of vocal tract showing terminology used to indicate place of articulation (after [4])

Manner of articulation refers to the type of obstruction used in the production of a consonant – whether, for example, the airflow is blocked completely for a brief time (yielding the manner known as plosive) or simply obstructed so that noisy turbulent flow occurs (the manner known as fricative).

Table 1. Manners of articulation (after [4])

manner	definition	comments
nasal	complete oral closure, soft palate lowered to allow air to escape nasally	
plosive	complete closure, soft palate closed	
affricate	plosive released into fricative at the same place of articulation	not always treated as a separate manner
fricative	close approximation of articulators, turbulent airflow	sibilants, having turbulence at the teeth, are an important sub-category
lateral fricative	complete closure on mid-line, turbulent flow at the side	
lateral approximant	complete closure on the mid-line, open approximation at the side	
approximant	open approximation, flow not turbulent	approximants which are within the vowel space are also called semivowels
trill	flexible articulator vibrates in the air stream	in trills and taps the brief closures do not raise intra-oral air pressure significantly flaps start with the tongue retroflexed
tap/flap	a single brief closure made by the tongue hitting the alveolar ridge	

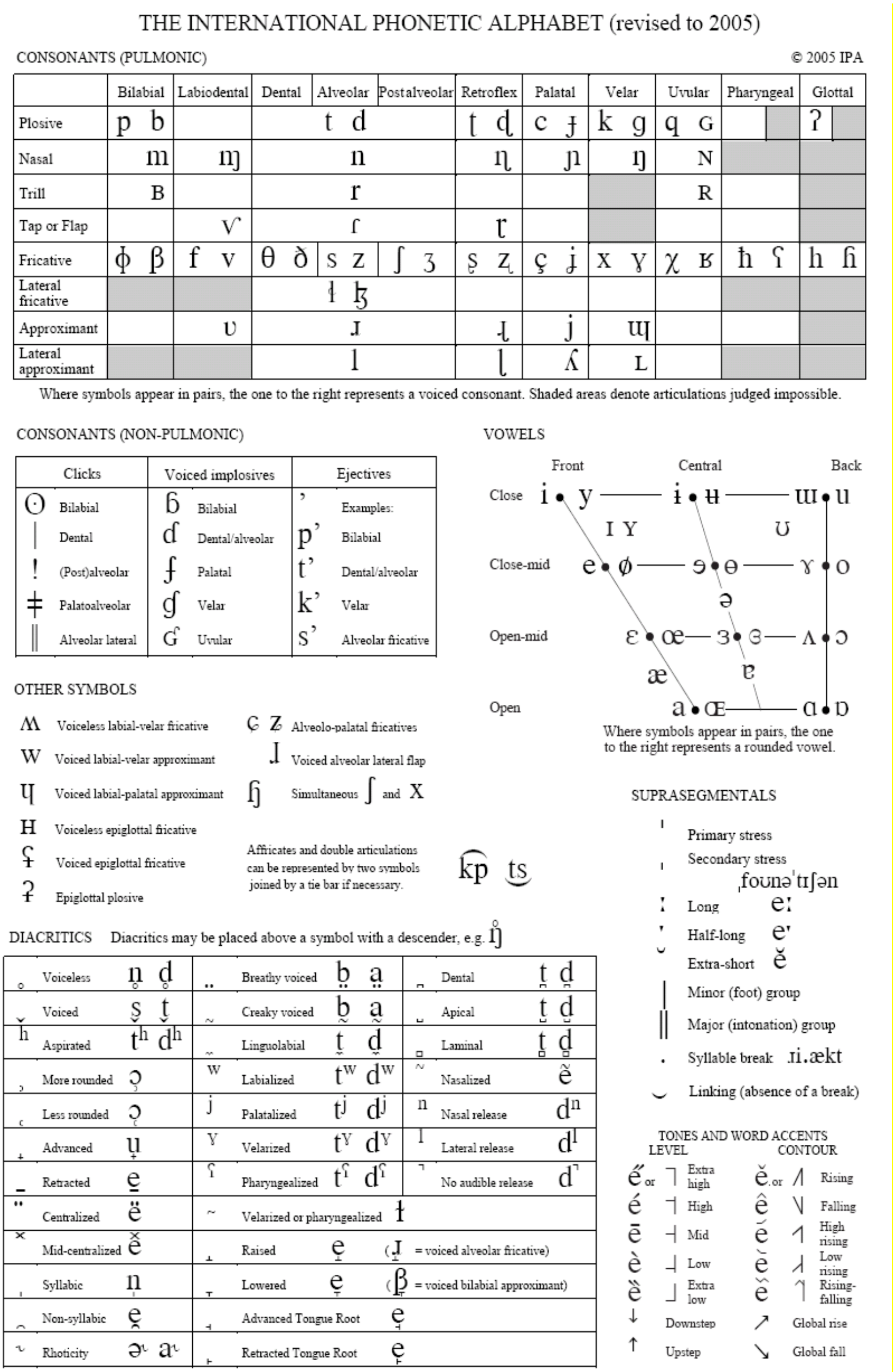
Manners of articulation are summarised in Table 1. Manners differ chiefly in the degree of obstruction, but also involved are the nasal/oral distinction and the central/lateral distinction. The rate of an articulatory manoeuvre is also relevant: for instance, if the tongue tip and blade make one brief closure against the alveolar ridge the result is called a tap, symbolised [ɾ], but a similar closure made at a slower rate will be a plosive [d].

3.2 The IPA Chart

Almost any sound may be voiceless or voiced regardless of its place or manner of production; and places and manners may be (with some restrictions) combined. The IPA chart takes the form of an array, with the columns being places of articulation, and the rows being manners. Voiceless and voiced symbols are put in that order in each cell. Blank cells on the IPA chart correspond to possible though unattested sound types, while shaded cells show impossible combinations

3.3 Vowel Classification

For vowels, the arched tongue body takes up various positions within the oral cavity. In the vowel [i] the tongue body is well forward in the mouth, beneath the hard palate, whereas in [u] it is pulled back. Both [i] and [u] have the tongue relatively high in the



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	ʼ	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

VOWELS

Front Central Back

Close i • y — ɨ • ʉ — ɯ • u

ɪ ʏ ʊ

Close-mid e • ø — ɘ • ɵ — ɤ • o

ə

Open-mid ɛ • œ — ɜ • ɞ — ʌ • ɔ

æ ɐ

Open a • ɶ — ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͡s

SUPRASEGMENTALS

ˈ Primary stress

ˌ Secondary stress

ˌfoʊnəˈtɪʃən

ː Long eː

ˑ Half-long eˑ

˘ Extra-short ĕ

| Minor (foot) group

‖ Major (intonation) group

. Syllable break ɹi.ækt

‿ Linking (absence of a break)

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

̥	Voiceless	n̥ d̥	̤	Breathy voiced	b̤ a̤	̪	Dental	t̪ d̪
̬	Voiced	s̬ t̬	̰	Creaky voiced	b̰ a̰	̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	̼	Linguolabial	t̼ d̼	̻	Laminal	t̻ d̻
̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	̃	Nasalized	ẽ
̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	̚	No audible release	d̚
̈	Centralized	ë	̴	Velarized or pharyngealized ɫ				
̽	Mid-centralized	e̽	̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
̩	Syllabic	n̩	̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
̯	Non-syllabic	e̯	̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ə˞ a˞	̙	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
e̋ or	˥	Extra high	ě or	˩˥	Rising
é	˦	High	ê	˥˩	Falling
ē	˧	Mid	e᷄	˧˥	High rising
è	˨	Low	e᷅	˩˨	Low rising
ȅ	˩	Extra low	e᷈	˧˦˧	Rising-falling
↓		Downstep	↗		Global rise
↑		Upstep	↘		Global fall

oral cavity, while the vowel [a] requires it to be lowered (the jaw may open to assist). This provides a two-dimensional vowel "space" in the oral cavity, with the dimensions high-low (also called close-open) and front-back. The lips provide a third independent factor. They may form a spread orifice, as in [i], or be rounded and protruded into a small opening, as in [u]. Using tongue position for [i] but adding lip-rounding in place of lip-spreading yields a vowel which the IPA symbolises [y] as heard in a French word such as *rue* [ry] "street".

The IPA presents the vowel space as a quadrilateral of standardised proportions, based partly on X-ray studies of tongue position during sustained vowel production. Steady-state vowels can be represented as points within this space and symbolised appropriately. Diphthongs (such as those heard in English *sound* or *noise*) can be represented as a movement within the space.

3.4 Further Aspects of Vowel Classification

Certain languages use nasalization to differentiate otherwise similar vowels. A nasalised vowel is produced with a lowered velum, adding the acoustic resonances of the nasal cavity and giving a distinct auditory effect. For example, French [sɛ] *sait* "(he/she) knows" has a non-nasalised (oral) vowel, while [sɛ̃] *saint* "saint" has the nasalized counterpart.

Vowels may also have 'r-colouring' or rhotacisation, produced by a modification of tongue shape, typically by combining a curled-back tongue-tip gesture with an otherwise normal vowel. It is heard in North American English in such vowels as that in *nurse*, or the second syllable of *letter*.

3.5 Multiple Articulation

Languages may make use of pairs of segments which are alike in voice, place, and manner but distinct in sound because of an accompanying secondary adjustment. In the RP variety of English, for instance, a voiced alveolar lateral consonant which precedes a vowel (as in *look*) is different from one which occurs after a vowel (such as *cool*). The second one has a raising of the back of the tongue (the sound is said to be velarised). For English the difference is automatically conditioned and carries no meaning but in many languages (for example, Russian) this type of difference is applied to numerous pairs of consonants, and utilised to create linguistic contrasts. Apart from this velarization, other common types of secondary articulation found in languages are labialisation (the addition of a labial stricture, usually lip-rounding), and palatalisation (simultaneous raising of the front of the tongue towards the palate). Constriction of the pharynx gives pharyngealisation, which is used in the "emphatic" consonants of Arabic.

If there are two simultaneous gestures of equal degree by independent articulators, the result is termed double articulation. For example the Yoruba word for "arm" (part of the body) is [akpá], where [kp] indicates a voiceless plosive formed and released at the bilabial and velar places simultaneously. This is termed a labial-velar plosive. The widespread approximant sound [w] is also a labial-velar double articulation.

3.6 Non-pulmonic Airstreams

The egressive pulmonic airstream is the basis of speech in all languages, but certain languages supplement this with non-pulmonic airstreams – mechanisms which produce short-term compressions or rarefactions effected in the vocal tract itself, used for certain of their consonant sounds. Non-pulmonic sound types found in languages are ejectives, implosives and clicks. Ejectives, symbolised with an apostrophe [p’ t’ k’ tʃ’] are the commonest type. Their articulation resembles that of ordinary voiceless plosives and affricates, but they are produced with a closed glottis, which is moved upwards during the production, shortening the vocal tract and compressing the air trapped behind the articulatory constriction. Release of the articulatory closure takes place (generally with a characteristic auditory effect, which can be relatively powerful) and the glottal closure is then maintained for at least a further short interval. The speaker will then generally return to the pulmonic airstream for the following sound. This mechanism can be called the egressive glottalic mechanism. By definition, the vocal folds are closed, so all ejectives must lack vocal fold vibration.

Implosives are made by moving the closed glottis down rather than up, giving the ingressive glottalic mechanism. The implosives commonly encountered in languages are voiced, rather than being simple reversals of the ejective type. In these, egressive lung air passes between the vibrating vocal folds at the same time as the larynx is in the process of lowering. They are symbolised with a rightwards hook attached to the symbol for a voiced plosive, as in [ɓ ɗ ɠ].

Clicks are widespread as paralinguistic noises (such as the "tut-tut" of disapproval) but found as linguistic sounds in relatively few languages. They are suction sounds formed by enclosing a volume of air in the mouth and then enlarging that volume by tongue movement, with a consequent reduction in pressure. The back of the enclosed volume is formed by the tongue in contact with the velum; the front closure may be completed by the lips, or by the tongue tip and blade. Clicks are categorised as bilabial, dental, (post)alveolar, palatoalveolar and alveolar lateral. The basic clicks mechanism is voiceless, but the remainder of the vocal tract may perform a wide range of click accompaniments, including voicing, aspiration, voice-plus-nasality, glottal closure and others

3.7 Beyond the Segment

Length, stress and pitch are classified as suprasegmental (or prosodic) features on the current version of the IPA chart, This means that that they do not apply to single segments, but to sequences of segments, or entire syllables.

Languages often distinguish short and long vowels, and (less commonly) short and long consonants (the latter equivalently termed geminates). Vowels paired as "long" and "short" within a language do not necessarily differ only in duration. In English, for example, the "short " vowel of *bit* is typically lower, and centralised, compared with the "long" vowel of *beat*, and there are similar quality differences in the other pairs.

Duration, loudness and pitch are all relevant to the marking of "stress", which is generally considered to be a property of a whole syllable. Languages which make use of stress typically use it to render one syllable within a polysyllabic word more prominent. The position of this syllable may be fixed for a particular language (so in Czech, the initial syllable is stressed, whereas in Polish the penultimate syllable is stressed) or alternatively free to occupy various positions and thus differentiate words (so in English the noun *import* is stressed on the first syllable, but the verb is stressed on the second). Stresses are important to listeners in the task of parsing the incoming string into words.

The IPA also provides a range of marks for indicating pitch levels and movements on syllables. In tone languages (such as the various kinds of Chinese), the pitch level or pitch movement applied to each syllable is fixed as part of the makeup of each word. In addition, all languages (whether or not they employ lexical tone) make use of intonation, which is pitch variation applied to utterances of phrase length. It typically shows the grouping of words into "chunks" of information, the relative importance of words within the phrase, and affects interpretation (for example by marking utterances as questions or statements). Intonation can be modelled by locating a small number of tone levels (or movements) at specific points in a phrase, from which the overall pitch contour can be derived by interpolation. It is very common for the endings of intonation patterns (terminal contours) to carry particular significance in interpretation.

4 Expressions of Individuality

The production of a spoken utterance can be described in terms of a sequence of processing stages in the speaker, starting from a desire to achieve a communicative goal, and ending with sound generation in the vocal system. In this section we will build on the discussion of the human organs of speech given in section 2, and the discussion of how they are exploited to create different phonetic gestures given in section 3, to present a catalogue of the ways in which speaker-specific characteristics influence these generation stages.

In section 4.1 we'll look at how a speaker uses language to achieve a particular communicative goal – here a speaker will show preferences for which language to use, which words to choose, which grammatical structures are most appropriate for the circumstances. In section 4.2 we'll look at the phonological stage of production - given an utterance, the speaker must plan which phonetic segments and which form of intonation and rhythm would be most appropriate. In section 4.3 we'll look at how the sequence of segments must in turn be realised as a continuous and dynamic series of phonetic gestures whereby the articulators move to realise the phonological form. In section 4.4 we'll see how the movements of the articulators, particularly the jaw, lips, tongue, soft palate and larynx creates sounds which themselves carry speaker information as well as the spoken message. In section 4.5 we'll see how all of these stages can be influenced by the mental and physiological state of the speaker, for example whether they are emotionally aroused, or whether they are intoxicated. In section 4.6 we'll describe ways in which speaker characteristics can be used to place an individual as a member of a number of groups. Finally, in section 4.7 we'll introduce two main application areas for information about speaker identity extracted from speech.

4.1 Individuality in Language and Language Use

Since there are about 5000 different languages in the world (the number varies depending on the definition of 'language') the choice of language used by a speaker can already be considered a distinguishing characteristic. While a few languages have large populations of speakers, many are very small, and often geographically isolated. Icelandic for example, has only approximately 280 000 speakers in the world. Many speakers also have competence in more than one language, so a speaker may be able to decide which language best suits a given circumstance. When a speaker uses a second language, it is also very common for their language use to be influenced by properties of their first language.

Even within a language, there can be variations in dialect – relative differences in the frequency of use of words and grammatical forms as well as variations in the pronunciation of words. For example, Scottish speakers of English might more frequently use the word *wee* to mean small, where other English speakers may use *little*. Or *I don't know* in British English, might be produced as *Me no know* in Jamaican English. Dialects do not only vary in a geographical sense; since language changes with time, older speakers may use different forms to younger speakers. Similarly dialect use may be indicative of a social grouping, perhaps related to socio-economic class, education or gender. Thus speaker A may tend to use *that's right* where speaker B prefers to say *OK* or speaker C says *cool*. Such words will then occur in the spoken discourse of a speaker with a higher frequency and are thus an individual feature of this particular speaker.

Of course different speakers will also react differently in different situations, so the very way in which a speaker chooses to use language to communicate in a specific situation can also be indicative of their identity.

4.2 Individuality in the Sound System

After an utterance has been constructed as a sequence of words, it is mapped into a set of speech sounds spread out across time. One component of this process deals with which phonetic segments are needed (this is called the segmental component), and the other is related to the stress pattern, timing and intonation of the sequence (this is called the supra-segmental component).

On a segmental level, the sound inventory used in the mental lexicon to represent the pronunciation of words can vary from speaker to speaker even for one language. This is one part of what is called the 'accent' of the speaker. For example most British English accents differentiate the words *Kahn*, *con* and *corn* using three different back open vowel qualities; however many American English accents use only two vowels in the three words (e.g. *Kahn* and *con* become homophones). Similarly, older speakers of English may differentiate the word *poor* from the word *paw* through the use of a [ʊə] diphthong that is not used by younger speakers.

As well as differences in the number and identity of segments in the sound inventory, accents may also differ in terms of the distribution of those segments across words in the lexicon. Both Northern and Southern British English have the vowels of *trap* and *palm*, but for the word *bath*, the Northern speakers use the first vowel, while Southern speakers use the second. Another frequently observed phonological variation across English accents is rhoticity – whether the accent expresses post-vocalic 'r' letters in the spelling of words as [r] sounds in their pronunciation.

On the supra-segmental level, speakers can vary in the way in which they seek to use intonation to select sentence functions (e.g. questions versus statements) or in the way in which particular words and phrases are highlighted (e.g. putting a focus on particular elements). For example, "uptalk" (the use of a rising terminal intonation on utterances that might otherwise be expected to have a simple fall) has recently become characteristic of younger speakers of English across a range of locations (Britain, America and Australia/New Zealand).

4.3 Individuality in Controlling the Speech Production Process

Given the phonological form of the utterance at the segmental and supra-segmental levels, the next stage in the process of speaking is the execution of the utterance through movements of the articulators. This process involves a sophisticated neural control system for the co-ordination of the many muscles involved in moving the tongue, jaw, lips, soft palate and larynx through the utterance. The complexity of this task introduces many possibilities for the expression of speaker characteristics.

A common way of modelling the motor control problem in speaking is to think of each phonological segment as specifying an articulatory gesture involving one or more articulators. In this model, speaking is then executing phonetic gestures in sequence. Fundamentally however, one gesture overlaps in time with the next, so that gestures can influence each other's form, a process called coarticulation. Importantly, the degree of coarticulation depends on the degree of gestural overlap, which may in turn depend on the amount of time available for the gestures to complete. Thus a speaker who speaks more quickly, or who decides to de-accent part of a particular utterance may show more coarticulatory behaviour than another. The consequences of coarticulation may be that the articulators do not reach the intended target position for the segment, or even that the segment has no measurable physical effect on the final production.

Speakers also vary in the particular form of gesture they use to implement a given underlying sound segment. This is another aspect of accent and gives rise to a lot of variation across individuals. Vowels are particularly variable: the exact gesture and hence the exact sound quality used to realise a particular vowel segment can vary widely. So, for example, the *trap* vowel in American English accents can vary widely in height. Consonants are affected too; for example, a recent innovation in English is the use by some speakers of a labio-dental approximant [ʋ] to implement phoneme /r/, more usually realized as a postalveolar approximant [ɹ].

The particular articulation used to realise a speech sound also varies according to context, and of course different speakers vary too in how much they are affected by the context. For example, some Southern British English speakers realise the velarised or "dark" variety of phoneme /l/ as a back rounded vowel rather than as a velarized lateral consonant in words like [bɪod] *build*.

Larynx settings are a rich source of speaker variation. As we have seen, changes in the degree of adduction and tension of the vocal folds in combination with changes in sub-glottal pressure lead to variations in voice quality. Speakers vary in their preferred, or default voice quality – particularly the degree of breathiness in the voice. Speakers can also vary the voice quality they use depending on the context, so that a breathier voice may be used for intimate communication, while increased vocal effort may be required for a noisy environment. Creaky voice can have discourse use, for example to mark the ends of topics or of dialogue turns.

Variations in vocal fold tension are used to manipulate the pitch of voiced sounds, and are the main means of implementing intonational changes in speech. Again the default pitch and pitch range used to realise particular intonation changes will vary from speaker to speaker. The mean pitch used by a speaker can vary according to the communicative context: people famously raise their pitch to talk to small children. The range used for an utterance can be quite small – giving a monotonous pitch – or quite large – giving a dramatic quality to the speech.

Speaking rate also varies from speaker to speaker, and this not only has consequences for the duration of individual syllables: a faster rate may also increase the degree of coarticulation between adjacent gestures.

4.4 Anatomical Influences on Individuality

The physical size of the organs of speech is a significant source of inter-speaker variation in the speech signal. The length of the vocal tract affects its acoustic properties as a resonance chamber and hence how it functions in shaping sound sources (see section 2.4). The length and mass of the vocal folds in the larynx influence the default pitch, pitch range and voice quality available to the speaker.

The influence of vocal tract size can be seen by considering the frequency response of a simple tube, closed at one end as we change its length. For a tube of length 17.6cm – the typical length of an adult male vocal tract – the first three resonances of the tube are close to 500, 1500 and 2500Hz. These frequencies are similar to the formant frequencies for a central vowel quality. These resonant frequencies scale inversely with the length of the tube, so that a 10% increase in the length of the tube leads to a 10% decrease in the resonant frequencies. This explains why adult female formant frequencies are higher than men on average, since a typical adult female vocal tract is shorter than that of an adult male. Of course there is considerable inter-speaker variation in vocal tract length, and some women have longer vocal tracts than some men. In addition, vocal tract length can be varied within a single speaker through adjustments to the height of the larynx and to the degree of lip protrusion.

Vocal folds vary across individuals in both size and mass, and this impacts the range of frequencies for which they can vibrate. Post-pubertal men have longer and thicker folds with a lower modal frequency compared to women and children. The range of frequencies available is indicated by the range used in singing, which for men is about 87-415Hz, while for women it is 184-880Hz. While it is possible to achieve vibrational frequencies outside these ranges, this usually involves changes in the quality of vibration: irregular creaky voice at the lower end, and falsetto voice (made with tense, rigid folds) at the upper. Individual speakers vary in both the range of frequencies they are capable of producing, and in the range of frequencies used in everyday speech. More typically, speakers use a range of only about one octave of fundamental frequency in normal speaking.

Phonation builds on the capability of the respiratory system to provide a large volume of air at a suitable, steady sub-glottal pressure. Respiratory volumes vary across individuals, and hence the quantity of speech that can be produced on one breath varies. This is also strongly influenced by the efficiency of phonation, with weaker adduction causing greater air loss.

The soft palate acts as a valve that isolates the nasal cavity from the oral cavity. Differences in the effectiveness of the valve and the way this is used in speaking can lead to changes in observed nasality of a speaker's voice.

4.5 Other Influences on Individuality

In the previous four sections we have considered how a speaker can impose his or her individuality on speech at different processing stages in speech production. In this section we look at how changes in the state of the speaker can also affect his or her speech. We'll look at changes over time, changes in emotion or changes in pathology.

A speaker’s voice does not remain constant since the speaker's vocal anatomy and physiology is affected by age. The larynx develops in children as they are growing, and its size and shape is particularly affected by hormonal changes at puberty, both for men and women. For men, the vocal folds can grow in size and mass over a short period, leading to a period of phonation instability as the speaker learns to control the new system. The vocal folds and their control are also affected by advancing age, and modal pitch, the degree of breathiness, and the degree of creakiness can all be affected.

The vocal tract itself also changes in size as a child grows, and this of course changes the range of resonant frequencies available. Control over the vocal tract, reflected in the degree of articulatory precision also develops in the first ten or so years of life. While vocal tract size remains relatively stable with advancing age, there may be significant degeneration in muscles, in the control process, and indeed in the ability to use language, such that speech becomes slower and less well articulated. Similar changes in speech can be brought about by physiological changes, such as tiredness or intoxication.

The emotional state of the speaker can have a great influence on the way in which speech is produced as well as on the content of the messages communicated. Increasing emotional arousal can raise the mean pitch and the pitch range as well as causing changes in loudness. Different emotions can have differing effects, so that it may be possible to differentiate emotional states such as anger, fear, sadness, joy and disgust, although speakers vary in exactly how these affect speech [5].

The health of the speaker can also influence his or her speech. Minor pathologies such as upper respiratory tract infections influence the larynx and the nasal cavities. Laryngitis is a swelling of the folds in response to infection which causes a lowering in pitch (due to the increase in mass of the folds), and can even prevent phonation occurring. Blocked nasal cavities can create a hypo-nasal form of speech that listeners recognise in speakers with a cold.

More serious pathological conditions, particularly stroke, can have effect on the parts of the brain responsible for speech planning and motor control – realised as aphasia and dysarthia. Damage to the vocal folds, such as swelling and the growth of nodules and polyps can also affect phonation and hence voice quality. Smoking and alcohol consumption have both been shown to cause vocal fold pathology.

4.6 From Individuality to Identity

We have shown that individuality can be expressed in many ways in speaking, at all levels of the message generation process. But while an individual speaker may exhibit a combination of characteristics that may make his or her own speech unique, it is very likely that each one of the characteristics is also used by other speakers. Thus another way of describing speaker characteristics is in terms of the groups of individuals which share a given feature. And another way of defining a speaker is in terms of the groups that the speaker is a member of.

We are used to grouping speakers on the basis of categories such as language, dialect and accent. However these may be much less well defined in practice than they are in theory. What differentiates a language from a dialect is not always clear. Sometimes, geopolitical factors, like a country's borders, influence the definition of the language of a speaker. Accents may be defined in both geographical and social terms; and people can be both geographically and socially mobile. A speaker might use a blend of languages or vary their accent according to circumstance.

Even if the groups are well defined, it may not always be easy to assign a speaker to a group. The very measurements we make of speech are prone to error, and the particular speech we measure may be unrepresentative of the speaker as a whole. For example it is not always the case that we can determine the sex of a speaker from their speech; and the estimation of age or physique can be quite difficult.

On the other hand, when the context is constrained, speaker characteristics can sometimes be used quite reliably to identify individuals. So when a friend on the telephone says *Hi, it’s me* then the combination of the observable speaker characteristics and the limited number of speakers known to you that might introduce themselves to you in this way may mean that the speaker can be accurately identified. This is not to say however, that you wouldn't be fooled by an impostor.

4.7 Applications

Within the field of Speech Science, much more emphasis has been placed on the scientific investigation of the linguistic content of utterances than on the investigation of speaker characteristics. To build an automatic speech recognition system that converts speech signals to text, for example, the speaker information must be discarded or ignored. Theories of production and perception focus on the strategies required to facilitate communication of words and meanings rather than on speaker identity. However two application areas for speaker information have emerged and these have led to an increasing interest in the individuality of voices. One is the field of forensic phonetics that deals with speaker identification in legal cases, the other is the technological field of speaker verification for voice access systems.

Forensic phonetics is a field in which phonetic knowledge is applied in legal cases where the identity of the speaker in a recording is disputed. Forensic phonetics distinguishes between two methodologies: identification of the speaker a) through the use of linguistically/phonetically naïve subjects, or b) through a trained expert witness

[6]. In method a) a witness of the crime (e.g. a person who has received sexually harassing phone calls) is asked to identify an alleged perpetrator by his/her voice by picking the speaker out from a 'line-up' of similar voices. In method b) a trained speech expert carries out an identification between the recorded voice and that of a specific suspect. This process is often done with a combination of auditory phonetic comparisons (the expert witness judges on the basis of his/her expert perception) or technical comparisons (analysis of fundamental frequency, formant frequencies, etc.). Speaker specific characteristics at all levels may be appropriate for forensic applications. The number of characteristics shared between recording and suspect increases the likelihood that the suspect made the recording, although it is much harder to estimate the likelihood that a person other than the suspect could have made the recording. Thus expert evidence in forensic speaker identification is better at eliminating suspects than in confirming them.

A second application that relies on speaker characteristic information is the field of speaker verification. Such systems can be used to secure access to a facility or resource, such as a building or a bank account. Typically a speaker is enrolled into the system using some known speech materials, and then the speaker is asked to verify his identity by producing a spoken utterance on demand. The main difference to forensic applications is that the speaker hopes to be identified successfully and is therefore willing to co-operate. To ensure that recordings of the speaker cannot be used to fool the system, a speaker verification system will typically request novel phrases to be produced to gain access – random digit strings for example. Most systems exploit low-level, speaker-specific spectral information found in the signal – that relating to pitch, voice quality, vowel quality and vocal tract length. This is because it is harder to extract robust speaker-specific information at higher levels. The restriction to low-level features also enables the possibility of text-independent verification, where speaker identity is verified without knowledge of what they are saying. It is hard to make speaker verification systems particularly accurate: false acceptance rates and false rejection rates of 5% or more are common. When a system is modified to accommodate the variability in production that occurs within the true speaker (when they have a cold or are tired, for example), this inevitably increases the success rate of impostors. The possibility that there is unused information present in the speech signal that would improve the performance of such systems is still open to investigation.

Acknowledgments

The IPA chart is used by permission of the International Phonetic Association.

References

1. Flanagan, J.: Speech Analysis, Synthesis and Perception. Springer, Heidelberg (1972)
2. Atkinson, M., McHanwell, S.: Basic Medical Science for Speech and Language Pathology Students. Whurr (2002)
3. International Phonetic Association: Handbook of the international phonetic association. Cambridge University Press (1999)

4. Ashby, M.: Phonetic classification. In: Brown, et al. (eds.) Encyclopedia of Language and Linguistics, 2nd edn., Elsevier, Amsterdam (2005)
5. Banse, R., Scherer, K.: Acoustic profiles in vocal emotional expression. Journal of Personality and Social Psychology 70, 614–636 (1996)
6. Nolan, F.: Speaker identification evidence: its forms, limitations, and roles. In: Proceedings of the conference 'Law and Language: Prospect and Retrospect', Levi Finland (2001)

Speaker Classification Concepts: Past, Present and Future

David R. Hill

University of Calgary

Dedicated to the memory of Walter Lawrence and Peter Ladefoged

Abstract. Speaker classification requires a sufficiently accurate functional description of speaker attributes and the resources used in speaking, to be able to produce new utterances mimicking the speaker's current physical, emotional and cognitive state, with the correct dialect, social class markers and speech habits. We lack adequate functional knowledge of why and how speakers produce the utterances they do, as well as adequate theoretical frameworks embodying the kinds of knowledge, resources and intentions they use. Rhythm and intonation - intimately linked in most language - provide a wealth of information relevant to speaker classification. Functional - as opposed to descriptive - models are needed. Segmental cues to speaker category, and markers for categories like fear, uncertainty, urgency, and confidence are largely un-researched. What Eckman and Friesen did for facial expression must be done for verbal expression. The chapter examines some potentially profitable research possibilities in context.

Keywords: voice morphing, impersonation, mimicry, socio-phonetics, speech forensics, speech research tools, speaker classification, speech segments, speech prosody, intonation, rhythm, formant sensitivity analysis, face recognition, emotional intelligence, dialogue dynamics, gnuspeech.

1 Introduction

Preamble. In an article in the Washington Post published February 1st 1999 science correspondent William Arkin reported on work at the Los Alamos National Laboratory in the US that claimed success in allowing arbitrary voices to be mimicked by computer [1]. One example was a recording purportedly by General Carl Steiner, former Commander-in-chief, U.S. Special Operations Command, in which he said:

> "Gentlemen! We have called you together to inform you that we are going to overthrow the United States government."

But it was not Steiner. Arkin reports it was the result of voice "morphing" technology developed at the Los Alamos National Laboratory in New Mexico, using just ten minutes of high quality digital recording of Steiner's voice to

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 21–46, 2007.

"clone" the speech patterns followed by the generation of the required speech in a research project led by George Papcun, one time member of the Computalker team. Steiner was sufficiently impressed, apparently, that he asked for a copy. Former Secretary of State, Colin Powell, was also mimicked saying something alleged to be equally unlikely.

The process was "near real-time" in 1999, and processor speeds have increased dramatically since then (say two orders of magnitude). This might suggest that the problem of characterizing a speaker's voice and using the characterization to fake arbitrary utterances is so well understood that the speaker classification problem is largely solved, and real-time mimicry no problem. This would be over-optimistic. Of course, the Los Alamos project appears to be secret, and further details are hard to obtain, but a recent job posting in the "foNETiks" newsletter [2] - requiring US Secret security clearance - makes it clear that mimicking human agents in all respects, including even things like eye movements, is an ongoing military research project.

> "Working as part of an interdisciplinary team, this research scientist will help create language-enabled synthetic entities that closely match human behavior. These synthetic entities will be integrated into training simulations to enhance training while reducing resource requirements. These synthetic entities are not gaming systems based on black-box AI techniques. They are cognitively transparent entities that exhibit human-like behavior at a fine-grained level of detail as supported by their implementation in the ACT-R cognitive architecture and validated via psychological measures like eye movements, reaction times and error rates." (From the job description)

1.1 Reasons for Wanting Speaker Classification

In approaching the problem of speaker classification, the first question has to be: "What is the purpose of the classification?" because the purpose sets the criteria for the task - a necessary precursor to choosing the tools, techniques and measures of success to be used.

Some of the reasons for attempting speaker classification in one form or another include: (1) a clustering exercise to allow automatic indexing of audio material; (2) identification or verification of individuals for ensuring secure access, including text-dependent and text-independent approaches; (3) determination of speaker characteristics and acoustic environment to facilitate a speech recognition task or tailor a machine dialogue to the needs and situation of the user; (4) a general characterization of a speaker type to allow a synthetic voice with similar characteristics (accent, gender, age...) to be generated - what we might term "Pronunciation Modelling"; or (5) a very specific characterization of a particular speaker to allow arbitrary speech to be generated that is indistinguishable from the speech expected from that speaker under a variety of external (physical and acoustic environment) and affective (emotional) conditions - or what we might term "Impersonation". It is interesting to note that, if impersonation could be carried out in real time, it could form the basis of very low bandwidth communication channels, along with other interesting military and commercial applications, including PsyOps. There are other reasons, including forensic analysis for purposes of law enforcement and categorizing speakers as part of the methodology in research on dialects and language acquisition. It is in such areas, particularly work by Paul Foulkes (York University, UK) and his co-workers at various universities, that some of the more interesting new directions have emerged in the field of what Foulkes calls sociophonetics [3].

Some reasons for speaker classification, such as identifying the language the speaker is using, can fit into more than one of the categories listed above. The same is true of modelling cognitive and emotional states. In any case, the list is not exhaustive.

More recently, a rising demand for human-like response systems has led to an increasing requirement for the ability to classify speakers in more general ways to permit, for example, machine dialogues to be tailored to clients - allowing shopping systems to recommend suitable goods appropriate to the age and sex of the shopper, or systems that can recognize urgency, confusion or perhaps language impairment. Such systems are reminiscent of the goals of the Air Force Research Laboratory project noted above [2].

When user modeling is important, and interaction is by voice, speaker classification becomes a very broad topic indeed, extending to the environment in which a speaker is operating, as well as the speaker's goals. Paralinguistic and context cues must be extracted along with more traditional speech analysis information and used in the procedures for categorizing speakers.

Recent papers providing an entry into the speaker recognition & speaker verification literature include [4], [5], [6], [7], [8], [9], [10], [11], [12] and [13]. Two early papers - still highly relevant to speaker identification and verification as well as offering insight into other speaker-dependent attributes of speech - are [14] & [15]. Many of these approaches are reminiscent of those used to try and solve the complementary problem of speech recognition, and typically involve: (a) slicing the speech into consecutive samples of a few milliseconds (the "salami" technique, involving slicing); and (b) carrying out some kind of automated statistical analysis (decision/pattern recognition strategy), on the data collected. Whilst these perhaps pay attention to some knowledge of speech structure - such as formant structure or underlying pitch - they have an unfortunate tendency to take a low-level approach to the treatment of the resulting data rather than an insightful approach (undoubtedly a result of pursuing goals of automating the process whilst minimizing the algorithms and CPU cycles involved). The tendency may also result from a necessary de-skilling of the classification task, so that only computing and mathematical skills are needed, even if linguists are consulted. It is tough to create teams, let alone find individuals, who possess all the skills that should be applied to an integrated and informed approach. Given the volumes of data involved, pressure to cast the problem into terms that facilitate uniform machine procedures is considerable. In this book we are concerned with speaker classification rather than verification and/or recognition.

It is necessary to turn our attention to speech structures, as these must be dealt with explicitly for more general classification tasks.

1.2 More Structured Approaches

If you are carrying out a statistical analysis of dinosaur bones when studying the characteristics of those prehistoric creatures, your results will not relate to any real dinosaur bones - let alone real dinosaurs- if you disregard the differences between bones due to gender, variety, and so on. Equally, if you wish to gather data relevant to classifying speakers, for whatever reason, you need to understand the attributes of speakers relevant to your required classification, rather than simply hoping that a genetic algorithm, neural net, Gaussian Mixture Model, or whatever will do the job for you. It might, but then again, it very well might not - at least, the classification will be nothing like as good as a properly informed discrimination that takes account of what you know about the populations of interest. This is why I am impressed by Paul Foulkes' work at York. He is doing for speaker classification what I have always regarded as the proper approach for speech recognition, namely carrying out careful experiments to reveal relevant speech structure.

By way of a very simple illustration of what I am talking about, consider what is perhaps the earliest recorded example of a solution to, and application of, the speaker classification problem - as recorded in the Bible:

> The Gileadites seized the fords of the Jordan and held them against Ephraim. When any Ephraimite who had escaped begged leave to cross, the men of Gilead asked him, 'Are you an Ephraimite?', and if he said, 'No', they would retort, 'Say Shibboleth'. He would say 'Sibboleth', and because he could not pronounce the word properly, they seized him and killed him at the fords of the Jordan. At that time forty-two thousand men of Ephraim lost their lives. (Judges 12, verses 5-6 [16])

One has to wonder how the error rate in this classification task compared to the error rate that might have been achieved with modern equipment doing statistical analyses of spectral sections and using Gaussian Mixture Models or similar techniques, but this early classification used very specific information about the population of interest, however flawed the science.

It is essential to use tests that are better than what amount to arbitrary statistical descriptions in order to begin improving our speaker classification, verification and identification techniques and success rates. Much of what is done these days seems little better than recognizing faces by comparing captured and stored photographic images, with some kind of statistical analysis of pixel groupings. Some success may be expected, but the approach is inherently limited unless a knowledge of the structure of faces, and the way the substructures vary, and relate to recognition/classification targets, is used.

This is the important step that Paul Foulkes has taken in the speech context, and is the issue to be addressed - in a rather specialized context - in the rest of this chapter. A full treatment is both impossible (as the research has yet to be done), and is beyond any reasonable scope.

2 Some Comments on Speaker Classification in the Context of Verification/Identification

People vary widely in their ability to recognize speakers, even those speakers they know well. Surprisingly, phoneticians seem no better than untrained listeners in distinguishing between different speakers ([17], quoted by [18]) so it is an open question as to just how the task is being performed. How do listeners who perform well recognize an idiosynchratic accent and idiolect, as opposed to making an accurate phonetic transcription of it? What other subtle cues do listeners use when identifying and distinguishing speakers, and gauging their affective and cognitive state, age and so on that are necessarily unrepresented in our phonetic/phonological models? These are important questions that are worth answering in order to make progress in improving machine procedures and performance at this difficult task. If accurate dialogue models can be constructed to include the use of pauses in turn taking, rhythm, and changes in pitch levels, intonation resources and the like, the give and take in dialogue may offer important clues for speaker classification that are outside current descriptive frameworks aimed at the phonological and semantic record. Research in such areas has been ongoing, though with a view to understanding dialogue behavior and revisiting meetings rather than classifying the participants - for example, the M4 project [19], as well as [20] and [21].

It is not even clear what categories of speaker and speaker state we could observe, if we knew how to, let alone what characteristics relate to these categories. We come back to the question of what categories are important, and how can we distinguish them - and thence to the question: “Why do you want to categorize the speakers, and what error rates are acceptable.” Presumably those wishing to pass by the Gileadites would have had rather strong views on that topic - especially if they had a lisp!

In their experiment on open set speaker identification by human listeners, Foulkes & Barron [18] found that voices which were less well identified nevertheless contained phonetic cues which were not found in some or even any of the other samples. This suggests that some cues, however salient they may appear to phoneticians, are not particularly useful diagnostics in the process of “live” speaker recognition. This suggests that if the “right” phonetic knowledge is used in structuring the cue determination for automatic speaker recognition it could be even more successful than recognition by listeners who know the speakers well. In this experiment, even the apparently obvious and well documented clue of “up talk” (high rising terminal intonation, or HRT) was not properly utilized by the listeners attempting to identify their friends, even while they made comments showing that they paid attention to pitch variation. It is interesting that, in pursuing this, the authors carried out a statistical analysis of pitch variation and tied the mean and standard deviation to the results as a basis for their interpretation, rather than examining the intonation patterns in more detail. Part of the reason for this is that we still do not have adequate models for the use of intonational resources. This is one area with which our own research has been concerned, but the research is based on a moving target. For example, Halliday’s model of British English intonation (and the associated underlying framework for rhythm) does not include the use of HRT in modern terms, even though the model includes it as a basic option for questions. This introduces some of the questions that have plagued us as we attempted to create a high-quality text-to-speech system - questions that lead naturally into some of the topics I feel are worth addressing. For example, if we had a good functional description of how people use intonational and rhythmic resources to achieve their goals, and a good way of recognizing these goals (rather than a description of specific patterns for a particular accent, such as British RP English), such information could likely be used to compare the meaningful differences between different speakers in different situations, and thereby effect a useful categorization. There is a great deal more choice in the use of intonational and rhythmic resources - at least for speakers without special training - than there is for voice quality, long-term spectral features, vowel quality and the like. Such features, properly extracted, and based on a well understood structure, would be valuable in some forms of speaker categorization. What we don't want to do is collect unstructured statistics in the hope that something will "pop out" of the data. That way we would simply wind up with a pile of "mythical dinosaur bones"!

As noted above, this chapter is not intended as a survey of speaker classification techniques, but as an outline of problems, possible solutions, and suggested directions for new research.

2.1 Some Comments on the Foulkes & Barron [18] Experiment

In describing their experimental design, Foulkes & Barron write:

> "Like McClelland [22], our study assesses SR by a group of people who know each other very well. Our group, however, consists of a set of young men who are university friends. This group was selected to investigate SR in a situation where the social profile of all group members is very similar in terms of age and gender (compare with McClelland's study, which involved men and women of various ages). ... "All were male, aged twenty or twenty-one, and formed a close social network. During their first year as students the ten had all lived together in shared student accommodation. Some of the network members spent large proportions of their academic time together, and they had all socialized with each other on a regular basis."

It is well known that one sure way to acquire the "right" accent in Britain is to attend the "right" school. That is how I acquired my own archaic RP accent. I was always amused when - on sabbatical - I crossed the Atlantic (both ways) in the Polish ocean liner "Stefan Batory" in company of other varied academics, diplomats and the like. I came across a family from the United States whose son, after one year at a British public school (British "public schools" are actually exclusive private boarding schools), had acquired an impeccable RP accent - indistinguishable from my own. His parents were quite embarrassed and puzzled. Peer group pressure - especially in a strange environment where the threats are unknown but often quite real - creates a tremendous and largely unconscious pressure to conform in all possible ways, including manner of speech, as I know from personal experience as well as observation. Note that all kinds of subtle cues are assimilated by the newcomer who adapts, and these combine to create a new accent which is acceptable to the group. This undoubtedly makes it more difficult to distinguish individuals.

The Foulkes and Barron experiment was well and insightfully designed, as experiments must be, to maximize the chance of revealing information relevant to the prior hypotheses. If you wish to maximize the chances of confusion, and eliminate the possibility that lack of familiarity contaminates the results with unknown factors, choosing a tightly knit social group - the members of which have actually lived together for a year in new surroundings and have acquired similar speaking habits whilst learning to ignore many of the differences - is exactly the right thing to do. You thereby gather useful information that might otherwise have been lost amongst the many possible confounding factors. Note that the chances of confusion were further increased by using telephone speech. In the context of forensics, and psychophysical experimentation, this is entirely appropriate. Working with a group of young people actually at a "public school" might have been even better but perhaps less practical.

Under such circumstances, the perceptually obvious cues that might identify individuals within the group will be considerably attenuated as the members aim to keep a low profile and fit in with the group. The differences that persist, however obvious to a linguist or a machine, are likely to be exactly those cues that are less important for identifying the speakers as different amongst themselves. At the same time, cues which are not so perceptually salient may well still be useful in categorizing or recognizing speakers by existing machine strategies. More than one approach is appropriate.

In speech recognition and synthesis, the arbiter of acceptable performance is ultimately the human listener - necessarily subjective, which, in turn, means that the perceptual consequences of any given speech determine whether the synthesis or recognition procedures were effective. Consequently, psychophysical studies of speech perception have provided a great deal of important information for those working in both areas.

In speaker verification and recognition, as well as some classification tasks, the ultimate arbiter of success is objective accuracy. Perceptual experiments have received less attention most likely because the classification procedures are amenable to objective measurement - for example, in the case of speaker identification, including foil rejection, classification by age, and classification by gender.

However, if the object is classification by some other criteria, such as accent, emotional state, cognitive state, and environmental effects, the problem once again becomes more subjective since objective measures of success are simply unavailable and success, or lack of it, must once again be based on perceptual judgments, whether by speaker or listener.

The same consideration also applies to a question like: “Does the speaker belong to the group who lived and studied together at university” rather than which particular individual is speaking - the kind of question that may be of considerable forensic interest these days and the answer may not be readily obtained by objective means until after the fact. Foulkes and Barron’s [18] experiment shows this very clearly since there were quite salient differences between the speakers in their experimental group that were apparently ignored by human listeners who must therefore have categorized the different speakers using only perceptually relevant cues that somehow had converged considerably towards a uniform state that caused significant confusion even amongst the in-group itself. The features needed to decide that a person was a member of the group were clearly different from the features needed to identify the same person within the group. Perceptual studies may help throw some light on the differences between these features, in concert with other approaches.

Perceptual studies are of interest in their own right, simply as a way of exploring the cues that arise from various factors, such as age, sexual orientation, or in-group membership.

3 Perceptual Studies

In research on recognition and synthesis, perceptual experiments have been powerful tools in resolving the important issues concerned with speech structure.

It might seem, in light of the reports by Shirt [17] & [23] that, since people - even trained linguists - have difficulty recognizing speakers, that the perceptual (i.e. subjective) effects of differences between speakers are less important than what might be termed the objective differences. This would be a misunderstanding as argued in the previous section.

For example, why are some listeners able to hear differences that other listeners cannot? Is there a continuous dimension for decision making or is it categorical? If the latter, how does the categorization threshold vary? Are listeners able to hear differences in the dimensions of interest and, if so what are the difference limens? When there are competing dimensions, which cues are the more powerful? If differences that are theoretically perceptible exist, which are ignored by listeners attempting to recognize speakers, what is the reason? It would seem that conventional linguistic training is not necessarily the issue in these and other investigative questions.

Field work is detailed, painstaking and demanding. Perceptual experiments are no less demanding, even if different skills, methods and tools are required. The work equally requires guidance from all the other sub-disciplines of linguistics and psychology. This in itself is demanding but, perhaps, the biggest stumbling block has been the absence of suitable instruments to pursue the experimental work since the cues sought are subtle and not well-understood which therefore demands a high-quality system to generate the experimental material in a controlled, accurate manner, within a matrix that is as natural as possible.

The invention of the sound spectrograph [24] Pattern Playback [25], the Parametric Artificial Talker [26], OVE II [27] and their successors were seminal inventions for our modern understanding of speech structure and speech perception for purposes of speech recognition and speech synthesis.

When the sound spectrograph first appeared, many considered that the problem of speech recognition was close to solution, if not solved, and it was quite a surprise that it took around two years to train observers well enough for them to recognize the "obvious" patterns revealed by the machine. Machine recognition remains a relatively unsolved problem, though programs like Dragon Naturally Speaking do a reasonably useful job by using a well designed interactive dialogue coupled with training to particular voices. Replicating human abilities is still a dream (and will remain so until we have better ways of representing the real world and using the information effectively - a core AI problem).

Pattern Playback could play back spectrograms of real speech and it wasn't long before Pierre Delattre, at the Haskins Labs in New York, hand-painted such spectrograms based on his experience - with real speech versions and perceptual experiments - to produce the first "real" [sic] synthetic speech. The rules "in his head" were soon made explicit [28], but the speech could not be called natural. It was not long before Holmes and his colleagues produced a completely automated text-to-segmental-level-speech synthesis system [29] to which Mattingly then added automated prosody [30].

With the invention of synthesizers that modeled some of the constraints on the human vocal tract, and Fant's seminal book (based on his thesis presented before the King of Sweden, with Walter Lawrence as the "third opponent") [31], progress in all speech areas accelerated and John Holmes, at the UK government "Joint Speech Research Unit" showed that it was possible to mimic human speech quite closely given enough care in preparing the input data [32]. But this exercise did not formalize just what it was that characterized a particular voice in any way that would be useful for carrying out general categorization tasks.

Many of the problems that plague solution of speech recognition and synthesis arise from exactly those aspects of speech that reveal information about the speaker. This provides one important reason for wishing to classify speakers, as noted in the introduction. There is also the problem that speech is a continuous acoustic recoding of articulatory gestures possessing no consistently clear boundaries in the ongoing spectrum of sounds - likened in the early days to an egg that has been scrambled. Both speech recognition and synthesis are still very much influenced by the linguists' phonetic analysis which determinedly inserts segment boundaries - an exercise that can be performed fairly consistently by those with suitable training, but an exercise that ignores the fact that many segmental boundaries are determined more by convention than acoustic reality. Those with a more phonological mind-set are much more concerned with the interdependency of successive segments and the succession of unsynchronized acoustic features. Where do you insert the boundary between, say, a stop and a following vowel to separate the acoustic features of the vowel from those of the stop, when important cues for the stop are embodied in the course of the formant transitions to the vowel, which don't even begin and end at the same time, and, as [33] showed, formant tracks are not consistent between different contexts. Boundaries in glide, vowel and liquid sequences tend to be even more arbitrary.

When Shearme and Holmes [33] examined the vowels in continuous speech in the hope that clusters characterizing them would emerge, they found there was a complete absence of such clustering. Thus speaker classification, to the extent that it needs to use the formant structure of vowels in the classification process, depends - at least to some extent - on speech recognition, so that the underlying phonetic structure can be used to recover useful vowel data. This puts both speech recognition and speaker classification into an interdependent relationship. You can help recognition by using information about the speaker, but speech recognition is needed to help classify the speaker.

It is possible that, in the cue-reduced environment of the telephone speech experiments referenced above [18], that this mutual support is sufficiently reduced that the listener's attention becomes focused on the primary task of recognizing what has been said, even though the ostensible task is to recognize who might be the speaker. It is also possible that, within the "in-group", listeners have learned to pay attention to some cues at the expense of others, and the cues that are used are simply absent from the band-limited telephone speech. Such questions need to be answered.

Fant [27] pointed out that the sound spectrograph was limited to an upper frequency of 3400 Hz whereas an upper limit of at least 8000 Hz was needed for unambiguous comparative description of unvoiced continuants and stops - a limitation with early research on recognition and synthesis. Voice quality - important in speaker classification - may need an even higher upper bound on the frequencies considered in assessing voice quality, because higher formant and other spectral cues are likely to be important. The interest in telephone quality speech arises because the telephone is ubiquitous, and forensic cases may have nothing other than sample of telephone speech. Also, by increasing the difficulty of the task in psychophysical experiments, there is a greater chance of producing statistically measurable results. However, the down side is that, since we don't really know what we are looking for, we may eliminate, or at least attenuate, the relative significance of the very factors we should actually be exploring.

Foulkes et al. [34] note:

> "it has also been shown that children learning different languages display subtle differences in the phonetic forms they use to realize a phonological category. For example, American and Swedish children aged 2-6 differ in place and manner of /t/ production, in accordance with differences found in the speech of American and Swedish adults [35]. Similar differences were found for vowel duration among the same children (page 2)" and "The (t) variants therefore involve subtle and highly complex differences in the coordination of oral and laryngeal gestures. (page 14)"

A useful additional tool for investigating and understanding the nature and importance of such differences in the speech of different speakers would be an articulatory synthesizer, since experimenters could use artificial stimuli, systematically varying such subtle cues under controlled conditions, to determine their perceptual effect. This author believes it is time to consider perceptual experiments using artificial stimuli that can closely mimic real human speech with a full spectral range.

A good quality articulatory synthesizer that is easily but appropriately controlled, and which is inherently restricted to the potential acoustic output of a real human vocal tract, with supporting models to provide a foundation for manipulating speech production from the lowest sub-phonetic articulatory level up to the prosodic level of rhythm and intonation would go a long way to providing the tool needed for such perceptual experiments. As Cooper et al. [25] pointed out in connection with earlier speech research, there are "many questions about the relation between acoustic stimulus and auditory perception which cannot be answered merely by an inspection of of spectrograms, no matter how numerous or varied these may be"

3.1 The Gnuspeech System

It has been the goal of building such an articulatory synthesizer and the necessary supporting models that has formed the subject of ongoing research by the present author and his colleagues [36]. The synthesis research has been ongoing for many years, first in the author's laboratory at the University of Calgary, then in 1990 it became the subject of a technology transfer exercise (to the now-defunct Trillium Sound Research - killed by the demise of NeXT Computer), and is now available to all under a General Public License as Gnuspeech - an ongoing GNU project [37]. Significant components of the complete, successful experimental system for articulatory synthesis that was developed on the NeXT have been ported to the Macintosh computer under OS/X and work is also under way to port it to GNU/Linux.

The complete system has been described elsewhere [36], [38], [39], and the source code is available for both the NeXT and the Macintosh (which is being modified to compile under GNUStep for GNU/Linux - though this port is not yet complete). Suffice it so say here that the approach builds on work carried out by Fant and Pauli [40] and by Carré [41]. By applying formant sensitivity analysis, and understanding the relationship of the resulting "Distinctive Regions" (Carré's term) to the articulatory possibilities inherent in the human vocal apparatus, the control problem for an articulatory synthesizer has been largely solved. In addition, the speed of modern computers allows the necessary complex computations for artificial speech based on the waveguide acoustic tube model to be carried out at a higher rate than is needed for real-time performance. Thus a tool is now available that allows experiments with the timing and form of articulatory events - with the caveat that the transformation between explicit articulatory specifications (such as tongue and jaw movements) and the Distinctive Region Model (DRM) equivalents has not yet been implemented, though the transformations are considered to be relatively straightforward.

3.2 Possibilities and Current Limitations of the Experimental System

The articulatory synthesis tools that have been developed do enable significant experiments on the effects of learned speech articulations to be performed. The tools could be improved and extended in a number of ways to make such work easier by enabling easier control of some of the characteristics to be investigated (for example, by implementing the transformation between articulator movements and the DRM equivalents). The tools and interfaces were only an initial implementation of what was needed to develop databases for a complete English text-to-speech system based on the articulatory synthesizer, rather than full a psychophysical/linguistic laboratory tool-set. However, the system as it stands allows experiments with the timing of articulatory events to be performed, based on the observation that the DRM captures the essence of human articulatory possibilities. It has been used already to look at geriatric articulation and the timing of stops and stop bursts.

Perhaps most importantly, since the system provides a complete text-to-speech system based on better, effective models of speech production, rhythm and intonation, experiments on particular characteristics will be embedded in a context that provides natural variation of *all* formants, with good rhythm and intonation and with accurate records of what variations were used. Making all variations, rules and data involved in any synthesis formally explicit and editable was an important goal of the system development.

An important limitation of the system is the reality that it is actually still a hybrid system. Though the acoustic tubes representing the oral and nasal cavities give a true simulation of the acoustic behavior of the appropriate human anatomy, with higher formants properly represented and variable, with inherently correct energy balances, and with simulation of oral and nasal radiation characteristics, the larynx waveform is injected directly - albeit from a wave-table that can be dynamically varied - and the fricative, aspiration and other noises (such as bursts) are also injected at appropriate places in the tube model. This latter arrangement provides the basis for appropriate fricative formant transitional behavior but the spectra of the injected noises are generic and approximate rather than individual and detailed. A better model would emulate the vibrating vocal folds, and oral tract constrictions, to generate the glottal waveform and noise spectra aerodynamically, based on accurate physiological models. The properties of all these noise spectra are characteristic of individual speakers.

The rhythm and intonation components of the system are based on the work of Jones [42], Pike [43], Jassem [44], Lehiste & Peterson [45], Abercrombie [46], Halliday [47], Allen [48], Ladefoged [49], Pierrehumbert [50], and Willems et al. [51], amongst many others as well as work in the author's laboratory at the U of Calgary. Wiktor Jassem spent a year in that lab and the results of the joint rhythm studies carried out are reported in [52] and [53]. Some of the intonation studies are reported in [54], [55] and [56]. Subsequent unpublished work on the intonation patterns has achieved significant improvement by using smoothed intonation contours, based on the timing events suggested by Allen [48]. Note that the original formant synthesizer used in our early research was replaced, in 1993-4, by the articulatory synthesizer described in [36], previously cited.

4 Face Recognition as an Analog of the Speaker Recognition Problem

Zhao and his colleagues [57] have provided a good summary of the state of the art in facial recognition. They conclude that face recognition is a dedicated process that is separate from normal object recognition and that both holistic and feature recognition strategies play a part, and facial expression seems to be recognized by separate mechanisms (somewhat as identity and location of objects are processed by different mechanisms in visual processing generally). Hair, face outline, eyes and mouth are significant, while the nose appears to play an insignificant role. Low spatial frequency components, bandpass components, and high frequency components seem to play different roles, with low frequency components permitting judgments concerning sex, but high frequency components being needed for individual identification. Other factors play a role, including lighting direction and being able to observe moving images rather than still photographs. Such observations may contain clues concerning how to approach speaker recognition and classification, with the major observation that the process is undoubtedly more complex than might be thought, and almost certainly involves different specialized mechanisms performing different tasks. Dynamic aspects are almost certainly important, and some kind of functional feature analysis, in addition to holistic measures, is likely to help.

Simply taking statistical measures of energy variation in the spectrum, or pitch values, and the like, is akin to trying to recognize faces from photographs based on a statistical comparison of pixel characteristics (spatial frequencies, distribution of pixel densities, and the like), without trying to identify features such as hair, eyes, mouth and so on, as well as relevant dynamic clues. It is the dinosaur bone problem. If you don't take account of the underlying structure of the data, your statistics become too unfocused to relate to reality in any precise way. As the Zhao et al. [57] state, quite clearly, for face recognition:

> "Feature extraction is the key to both face segmentation and recognition, as it is to any pattern classification task. For a comprehensive review of this subject see [58]."

It is also worth noting that the ability to observe a speaker's face affects the listener's ability to understand what the speaker is saying - from the fused perception of the McGurk effect [59] to the extreme form of lip-reading. Speech recognition is clearly multi-modal which, if nothing else, helps to illustrate some of the complexity of perception, and indicates that speaker classification is also likely multi-modal to the extent that cues other than voice are available.

In achieving good mimicry of a speaker, whether by voice alone, or using additional cues such as facial expression and body language, the speaker mimic needs to do more than imitate voice quality, intonation, and accent. The mimic succeeds best if he or she captures the appropriate “persona” of the person being mimicked - cool, excited, in control, sympathetic and so on - relating closely to *how the target would be expected to act* in the same circumstances.

Perception is a complex, active, organizing process based on assumptions that work, in the real world. It is not passive. We see the moon as increasing in size, the nearer it is to the horizon, because we are increasingly compelled to see it as increasingly far away. The fusion of the McGurk effect (hearing /b/, seeing /k/ and perceiving /d/, for example) arises because we have to reconcile the sound we hear with the conflicting appearance of the speakers lips and jaw. Close the eyes, and we perceive the sound that was actually produced. This is not the place to become diverted into a treatise on perception, but its active, organizing nature is well documented in the literature. There is no reason to suppose that our approach to recognizing speakers is any different, whether the categorization is broad or narrow. The extent we succeed or fail in the task is a measure of the cues to which we learn to pay attention and those we learn to ignore, just as with learning to recognize the sounds of our native language as infants [60].

5 Back to the Main Goal - Speaker Classification

In his survey of speaker recognition, Atal ([14] p 460) asks: “How do listeners differentiate among speakers?” and states that a satisfactory answer is not easy. In his review of automatic speaker verification in the same special issue of the IEEE proceedings, Rosenberg ([15] p 480) says that, for foils, mimicking behavior and learned characteristics is less successful than obtaining a strong physiological correlation, but then quotes an experiment showing that even an identical twin was unable to imitate the enrolled sibling well enough to get accepted by a verification system when attempting to foil the system.

This tells us three things. First that, even in 1976, verification techniques were amazingly effective; secondly that possessing identical physiology did not give the advantage that might have been expected, given his earlier remarks; and thirdly, that the verification methods used must have captured some aspects of the speaker twins other than physiologically determined characteristics - somewhat refuting the notion that physiology was the core characteristic for discrimination. It also tells us that we have to look at learned speaking behavior and other factors even for speaker recognition and verification, let alone for speaker classification.

Chollet and Homayounpour [7] carried out an extensive study to test the ability of listeners to discriminate the voices of twins. Family members were significantly better at the task than listeners not familiar with the twins, and the latter did not perform significantly differently from the two automatic procedures based on low-level acoustic features that were also tested. The authors conclude, amongst other things, that a speaker verification system which takes account of a speaker’s behavioral characteristics will be more robust against foiling by a twin with a similar voice.

How does the problem change if, instead of wishing to verify an enrolled speaker, or identify a speaker from a group of speakers, the task is to determine something about the speaker such as age, sex, emotional state, the speaker's feeling of confidence, and so on? The reasons for wanting such information can be quite varied. Müller, in his thesis and recent papers, [61] and [62] describes his AGENDER system, designed to obtain information about age and sex of speakers to allow an automatic shopping assistant to help a customer more effectively by tailoring its purchase recommendations based on the information extracted. A more ambitious goal is to understand the cues and behavior in speech for purposes of synthesis, to create more realistic artificial agents. The German Research Center for Artificial Intelligence in Bremen has a project to create a "Virtual Human" [63]:

> "Creating a virtual figure as a conversation partner requires detailed, anthropomorphic design of the character, realistic speech, and emotional interactions, as well as, the exact simulation of movement in real time." (from the web site)

Such ambition goes beyond the scope of this chapter, but illustrates the directions of research of interest for both speaker classification and speech synthesis, and ties together the synthesis of speech, facial expression and body language - a topic that has also been of interest in this author's lab [64] and [65]. It also parallels the work at the US Air Force Research Lab in Mesa, Arizona [2] previously cited.

Understanding the basis for adapting to speakers, according to their condition, type and situation, and responding appropriately, are increasingly important as machines become more involved in significant dialogues with people. The many reasons that people detest current voice response systems comprise their one-size-fits-all approach to dialogue, coupled with painfully slow exploration of many possible choices, most of which are irrelevant to the caller, together with their total lack of natural dialogue and empathy, including their inability to assess urgency, puzzlement, or other dialogue conditions, as well as their stereotypical and mechanical approach to even the lowest levels of social nicety. Machines currently exhibit low emotional intelligence [66] - at least partly because they have little basis for performing appropriate speaker classification at present.

Although there has been considerable success in using multidimensional classification methods on "feature vectors" derived from various kinds of speech analysis, less work has been done on approaches to classification involving the explicit extraction of features known to be associated with the distinctions that the classifier is expected to make. In order to extend this work more information concerning such features and their relevance is needed. Given that - unlike identification, verification, sex or age determination - the judgments are less objective, it is necessary to understand the subjective aspects of speaker classification, at least as a precursor to or test for the development of more objective measures.

5.1 Intonation and Rhythm

Pitch has been successfully used in speaker identification and verification, but only at a statistical level, without a lot of attention to its *functional* patterning - where "functional" includes involuntary effects (such as the effects of fear on larynx performance) that would be important for more general classification tasks. It would be hard to get ethics approval for an experiment in which people were made so afraid that it affected their voice, and perhaps just as hard to make people that afraid in an experimental setting. This is an example where a good synthetic speech emulation of relevant factors could produce output for judgment by listeners as a means of exploring the markers for fear. Of course, this raises the question of what listeners are able to perceive versus what changes occur when the speaker is afraid. Given the reports already cited concerning speaker identification, it would seem that even trained listeners do not hear differences in speech characteristics that are measurable. At the same time, not all measurable differences are necessarily relevant to either speech recognition or speaker classification.

It is worth reiterating that picking up clues relevant to speaker classification may be made easier if the speech can be recognized, just as recognizing speech may be made easier if there has been at least some degree of speaker classification. For example, recognizing an accent is likely to be enhanced by a procedure that identifies vocalic segments, or recognizes words that often contain glottal stops substituted for other stops in particular accents/dialects. Reductionism is no longer the best approach.

The larynx, which creates pitch pulses, functions at both segmental and suprasegmental levels, and both levels are relevant to various aspects of speaker classification. At the segmental level, for example, relative variation in voice onset time (VOT) in the transition from voiceless stops to voiced segments or for initial voiced stops can provide information relevant to linguistic background, sex and age [67]. Much of the work to date has focused on the relevance of VOT at the segmental level, rather than as a cue for speaker classification. Such a measurement depends on identification of the segments concerned - that is, on speech recognition.

At the suprasegmental level, the frequency and amplitude of pitch pulses vary, and give information about intonation pattern and rhythm. These represent some of the dynamic features of speech in which we need to take an interest for speaker classification. By far the most important determinant of rhythm is the relative *duration* of the underlying segments - particularly nuclear vowels - which are also associated with significant features of pitch change. The precise timing of the pitch changes relative to the segmental structure is almost certainly a useful clue to speakers and their characteristics (see also [48]). Again, speech recognition and speaker classification characteristics are mutually supportive and greater success in either is likely to be achieved if both are done in concert.

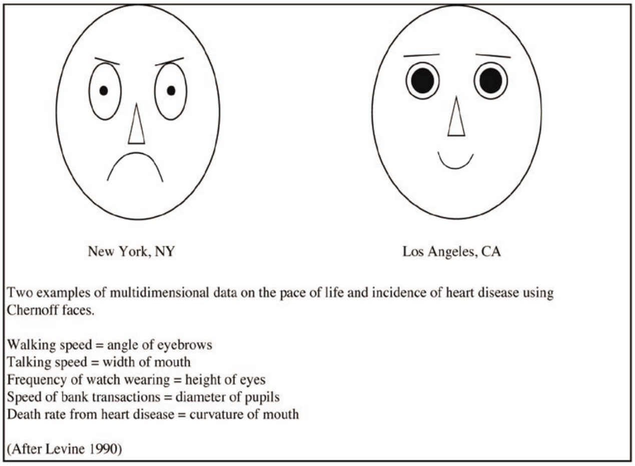
It is not clear to what extent the same is true of jitter (short-term pitch period variation) and shimmer (short-term pitch amplitude variation) nor how are these might be affected by rage, nervousness, age, illness, and so on. If they have relevance, is the effect more pronounced at semantically significant or phonetically significant places in the utterance. We simply don't know.

Modelling rhythm and intonation, even in general, have proved to be contentious topics for decades, and there is a plethora of different approaches to characterizing intonation patterns and describing rhythm. One of the more obvious splits on rhythmic description is between those who consider English to have a tendency towards isochrony (equal durations between "beats" [46] and [47]), and those who say such a phenomenon is a fiction - an artifact of perception and the phonetic structure of words. Though I have not seen anyone say this explicitly, it could be that this is a difference between American English and British English. Certainly we found that "a tendency towards isochrony" accounted for 10% of the variance in segment duration in the body of British RP English that we examined in detail [52], [53], and our results have more recently been supported by work at the Centre for Speech Technology Research in Edinburgh [68]. The degree of "tendency towards isochrony" is another potential characteristic that may help categories speakers. There are other aspects of rhythmic patterning and rate of speech that provide cues to the speaker class and condition such as pause patterns (think of Churchill, the wartime UK prime minister, and the way he spoke).

Intonation patterns provide a varied and shifting target since there seem to be many varieties of English intonation, and new forms arise before the old models have been evaluated and tested - a case in point being the arrival of "up talk" in which a pattern with rising pitch at the end of utterances seems to have become very popular, and supposedly derives from the "valley talk" which had its genesis in the 1960s in California. Other intonational patterns, if recognized and described, would provide valuable information concerning speaker class and condition. Again, think of Churchill's intonation as well as his rhythm when indulging in his famous oratory.

The question is not so much: "What is a good model of British or American English intonation?"; as "What resources are available to speakers, and how can these be clearly characterized in order to detect similar patterns in different groups of speakers, or determine their emotional and other psychological states?". What Eckman and Friesen [69], [70], did for the face we must do for speech, at least for intonation and rhythm, which seems the obvious place to start. Prosody (rhythm and intonation) probably plays an auditory role akin to that played by facial expression and body language in the visual domain. Humans have a peculiar sensitivity to facial expression - which is what prompted Levine [71] to use facial caricatures developed by Chernoff [72] to present multidimensional statistical data, as illustrated in Figure 1.

The choice of correspondence was done so insightfully that not only are the expressions produced very appropriate to the data being illustrated, but the editors felt compelled to disclaim any suggestion that the expressions indicated the mood of people in the cities from which the data were drawn. Prosody may offer similar possibilities in the auditory domain but we simply don't know. We do know that if analogous correspondences were found, we should have found a



Two examples of multidimensional data on the pace of life and incidence of heart disease using Chernoff faces.

Walking speed = angle of eyebrows
Talking speed = width of mouth
Frequency of watch wearing = height of eyes
Speed of bank transactions = diameter of pupils
Death rate from heart disease = curvature of mouth

(After Levine 1990)

Fig. 1. Chernoff faces showing the pace of life and heart disease

powerful possibility for characterizing speaker state from prosody. Many readers will be familiar with the voice of Marvin "the metal man" in Douglas Adams' *Hitch-hikers Guide to the Galaxy* which provided an excellent auditory caricature of depression.

It would be interesting to perform experiments to characterize, and determine human sensitivity to, "tone of voice" - which would include both rhythm and intonation, as well as voice quality and facial expression. Understanding "tone of voice" in a formal sense would be an important step in dealing with a range of speaker classification tasks.

People with hearing impairments have identifiable differences in their rhythm and intonation - in fact, people with total hearing loss must undergo regular speech therapy to keep their voices in a reasonably normal state.

Abberton was one of the pioneers of speaker identification using solely information concerning intonation patterns [73]. She pointed out that intonation not only contains useful information for speaker identification, but also contains considerable information relevant to speaker classification. She included synthetic stimuli in the listening trials of her experiment to control for potential confounding factors. She quotes earlier experiments by Brown, Strong and Rencher [74] who also conducted listening experiments with synthetic speech to investigate the relationship between perceived "benevolence" and "competence" speaker characteristics. It is reasonable to suppose that clues may be obtained that relate to many factors such as: anger, enthusiasm, ethnicity/native language, fear, urgency, uncertainty, lying, submission, puzzlement, frustration, aggression, dominance and confidence. Both analytical and synthetic experiments would be appropriate. The ability to caricature any of these factors in synthetic speech would, as noted, be dramatic and informative. To the extent that subjective evaluation is important, hypotheses derived on the basis of analysis are best tested and validated by perceptual experiments using synthesis, provided the parameters of interest can be controlled in a reasonable way.

5.2 Lower Level Cues: Segmental Level and Below

Suitable speaker verification/identification techniques undoubtedly extract measures closely related to the shapes and rates of formant transitions. These are characteristic of individual speakers who have learned speech habits. However, although this ability may serve its purpose, it does not contribute to knowing how to categories speakers as opposed to how to recognize or verify them, partly because the nature of the features extracted are hidden inside the complexities of the automatic decision procedures that are the norm for such tasks these days; and partly because even if they were not hidden, or could be discovered, the information is not structured by, nor related to any knowledge of particular classification categories being sought. Determining that a particular speaker is who he or she claims to be, or identifying which individual in a group is the one speaking is not the same as assigning such a speaker to any one of the large variety of possible categories listed at the end of the previous section (5.1), which is not exhaustive.

The question, as for higher level features, is not how do individuals differ in their acoustic characteristics when speaking, but in what ways are the members of some category of interest similar. This is a very different problem. To solve the problem requires that we examine the speech of ingroup and out-group speakers, formulate hypotheses about similarities and differences, and then test these in various ways, including by synthesizing speech with and without the characteristics that seem to identify in-group versus out-group individuals. Systematic psychophysical experiments can also be used directly as a way of finding of what affects perception that the voice belongs to a particular group, just as perceptual tests in the early days allowed researchers to find out what acoustic characteristics were essential for the perception of particular categories of speech sound.

Some characteristics at the segmental level that may be of interest for categorizing speakers include: relative formant amplitudes; rates and shapes of formant change; rates and shapes of articulatory movements (closely related to the previous item, but wider); formant ratios and values in known vowels; quality of vowels in known words (degree of reduction, actual formant values ...); segment durations & statistics; segment ellipsis & substitution; use of markers such as glottal stops versus other stops; rate of speech; relative event timing at the segmental level (Voice Onset Time and stop durations are examples); spectral manifestations of sinuses and other physiological structures; nasalization; and nasal spectrum. Again, note that speech recognition is an essential adjunct to extracting the features relevant to speaker classification.

In their paper on vowel clustering, already cited [33], Shearme and Holmes identified three generalizations concerning vowel formants, apart from the lack of signs of clustering. One was that plotting the formant tracks for a given speaker for each vowel produced could be used to draw relatively small areas containing at least a small portion of every track. The second was that these areas were significantly different for each speaker. The third was that each speaker's derived vowel-track F1-F2 areas were considerably displaced from the F1-F2 areas for the same vowels they produced in isolated monosyllables.

In a related but different experiment, using Lawrence's Parametric Artificial Talker - PAT [26], Broadbent & Ladefoged [75] synthesized sentences with different mean formant frequencies, all saying: "Please say what this word is". They also synthesized single-word stimuli "bit", "bet", and "bat". The single-word stimuli were presented to listeners, accompanied by different versions of the sentence, in an experimental design that provided an hour's intelligence-testing between two presentations of a test sentence and a stimulus word. There were seven different groups of subjects in which the details of the sentence/stimulus-word presentation varied - especially the delay between each sentence and the stimulus word. Most groups heard the sentence, followed - after a delay (depending on the group) - by the stimulus word. In one group, the stimulus word was presented first. The latter group showed little effect of the sentence on perception. A second group that counted during a 10 second delay also showed little effect.

Except for the conditions noted, it was found overall that the way speakers categorized the stimulus words as "bit", "bet" or "bat" was related in a simple way suggestive of a typical perceptual adaptation effect to variation in the mean formant frequencies of the preceding sentence. The same stimulus would be perceived as a different word by the same listener, depending on the formant frequencies of the preceding sentence. There were some unexplained anomalies.

These experiments suggest: first, the actual spectral quality of the vowels is less important than the dynamics of the formant transitions from point of view of recognition; and secondly, if the appropriate small areas containing at least part of all a speaker's vowel formant tracks could be determined, these could be powerful clues to speaker classification - or at least speaker verification/identification.

5.3 Dynamics and Longer Term Effects

Adami et al. [76] comment that: "Most current state-of-the-art automatic speaker recognition systems extract speaker-dependent features by looking at short-term spectral information. This approach ignores long-term information that can convey supra-segmental information, such as prosodics and speaking style." Their system, which uses Gaussian Mixture Modeling claims 3.7% error rates in speaker recognition (presented as a 77% relative improvement over other approaches), and they plan work on formants. This represents a small departure from the common obsession with "feature" selection, as opposed to looking at function and underlying mechanisms, even though their goal is only speaker recognition rather than classification.

Part of recognizing a speaker is the dynamic *interactive* aspect - how they react in dialog, what choice of words and argument structure they use, how they signal how they are feeling, and so on. Similar characteristics are likely relevant to classifying speakers into groups but we need to understand how all these potential markers relate to the groups in which we are interested.

5.4 Recognizing Speakers Gender and Age, and Sexual Orientation

Carlson et al. [77] noted that “Special effort is invested in the creation of a female voice. Transformations by global rules of male parameters are not judged to be sufficient. Changes in definitions and rules are made according to data from a natural female voice.” Such differences arise at both the segmental [67] and suprasegmental levels.

In producing convincing female speech from the *Gnuspeech* synthesizer, we found similar problems. Early in the development, Leonard Manzara produced three versions of the utterance: “Hello”. By adding “breathiness” to the glottal excitation (one of the available utterance rate parameters) and by judiciously crafting the intonation and rhythm, a reasonably convincing female voice version was produced, using the standard rules for the segmental level synthesis. The male and child voices were less trouble, though the child voice is probably more like a boy than a girl. All the voices could probably be improved if we understood the markers better, and this would provide a better basis for making these important categorizations is speaker classification. The relevant speech synthesis examples are provided for listening in connection with [36] which is available online. A great deal more understanding of the differences between male, female, and child voices, as well as more general markers for age, is required - research that is likely best carried out using an articulatory synthesizer similar to the *Gnuspeech* system in concert with careful study of relevant spectrographic data and previous research.

It seems probable that the markers involved in these kinds of distinction also play a part in the voice quality and intonation often associated with speakers with specific sexual orientation - for example, some gay men. By casting the research net wider to encompass such speaker categorization, even more should be learned about the resources *all* speakers use to project their identity through speech, and assist with speaker classification.

6 Conclusions

A major conclusion is that speaker classification requires the isolation of features relevant to specific kinds of categorization task, and that many of these features can only be extracted on the basis of a reasonable capability for recognizing what has been said - that is, by speech recognition - and by using other knowledge about the structure of speech, with better ways of characterizing the resources used for such speech attributes as rhythm and intonation. Without such informed structuring of the data, and identification of the linguistic and paralinguistic structure, any statistics that are derived may allow reasonable success at identifying or verifying particular speakers (on the same principle that photographic comparisons may allow people to be identified or verified on the basis of pixel images), but the “bones” that have been identified will be very hard to classify into meaningful groups of the different kinds needed for useful speaker classification.

A second conclusion is that, for any speech or speaker recognition task, much greater explicit attention should be paid to dynamic properties of the speech signal, at both the segmental and the suprasegmental level. Additionally, dialogue structure information - also dynamic - can provide important information.

A third major conclusion is that only by paying attention to the underlying structure of speech, explicitly, shall we continue to make progress in both speech recognition and speaker classification. Ignoring the dinosaurs whose bones we are examining will only take us so far. Most modern pattern classification approaches deliberately hide this underlying structure inside automatic methods, if it is used at all. We need to expand our approaches and stop focussing on reductionist solutions.

Hollien [78] opens the section on speaker identification in his book by saying:

> "Almost anyone who has normal hearing, and who has lived long enough to read these words, has had the experience of recognizing some unseen speaker (usually someone familiar) solely from listening to his or her voice. It was from this everyday experience that the concept (or is it a myth?) of speaker identification was born."

Very similar remarks could be made about the everyday experience of judging mood, ethnicity, intent, and a host of other factors relevant to speaker classification just from hearing someone speak - whether seen or unseen. Campbell [5, p 1446] refers to "the curse of dimensionality" referring to the problem that automatic feature extraction (as in the speaker identification task) soon causes resources to be overwhelmed unless some kind of statistical model is used to manage and structure the plethora of data. This paper points out some of the avenues and directions towards which research for relevant structure in speaker classification may usefully be directed, and reminds the reader of the importance of experiments with synthetic speech in this quest.

Acknowledgments. The author would like to thank the Natural Sciences and Engineering Research Council of Canada for providing financial support for his research until 1992 under Grant OGP0005261. He would also like to thank the shareholders and principals of Trillium Sound Research Inc. for agreeing to release all claim on the *Gnuspeech* software and other materials related to the work so that they could be made available to the speech research and linguistics communities under a Free Software Foundation "General Public License" (GPL) (see: http://www.gnu.org/copyleft/gpl.html).

References

1. Arkin, W.: When seeing and hearing isn't believing. Washington Post (1999) `http://www.washingtonpost.com/wp-srv/national/dotmil/arkin020199.htm`
2. Shockey, L., Docherty, G., Foulkes, P., Lim, L.: Research Scientist Software Engineering Air Force Research Laboratory. In: foNETiks (A network newsletter for the International Phonetic Association and for the Phonetic Sciences). Lockheed Martin Operations Support (2006)

3. Foulkes, P.: The social life of phonetics and phonology. In: Proc. of the Aberdeen Symposium. University of York, Power Point Presentation (136 slides) (2005) http://www-users.york.ac.uk/~pf11/aberdeen05-webversion.ppt
4. Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Petrovska-Delacrétaz, Reynolds, D.: A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Applied Signal Processing 4, 430–451 (2004), http://www.hindawi.com/GetArticle.aspx?doi=10.1155/S1110865704310024
5. Campbell, J.J.: Speaker recognition: a tutorial. Proc. of the IEEE 85 9, 1437–1462 (1997)
6. Chollet, G., Homayounpour, M.: Neural net approaches to speaker verification: comparison with second-order statistical measures. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95) (1995)
7. Homayounpour, M., Chollet, G.: Discrimination of the voices of twins and siblings for speaker verification. In: Proc. of the Fourth European Conference on Speech Communication and Technology (EUROSPEECH '95), Madrid, pp. 345–348 (1995)
8. Genoud, D., Chollet, G.: Segmental approaches to automatic speaker verification. Digital Signal Processing: a Review Journal (2000)
9. Huang, R., Hansen, J.: Unsupervised Audio Segmentation and Classification for Robust Spoken Document Retrieval. In: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), vol. 1, pp. 741–744. Montreal (2004)
10. Lapidot, I., Guterman, H.: Resolution Limitation in Speakers Clustering and Segmentation Problems. In: Proc. of 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece (2001)
11. Meignier, S., Bonastre, J.F., Magrin-Chagnolleau, I.: Speaker utterances tying among speaker segmented audio documents using hierarchical classification: towards speaker indexing of audio databases. In: Proc. of the International Conference on Spoken Language Processong (ICSLP '02) (2002)
12. Meinedo, H., Neto, J.: A stream-based audio segmentation, classification and clustering preprocessing system for broadcast news using ANN models. In: Proc. of the 9th European Conference on Speech Communication and Technology (Interspeech '05), Lisbon, Portugal (2005)
13. Sanchez-Soto, E., Sigelle, M., Chollet, G.: Graphical Models for Text-Independent Speaker Verification. In: Chollet, G., Esposito, A., Faúndez-Zanuy, M., Marinaro, M. (eds.) Nonlinear Speech Modeling and Applications. LNCS (LNAI), vol. 3445, Springer, Heidelberg (2005)
14. Atal, B.S.: Automatic recognition of speakers from their voices. Proc. of the IEEE 64(4), 460–475 (1976)
15. Rosenberg, A.E.: Automatic speaker verification: a review. Proc. of the IEEE 64(4), 475–487 (1976)
16. Sandmel, S. (ed.): The New English Bible with apocrypha. Oxford study edn. Oxford University Press, New York (1976)
17. Shirt, M.: An Auditory Speaker-Recognition Experiment Comparing the Performance of Trained Phoneticians and Phonetically Naive Listeners. Working Papers in Linguistics 1, 115–117 (1983)
18. Foulkes, P., Barron, A.: Telephone speaker recognition amongst members of a close social network. Forensic Linguistics 7(2), 180–198 (2000)
19. M4-Project: Annual report (2004) http://www.m4project.org/M4-AnnualReport2004/

20. Mc Cowan, I., Gatica-Perez, D., Bengio, S., Moore, D., Bourlard, H.: Towards Computer Understanding of Human Interactions. In: Aarts, E., Collier, R.W., van Loenen, E., de Ruyter, B. (eds.) EUSAI 2003. LNCS, vol. 2875, Springer, Heidelberg (2003)
21. Ginzburg, J.: Dynamics and the semantics of dialogue. In: Seligman, J., Westerståhl, D. (eds.) Logic, Language, and Computation. CSLI, Stanford, CA, pp. 221–237 (1996)
22. Mc Clelland, E.: Familial similarity in voices. In: Proc. of the BAAP Colloquium, University of Glasgow (2000)
23. Shirt, M.: An auditory speaker recognition experiment. In: Proc. of the Institute of Acoustics Conference. vol. 6, pp. 101–104 (1984)
24. Koenig, W., Dunn, H.K., Lacy, L.Y.: The sound spectrograph. Journal of the Acoustic Society of America 18, 19–49 (1946)
25. Cooper, F.S., Liberman, A.M., Borst, J.M.: The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc. of the National Academy of Sciences 37, 318–325 (1951)
26. Lawrence, W.: The synthesis of speech from signals which have a low information rate. In: W., J. (ed.) Communication Theory, pp. 460–469. Butterworth & Co, London (1953)
27. Fant, C.G.M.: Modern instruments and methods for acoustic studies of speech. In: Proc. of the 8th. International Congress of Linguists, Oslo University Press, Oslo (1958)
28. Liberman, A., Ingemann, F., Lisker, L., Delattre, P., Cooper, F.S.: Minimal Rules for Synthesizing Speech. Journal of the Acoustic Society of America 31, 1490–1499 (1959)
29. Holmes, J.N.: Speech synthesis by rule. Language and Speech 7(3), 127–143 (1964)
30. Mattingly, I.G.: Synthesis by rule of prosodic features. Language and Speech 9, 1–13 (1966)
31. Fant, C.: Acoustic Theory of Speech Production. Mouton, The Hague, Netherlands (1960)
32. Holmes, J.: Research on speech synthesis carried out during a visit to the Royal Institute of Technology. Technical Report 20739, GPO Eng. Dept., Stockholm (1961)
33. Shearme, J., Holmes, J.: An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 - formant 2 plane. In: Proc. of the 4th International Congress of Phonetic Sciences, Helsinki (1961)
34. Foulkes, P., Docherty, G.J., Watt, D.J.L.: The emergence of structured variation. University of Pennsylvania Working Papers in Linguistics 7(3), 67–84 (2001)
35. Stoel-Gammon, C.K.W., Buder, E.: Cross-language differences in phonological acquisition: Swedish and American /t/. Phonetica 51, 146–158 (1994)
36. Hill, D., Manzara, L., Taube-Schock, C.R.: Real-time articulatory speech-synthesis-byrules. In: Proc. of the 14th International Conference on Voice Technology Applications of the American Voice Input/Output Society (AVIOS 95), San Jose, CA, - includes samples of synthetic speech, pp. 22–44 (1995) http://pages.cpsc.ucalgary.ca/~hill/papers/avios95
37. Hill, D.: GNUSpeech: Articulatory Speech Synthesis. A GNU project (2003), http://savannah.gnu.org/projects/gnuspeech/
38. Hill, D.: Manual for the “Monet” speech synthesis engine and parameter editor (2002), http://pages.cpsc.ucalgary.ca/~hill/papers/monman/index.html
39. Hill, D.: Synthesizer Tube Resonance Model user manual (2004) http://pages.cpsc.ucalgary.ca/~hill/papers/synthesizer/index.html

40. Fant, C., Pauli, S.: Stockholm, Sweden. In: Proc. of the Stockholm Speech Communication Seminar. KTH, Stockholm (1974)
41. Carré, R.: Distinctive regions in acoustic tubes. Speech production modelling. Journal d'Acoustique 5, 141–159 (1992)
42. Jones, D.: Cambridge, UK. Heffe (1918)
43. Pike, K.: Intonation in American English. University of Michigan Press, Ann Arbor (1945 reprinted 1970)
44. Jassem, W.: Stress in modern English. Bulletin de la Sociëté Linguistique Polonaise XII, 189–194 (1952)
45. Lehiste, I., Peterson, G.: Some basic considerations in the analysis of intonation. Journal of the Acoustic Society of America 33, 4 (1961)
46. Abercrombie, D.: Elements of general phonetics. Edinburgh University Press, Edinburgh (1967)
47. Halliday, M.A.K: A course in spoken English intonation. Oxford University Press, London (1970)
48. Allen, G.D.: The location of rhythmic stress beats in English: An experimental study I and II. Language and Speech 15, 72–100, 179–195 (1972)
49. Ladefoged, P.: A course in phonetics. Harcourt Brace Jovanovic, New York (1975)
50. Pierrehumbert, J.: Synthesizing intonation. Journal of the Acoustic Society of America 70, 4 (1981)
51. Willems, N.J., Collier, R., 't Hart, J.: A synthesis scheme for British English intonation.. Journal of the Acoustic Society of America 84, 1250–1261 (1988)
52. Hill, D., Witten, I., Jassem, W.: Some results from a preliminary study of British English speech rhythm. Technical Report Research Report Number 78/26/5 T2N 1N4, Department of Computer Science, University of Calgary, Alberta, Canada (1978)
53. Jassem, W., Hill, D.R., Witten, I.H.: Isochrony in English speech: its statistical validity and linguistic relevance. In: Gibbon, D. (ed.) Pattern, Process and Function in Discourse Phonology, pp. 203–225. de Gruyter, Berlin (1984)
54. Hill, D.R., Reid, N.A: An experiment on the perception of intonational features. International Journal of Man-Machine Studies 9, 337–347 (1977)
55. Hill, D., Schock, C.R.: Unrestricted text-to-speech revisite: rhythm and intonation. In: Proc. of the International Conference on Spoken Language Processing (ICSLP 92), Banff, Alberta, pp. 1219–1222 (1992)
56. Taube-Schock, C.R.: Synthesizing Intonation for Computer Speech Output. Master's thesis, University of Calgary, Thesis awarded Governor General's Gold Medal (1993)
57. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.: Face Recognition: A Literature Survey. ACM Computing Surveys 399–458 (2003)
58. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces, a surve. Proc. of the IEEE 83, 705–740 (1995)
59. Gurk, H.M., MacDonald, J.: Hearing lips and seeing voices. Nature 264, 746–748 (1976)
60. Kuhl, P.: Infants' perception and representation of speech: development of a new theory. In: Proc. of the International Conference on Spoken Language Processing (ICSLP '92), pp. 449–456 (1992)
61. Müller, C.: Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht. Akademische Verlags-Gesellschaft Aka, Berlin/Amsterdam (2006)

62. Müller, C.: Automatic Recognition of Speakers' Age and Gender on the Basis of Empirical Studies. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech '06 - ICSLP), Pittsburgh, PA (2006)
63. DFKI: Anthropomorphic Interaction Agents (2006) http://www.virtual-human.org
64. Wyvill, B., Hill, D.: Expression control using synthetic speech: The State of the Art in Facial Animation. In: Proc. of the 17st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '90), Dallas, pp. 186–212 (1990)
65. Hill, D.R., Pearce, A., Wyvill, B.L.M.: Animating speech: an automated approach using speech synthesised by rules. The Visual Computer 3, 277–289 (1988)
66. Goleman, D.: Emotional Intelligence: why it can matter more than IQ. Bantam Books, New York/Toronto/London/Sydney/Auckland (1995)
67. Stolten, K., Engstrand, O.: Effects of sex and age in the Arjeplog dialect: a listening test and measurements of preaspiration and VOT. In: Proc. of Fonetik 2002, Stockholm, Speech Technology Laboratory, KTH, Stockholm, pp. 29–32 (2002)
68. Williams, B., Hiller, S.M.: The question of randomness in English foot timin: a control experiment. Journal of Phonetics 22, 423–439 (1994)
69. Eckman, P., Friesen, W.: Unmasking the human face. Consulting Psychologist Press, Palo Alto, California (1975)
70. Eckman, P., Friesen, W.: Manual for the facial action coding system. Consulting Psychologist Press, Palo Alto (1977)
71. Levine, R.V.: The pace of life. American Scientist 78, 450–459 (1990)
72. Chernoff, H.: The use of faces to represent points in a k-dimensional space graphically. Journal of the American Statistical Association 68, 361–368 (1973)
73. Abberton, E., Fourcin, A.J.: Intonation and speaker identification. Language and Speech 21, 305–318 (1978)
74. Brown, B.L., Strong, W.J., Rencher, A.C.: Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. Journal of the Acoustic Society of America 55, 313–318 (1974)
75. Broadbent, D., Ladefoged, P.: Vowel judgements and adaptation level. Proc. of the Royal Society of London 151, 384–399 (1960)
76. Adami, A., Mihaescu, R., Reynolds, D., Godfrey, J.: Modeling prosodic dynamics for speaker recognition. In: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong-Kong (2003)
77. Carlson, R., Granstrom, B., Karlsson, I.: Experiments with voice modelling in speech synthesis. In: Proc. of the ESCA workshop on Speaker Characterization in Speech Technology, Edinburgh, pp. 26–28 (1990)
78. Hollien, H.: The Acoustics of Crime: The new science of forensic phonetics. Plenum Press, New York/London (1990)

Speaker Characteristics

Tanja Schultz

Carnegie Mellon University,
Pittsburgh, PA, USA
tanja@cs.cmu.edu
http://www.cs.cmu.edu/~tanja

Abstract. In this chapter, we give a brief introduction to speech-driven applications in order to motivate *why* it is desirable to automatically recognize particular speaker characteristics from speech. Starting from these applications, we derive *what* kind of characteristics might be useful. After categorizing relevant speaker characteristics, we describe in more detail language, accent, dialect, idiolect, and sociolect. Next, we briefly summarize classification approaches to illustrate *how* these characteristics can be recognized automatically, and conclude with a practical example of a system implementation that performs well on the classification of various speaker characteristics.

Keywords: language-dependent speaker characteristics, automatic speaker classification, real-world applications, multilingual phonetic recognition.

1 Introduction

When we talk to someone face-to-face, we can immediately tell if we met this person before or not. We are extremely fast and accurate when in comes to recognizing and memorizing people, even when they are less familiar or we did not see them for a long time. However, we can do much more than just discriminating familiar from unfamiliar people. Pretty quickly we assess a person's gender, age, native language, emotional or attentional state, and educational or cultural background. This is not too surprising when we consider our heritage, where our survival depends on distinguishing tribe members from enemies, liars from trustworthy people, prey from predators. In modern society we will not outright die from misjudging people, but our social behavior, and often our career and success relies on assessing people and their behavior. We are so accustomed to these skills that human beings who do not have this ability draw a lot of attention [1].

To size up a person, we use visual cues such as general appearance, health conditions, and clothing. The importance of the latter was expressed by the Roman rhetorician Quintilian, who said "vestis virum reddit -clothes make the man". However, humans also heavily rely on auditory cues when characterizing people. When we speak to a person over the phone, we identify a familiar voice. If we do not know the speaker, we still form an impression from the speaker's voice.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 47–74, 2007.

With surprising accuracy we can judge height, sex, and age from a speaker's voice [2], but we also sort speakers along categories, such as idiolect, language, dialect, credibility, confidence, educational background, and much more. Apparently, we classify people based on various characteristics and many of those can be derived from speech alone.

In this issue we classify speakers according to characteristics that are derived solely from their speech, as expressed in the definition of speaker classification to be "the process of assigning a vector of speech features to a discrete speaker class". This definition discriminates speaker classification from other biometrical classification techniques, in which intrinsic characteristics of a person are derived for example from fingerprints, retinal pattern, facial features, or DNA structure. It also differentiates speaker classification from techniques based on artifacts such as badges, business cards, or clothing. As mentioned above, humans are pretty good in this assignment process, however the objective of this chapter focus on an automatic assignment process performed by machines. While we see from the above argumentation that speaker characterization is crucial to our social life, it is not immediately clear which benefits we get from *automatic* speaker characterization performed by machines.

In the remainder of this chapter we will discuss why the classification of speaker characteristics is useful. This will be motivated by examples of real-world applications, which rely on the knowledge of characteristics of its users. We will highlight the most important speaker characteristics, categorize them according to some proposed schemes, and explain how these characteristics can be automatically derived by machines. The chapter concludes with a practical implementation example of a particular classification algorithm and its results.

2 Why? - Applications to Speaker Characteristics

Humans are "wired for speech". This term was coined by Clifford Nass and colleagues [3] and refers to the fact that even though people know that they are dealing with an automated system, if it takes speech as input and delivers speech as output, they treat machines as if they were people - with the same beliefs and prejudices. People behave and speak to the machine as if it were a person, they raise their voice to make the machine better understand, yell at it when they get angry, and say good-bye at the end of a "conversation". At the same time, people infer a certain *personality* from the way the machine talks and the words it uses although it is absurd to assume that the machine has a personality (see also [4]). Nass showed for example that it is crucial for a speech-driven interface to match the emotion in the output to the (expected) emotional state of the user [5], and that users regard a computer voice as more attractive, credible and informative if it matched their own personality [6].

Despite this need for personalized and customized **system output**, the body of research is rather small. This fact has recently been addressed in a special session on Speech Communication [7], and we expect that personalized output will get more attention in the future. In contrast, a large body of work has been dedicated to adapting speech-based systems to better match the expected **user input**. The aspect of personalization and customization has been proven to be highly effective in the context of real-world applications. In the following we will briefly introduce some of the research and highlight those speaker characteristics that turned out to be relevant to the process of adapting applications to the users' *spoken* input. Furthermore, we will describe applications that rely on the recognition of *speaker identity.* This brief overview is divided into work on classical human-computer interaction systems and human-centered or computer mediated human-human communication systems.

2.1 Human-Computer Interaction Systems

Human-Computer Interaction refers to the interaction between people (users) and computers taking place at the speech-driven user interface. Examples of applications are telephone-based services using dialog interfaces, authentication systems that assess the user's identity to perform (remote) banking or business transactions, and access control systems to allow physical entry to facilities or virtual rooms such as a computer network.

Today, many banking and other business transactions are done remotely over the phone or via internet. To avoid misuse it is critical to ensure that the user is who s/he claims to be. **Authentication** is the process of assessing the *identity* of a speaker and checking if it corresponds to the claimed identity. Only if the speaker's identity is verified, access is granted. Most of current authentication systems still use textual information provided by users such as passwords, Social Security Numbers, PINs and TANs. However, as the number of phone- and internet-based services increases, juggling numerous accounts and passwords becomes complicated and cumbersome for the user and the risks of fraud escalate. Performing identity verification based on the user's voice appears to be a possible alternative and therefore, service companies heavily investigate the potential of speaker verification. Different from authentication, the task in **Access Control** is to assess the identity of a speaker and to check if this particular speaker belongs to a group of people that get access to for example physical facilities or virtual rooms such as computer networks and websites. Both system types are based on the speaker characteristic *identity.* An early example of a real-world application was the voice verification system at Texas Instruments that controlled the physical entry into its main computer center [8].

Spoken Dialogs Systems play a major role in modern life, become increasingly pervasive, and provide services in a growing number of domains such as finance [9], travel [10], scheduling [11], tutoring [12], or weather [13]. In order to provide timely and relevant service, the systems need to collect information from the user. Therefore, a service dialog will be faster and more satisfying when such information can be gathered automatically. Hazen and colleagues [14] for example included automatic recognition of speaker *identity* to personalize the system according to pre-collected information from registered users and to prevent unauthorized access to sensitive information.

Most telephone-based services in the U.S. today use some sort of spoken dialog systems to either route calls to the appropriate agent or even handle the complete service by an automatic system. Muthusamy [15] developed a front-end system to the 911 emergency phone line, which automatically assessed the *language* of the speaker to route the call to a native agent. One of the early and successful dialog systems, with wide exposure in the U.S. was the AT&T customer care system "How May I Help You?" developed by Gorin and colleagues [16]. Their studies of vast amounts of recording, logs, and transcriptions, propelled research on dialog systems but also showed that automatic systems fail to predict dialog problems. Batliner and colleagues [17] looked at *emotion* as indicator of "trouble in communication" and developed a call routing system that automatically passes over to human operators when users get angry. Polzin [18] argued that human-computer interfaces should in general be sensitive to users' *emotion*. He created an interface that first detects emotion expressed by the user and then adjusts the prompting, feedback, and dialog flow of the system accordingly. The system prompts sound more apologetic when a user seemed annoyed, and feedback is more explicit when the user's voice indicates frustration. Raux [19] used speaker characteristics such as *agegroup* and *nativeness* to tailor the system output to elderly and non-native users with limited abilities in English to make the speech output more understandable. Nass [3] found that people infer a certain *personality* from the way the machine talks and have prejudices about *gender*, regional *dialects* or foreign *accents*, *geographical background*, and *race*. It is expected that these human factors will be taken into account in future systems.

Computer-aided Learning and Assessment tools are another example of human-computer interaction applications. Speech input functionality is particularly desirable in the context of language learning [20]. Advanced systems provide interactive recording and playback of user's input speech, feedback regarding acoustic speech features, recognizing the input, and interpreting interaction to act as a conversation partner. Especially the latter three functionalities are very challenging due to the naturally broad range of *accent* and *fluency* of its users. Learning systems are usually customized to the *native language L1* of the language learner to overcome robustness issues [21], but may have to be tailored towards particular *dialects*, especially in countries of diglossia. Automatic assessment of *proficiency level* is deemed important, particularly in the light of strong imbalance between number of learners and number of teachers, see for example the E-Language Learning System program between the U.S. Department of Education and the Chinese Ministry of Education [20].

New challenges arise when applications are brought to the **developing world** to users with limited access, exposure, and with a different cultural basis for understanding. Barnard and colleagues built a telephone-based service in rural South-Africa [22]. Some of their findings are surprising and not foreseen, such as the request for louder prompts (due to collectivsm bystanders who share the conversation) and the fact that silence after prompt does not elicit an answer due to uncertainty avoidance in this *cultural background*. The last example emphasizes that many aspects of speech-driven systems have not been fully understood or investigated. We expect that with the increasing application of these systems, the research on automatic classification of speaker characteristics will be intensified to make systems more useful for a large population of users.

2.2 Human-Centered Systems

Human-Centered Systems refer to computer services that are delivered in an implicit, indirect, and unobstrusive way to people whose primary goal is to interact with other people. Computers stay in the background - like electronic butlers - attempting to anticipate and serve people's needs. Thus, computers are introduced into a loop of humans interacting with humans, rather than condemning a human to operate in a loop of computers (see CHIL - Computers in the Human Interaction Loop [23]).

Emerging computer services are **Smart Room Environments** [24], in which computers watch and interpret people's actions and interactions in order to support communication goals. One implementation example is an automatic meeting support system, which tracks what was said, who said it, to whom, and how it was said [25]. By annotating speech recognition output with the speakers' *identity*, *attentional state*, and *emotional state*, the meeting notes can be properly indexed, skimmed, searched, and retrieved. Infrastructures such as socially-supportive workspaces [23] or augmented multiparty interactions [26] foster cooperation among meeting participants, including multimodal interface to enter and manipulate participants' contributions, and facilitator functionalities that monitor group activities. Other services implemented within the framework of CHIL [23] include better ways of connecting people and supporting human memory. For all of these services, computers need to automatically gather context- and content-aware information such as *topic*, *meeting type*, or environmental conditions, and participant characteristics such as *attentional state*.

An example of computer-mediated applications that support human-to-human communication is **Speech Translation** [27,28,29]. The task of speech translation is to recognize incoming speech from the source language, to translate the text of the recognizer output into text of the target language, and then synthesize the translated text to audible speech in the target language. Most applications are designed as two parallel one-directional systems, some systems perform automatic *language* identification to route the speech into the corresponding system [30]. Ideally, the translation should not only preserve the original meaning of the spoken input, but also reflect other aspects of the message such as level of politeness, respect, directness, or wittiness. Some of these aspects might be directly derived from speaker characteristics, such as the generation of appropriate synthesized output based on the speaker's *gender*, or based on the identification of the *emotional state* of a speaker in order to interpret emotional cues and wittiness. Beyond this, some aspects require knowledge about the *relationship between the speaker and the listener*. In some languages, the word usage changes significantly depending on the hierarchy between sender and receiver, and using the wrong form may offend the receiver. Japanese is such an example, where Dr. Sadaoki Tomuko would be addresses as Tomuko-san if he is a close friend or Tomuko-sensei if he is the boss of the sender. To address this problem, the English-Japanese JANUS translation system [31] was designed to switch between politeness levels.

2.3 Adaptation of System Components

As described above, the classification of speaker characteristics plays a crucial role in customization and personalization of applications. Beyond that, speaker characteristics need to be assessed in order to adapt system components, particularly the speech recognition front-end to the specific voice characteristics of the speaker and the content of what was spoken. This adaptation process has been proven to dramatically improve the recognition accuracy, which usually carries over favorably to the performance of the overall system.

Adaptation of speech recognition is traditionally mostly concerned with the adaptation of the acoustic and language model. In early days the acoustic model adaptation was performed by an enrollment procedure that asked the user to reading text prompts. This method might be quite helpful to power users of the system and allows to store and pre-load speaker-specific acoustic models. However, this enrollment procedure is time consuming. Therefore, more recent systems rely on speaker adaptive training methods, which first determine the speaker's *identity* and then apply acoustic model adaptation based on the assumed identity. Some applications rely on broader speaker classes such as *gender* or *agegroup* to load pre-trained models [32]. For the purpose of dictionary and language model adaptation, the *topic* or the *content* of the spoken input is analyzed and used for adaptation [33]. Beside the speech recognition front-end, other dialog components may benefit from this technique as well, by modeling various *dialog states*, or detecting *keywords* to trigger state switches.

Code switching, i.e. switching the language between utterances, can not be handled by monolingual speech recognition systems. Efforts have been made to develop multilingual speech recognition system [34], but so far it looks favorable to design dedicated language identification modules that direct the speech input to the appropriate monolingual recognition system [30]. Idiolect has shown to have a significant influence on speaker recognition [35] and accent is particularly known to have a detrimental effect on speech recognition performance. Consequently, much effort has been put into the classification of these characteristics and the appropriate adaptation of system components. For an overview, we refer the reader to [36].

2.4 Summary

We conclude this section with a table that summarizes those speaker characteristics, which are most relevant to human-computer and human-centered applications. In addition, it gives references to implementation examples, or studies thereof. Some of the referenced applications are not covered in this section, as they are described in large detail elsewhere in this issue. Among those are

Forensic applications, where the characteristics *gender*, *age*, *medical conditions*, *dialect*, *accent*, and *sociolect* play a pivotal role. An overview of forensic applications is provided by Jessen in this issue [37]. Furthermore, we did not discuss emerging applications for home parole, detection of deception, or fraud in the context of **Law Enforcement**, which are concerned with speaker's *identity* or *emotion.* An introduction to this field concerning the latter characteristic is given by Eriksson in this issue [38].

Table 1. Speaker Characteristics and Applications

Characteristic	Applications, Reference			
identity	Transaction Authentication	[39];	Access Control	[8]
	Dialog Systems	[14];	Meeting Browser	[25]
gender	Dialog Systems	[32];	Speech Synthesis	[3]
	Forensics	[37]		
age	Dialog Systems	[32];	Forensics	[37]
	Speech Synthesis	[19]		
health	Forensics	[37]		
language	Call Routing	[15];	Speech Translation	[30]
dialect	Forensics	[37]		
accent	Language Learning	[21];	Dialog Systems	
	Speech Synthesis	[19];	Forensics	[37]
	Assessment Systems	[20]		
sociolect	Forensics	[37]		
idiolect	Speaker Recognition	[35];	Forensics	[37]
emotional state	Speech Translation	[40];	Meeting Browser	[25]
	Law Enforcement	[38];	Dialog Systems	[18,17]
attentional state	Human-Robot Interaction	[41];	Smart Workspaces	[26,23,24]
relationship/role	Speech Translation	[31]		
cultural background	Dialog Systems	[22]		

3 What? A Taxonomy of Speaker Characteristics

The discrete speaker classes, to which vectors of speech features are assigned, characterize a speaker. We impose here a hierarchical structure on those characteristics, which we consider to be relevant to speech-based applications as described above.

Figure 1 shows the propose taxonomy, distinguishing first and foremost between physiological and psychological aspects of speaker characteristics. The latter ones are further divided into aspects which concern the individual speaker versus those that concern a speaker in a particular community or collective. For example, a speaker may be in the role of a professor for the students at university, a wife to her husband at home, or a mother to her child. The authority of a speaker may vary with the context he or she is talking about, the hierarchy depends on whom s/he talks to, the credibility may depend on whom s/he is doing

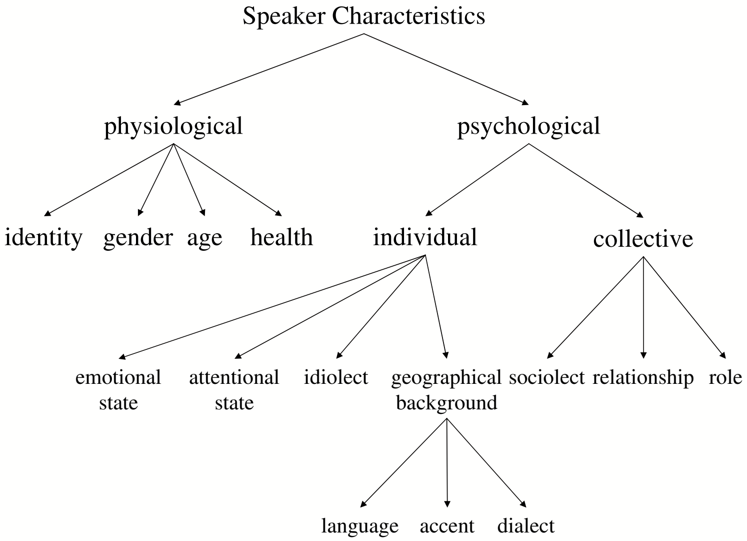


Fig. 1. Taxonomy of Speaker Characteristics

business with, and so on. That is, the category ”collective” requires a definition of a relation between sender and receiver.

This taxonomy is not without limitations, for example it does not cover all aspects of an individual (e.g. weight, height, smoking or drinking habits, demographics such as race, income, mobility, employment status) or special aspects such as speech pathologies, but rather focus on those characteristics we consider to be relevant (and assessable) in the context of typical speech applications.

Furthermore, the taxonomy does not indicate, which level of **linguistic information** are necessary to discriminate between characteristics. For example, low level acoustic features are usually sufficient to discriminate gender; phonetic, phonologic, and lexical knowledge might be required to discriminate idiolects, while it needs semantic and syntactic information to differentiate sociolects. Even pragmatics might be necessary to derive the role of speakers and their relationship to a collective. While low level physical aspects are relatively easy to automatically extract, high level cues are difficult to assess. As a consequence most automatic systems for speaker recognition still concentrate on the low-level cues.

Another aspect, which is not reflected in the taxonomy is the discrimination between **stable versus transient** characteristics. Examples for stable characteristics are speaker identity and gender. Transient characteristics change over time. This aspect may play an important role for practical applications, especially if a characteristic underlies dynamic changes over the duration of a single audio recording session. While locally stable characteristics such as age, health, language, accent, dialect, and idiolect may change very slowly compared to the duration of a recording session, characteristics such as attentional and emotional state of a speaker, as well as the context or topic change dynamically. Also, the relationship of a speaker to the listener may change over the course of an interaction. Other characteristics such as sociolect may depend on the collective. Idiolect, accent and dialect are functions of the spoken language, but are usually rather stable within the same language. Therefore, if a speaker switches languages within one recording session, the class assignments for idiolect, accent and dialect usually switch along.

3.1 Language-Dependent Speaker Characteristics

In the following we subsume the five characteristics language, accent, dialect, idiolect, and sociolect under the term **language-dependent** speaker characteristics as they are somewhat dependent on the actual language spoken by the speaker.

Drawing the line between genuinely different languages and dialects of the same language is a subject of various disputes. We define a **dialect** as a regional variant of a language that involves modifications at the lexical and grammatical level. In contrast **accent** is a regional variant affecting only the pronunciation, mostly phonetic realizations but also prosody, allophonic distribution, and fluency. British Received Pronunciation for example is an accent of English, whereas Scottish English would be considered a dialect since it often exhibits grammatical differences, such as ”Are ye no going?” for ”Aren't you going?” (see [42]). Dialects of the same language are assumed to be mutually intelligible, while different **languages** are not, i.e. languages need to be explicitly learned by speakers of other languages. In addition, languages have a distinct literary tradition, while dialects are primarily spoken varieties without literary tradition.

These definitions are greatly simplified. Many languages lack a writing system and thus do not have any literary tradition. Also, the distinction between languages and dialects is a continuum rather than a binary decision, and often motivated by sociopolitical rather than linguistic considerations. Chinese languages, for example are unified by a common writing system but have a large number of mutually unintelligible varieties that differ substantially in pronunciation, vocabulary, and grammar. While most linguists would argue that these variations are different languages, they are officially labeled as dialects to promote the concept of Chinese national unity (see [42]). The exact opposite happened for Serbo-Croatian, the official language of former Yugoslavia. After the breakup, the languages Croatian and Serbian became to be described as separate languages to emphasize national independence.

Apart from regional variations, languages exhibit idiolectal and sociolectal variation. The term **idiolect** describes consistent speech patterns in pronunciation, lexical choice, or grammar that are specific to a particular speaker. Idiolectal patterns may include speaker-specific recurrent phrases (e.g. a tendency to start sentences with *Well, to be honest...*), characteristic intonation patterns, or divergent pronunciations (e.g. *nucular* instead of *nuclear*) (see [42]). A **sociolect** is a set of variations that are characteristic of a group of speakers defined not by regional cohesion but by social parameters, such as economic status, age, profession, etc. Since dialects often have a particular social status, some variants may be considered simultaneously a dialect and a sociolect. For example, standard German has close similarities to dialects spoken in Hannover and the state of Saxony-Anhalt, the latter being the origin of Martin Luther whose bible translation formed the basis for the development of standard German. Thus, while being a dialect in these particular areas, standard German is also a sociolect in that it carries a certain prestige from being the national language of Germany, used throughout the country in broadcast, press, and by people of higher education.

Despite significant efforts to make speech recognition systems robust for real-world applications, the problem of regional variations remains to be a significant challenge. Word error rates increase significantly in the presence of non-native [43,44] and dialectal speech [45]. One of the main reasons for this performance degradation is that acoustic models and pronunciation dictionaries are tailored toward native speakers and lack the variety resulting from non-native pronunciations. In addition, the lexicon and language model lack the dialectal variety. The straight-forward solution of deploying dialect- or accent-specific speech recognizers is prohibited by two practical limitations: lack of platform resources and lack of data. Particularly embedded environments such as mobile or automotive applications limit the integration of multiple recognizers within one system. Even if resources permit the deployment of dialect or accent specific systems, the variety usually leads to very limited data resources. As a consequence real-world applications require cross-dialect or non-native recognition. The reader is referred to [36] for a comprehensive introduction into this area. Idiolectal features can be used for tailoring a speech application to a specific user, for instance in training a speech-based automated office assistant. In addition, idiolectal features have been shown to be helpful in automatic speaker identification [35]. Similarly, sociolectal features can be taken into account when developing an application for an entire user group.

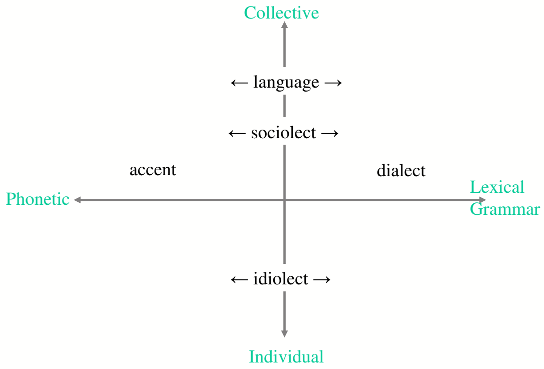


Fig. 2. Language-dependent Characteristics

Multilingual settings may impact idiolectal and sociolectal variations, for example [46] found evidence that bilingual speakers change their L1 speech after spending time in L2-speaking environment. Several techniques to improve speech recognition performance in the presence of **code-switching** have been investigated [47,48]. Code-switching refers to the act of using words or phrases from different languages in one sentence, a typical behavior of multilingual speakers engaged in informal conversations.

Figure 2 summarizes the similarities and differences among the language-dependent characteristics language, dialect, accent, idiolect, and sociolect. Main discriminating factors are the effects on linguistic aspects and whether these characteristics apply to individuals or a collective.

4 How? - Automatic Classification of Speaker Characteristics

Probably the most extensively studied and prominent tasks that investigate the "assignment of speech features to discrete speaker classes" are speaker recognition (who is speaking, class=identity) and language identification (which language is spoken, class=language). Speech recognition (what is said, class=content) tackles a much broader problem but could be viewed as part of "Speaker Classification" when high-level characteristics, such as content, topic, or role are investigated. Recently, the three tasks grow closer together, as it becomes evident that solutions to one task may benefit the performance of the other, and that all of them need to be studied in order to improve speech-based real-world applications. In the following we will briefly survey language identification and speaker recognition. This section it not meant to give a comprehensive introduction, for more details the reader is referred to in-depth overviews, such as [49] for language identification and [39,50] for speaker recognition. A good introduction into speech recognition can be found in [51].

4.1 Speaker Recognition

Classification approaches can be discriminated by the level of linguistic knowledge applied to the solution of the classification task. Reynolds defines a hierarchy of perceptual cues that humans apply for the purpose of recognizing speakers [39]. On the highest level, people use semantics, diction, idiolect, pronunciation and ideosynchrasies, which emerge from the socio-economic status, education, and place of birth of a speaker. On the second level are features such as prosodic, rhythm, speed, intonation, and volume of modulation, which discriminate personality and parental influence of a speaker. On the lowest linguistic level people use acoustic aspects of sounds, such as nasality, breathiness or roughness, which allow to draw conclusions about the anatomical structure of the speaker's vocal apparatus. While low level physical aspects are relatively easy to extract automatically, high level cues are difficult to assess. As a consequence most automatic systems for speaker recognition still concentrate on the low-level cues.

Conventional systems apply Gaussian Mixture Models (GMM) to capture frame-level characteristics [52]. Since the speech frames are assumed to be independent from each other, GMMs often fail to discriminate speaker-specific information that evolves over more than one frame. Therefore, GMMs are poorly suited for discriminating speakers based on higher-level differences, such as idiolect. Furthermore, GMMs are found to be challenged by mismatching acoustic conditions as they solely rely on low-level speech-signal features. To overcome these problems, speaker recognition recently focus on including higher-level linguistic features, such as phonetic information emerging from speaker ideosynchrasies [35]. This area is called phonetic speaker recognition and applies relative frequencies from phone n-grams [53]. This approach is currently intensively studied [39] and extended by different modeling strategies, variations of statistical n-gram models [54], variations of classifiers like Support Vector Machines [55], and modeling of cross-stream dimensions to discover underlying phone dependencies across multiple languages [54,56].

4.2 Language Identification

Similar to speaker recognition, language identification approaches can be categorized by the level of linguistic information, which is applied to the classification task. [49] discriminates the signal processing level, the unit level (e.g. phones), the word level, and the sentence level. According to these levels, he distinguishes between acoustic approaches to language identification that apply spectral features derived from speech segments [57], phonotactic approaches, which use the contraints of relative frequencies of sound units [58], along with various derivatives using multilingual phone recognizers as tokenizer [59], extended n-grams [60], cross-stream modeling [61], and combinations of GMMs and phonotactic models [62]. Furthermore, Navrátil [49] lists prosodic approaches, which use tone, intonation, and prominence [63], and those approaches that apply full speech recognizers to language identification [64].

5 A Classification System for Speaker Characteristics

In this section we present a general classification system, which applies one common framework to the classification of various speaker characteristics, namely identity, gender, language, accent, proficiency level, and attentional state of a speaker. The framework uses high-level phonetic information to capture speakers' ideosynchrasies, as initially proposed by [58] in the context of language identification and [35] in the context of speaker recognition. The basic idea is to decode speech by various phone recognizers and to use the relative frequencies of phone n-grams as features for training speaker characteristic models and for their classification. We enrich existing algorithms by applying the approach to various speaker characteristics, by using a larger number of language independent phone recognizers, and by modeling dependencies across multiple phone streams [54]. Furthermore, we investigate different decision rules, study the impact of the number of languages involved, and examine multilingual versus multi-engine approaches with respect to classification performance.

5.1 Multilingual Phone Sequences

Our experiments were conducted using phone recognizers of the GlobalPhone project [65] available in 12 languages Arabic (AR), Mandarin Chinese (CH), Croatian (KR), German (DE), French (FR), Japanese (JA), Korean (KO), Portuguese (PO), Russian (RU), Spanish (SP), Swedish (SW), and Turkish (TU). These phone recognizers were trained using the Janus Speech Recognition Toolkit. The acoustic model consists of a context-independent 3-state HMM system with 16 Gaussians per state. The Gaussians are based on 13 Mel-scale cepstral coefficients and power, with first and second order derivatives. Following cepstral mean subtraction, linear discriminant analysis reduces the input vector to 16 dimensions. Training includes vocal tract length normalization (VTLN) for speaker normalization. Decoding applies unsupervised MLLR to find the best matching warp factor for the test speaker. Decoding is performed with Viterbi search using a fully connected null-grammar network of mono-phones, i.e. no prior knowledge about phone statistics is used for the recognition process. Figure 3 shows the correlation between number of phone units and phone error rates for ten languages.

To train a model for a particular speaker characteristic, a language dependent phonetic n-gram model is generated based on the available training data. In our experiments we train phonetic bigram models created from the CMU-Cambridge Statistical Language Model Toolkit [19]. All phonetic bigram models are directly estimated from the data, rather than applying universal background models or adaptation with background models. No transcriptions of speech data are required at any step of model training. Figure 4 shows the procedure of training for a speaker identity model for speaker k. Each of the m phone recognizers $(PR_1, \ldots, PR_m)$ decode the training data of speaker k to produce m phone strings. Based on these phone strings m phonetic bigram models $(PM_{1,k}, \ldots, PM_{m,k})$ are estimated for speaker k. Therefore, if an audio segment needs to be classified into one of an n-class speaker characteristic, the m phone recognizers will produce $m \times n$ phonetic bigram models.

During classification, each of the m phone recognizers PR_i, as used for phonetic bigram model training, decodes the test audio segment. Each of the resulting m phone strings is scored against each of n bigram models $PM_{i,j}$. This results in a perplexity matrix PP, whose $PP_{i,j}$ element is the perplexity produced by phonetic bigram model $PM_{i,j}$ on the phone string output of phone recognizer PR_i. While we will explore some alternatives in later experiments, our default decision algorithm is to propose a class estimate C_j^* by selecting the lowest $\sum_i (PP)_{i,j}$. Figure 5 depicts this procedure, which we refer to as MPM-pp.

In the following we apply the described MPM-pp classification approach to a variety of classification tasks in the context of speaker characteristics, namely

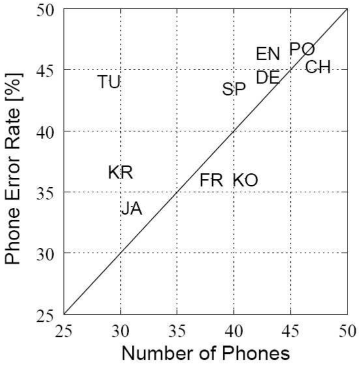


Fig. 3. Error rate vs number of phones for ten GlobalPhone languages

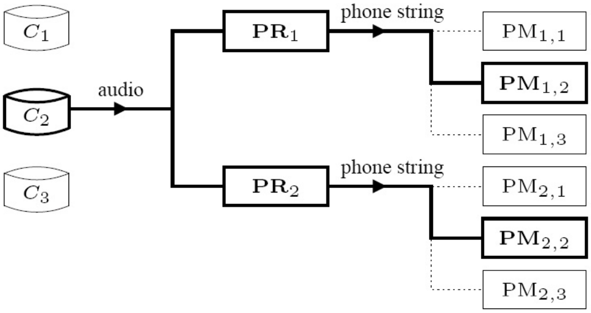


Fig. 4. Training of feature-specific phonetic models for 2 phone recognizers and a 3–class problem

to the classification of identity, gender, accent, proficiency level, language, and attentional state of a speaker.

5.2 Classification of Speaker Identity

In order to investigate robust speaker identification (SID) under far-field conditions, a distant-microphone database containing speech recorded from various microphone distances had been collected at the Interactive Systems Laboratory. The database contains 30 native English speakers reading different articles. Each of the five sessions per speaker are recorded using eight microphones in parallel: one close-speaking microphone (Dis 0), one lapel microphone (Dis L) worn by the speaker, and six other lapel microphones at distances of 1, 2, 4, 5, 6, and 8 feet from the speaker. About 7 minutes of spoken speech (approximately 5000 phones) is used for training phonetic bigram models.

Table 2 lists the identification results of each phone recognizer and the combination results for eight language phone recognizers for Dis 0 under matching

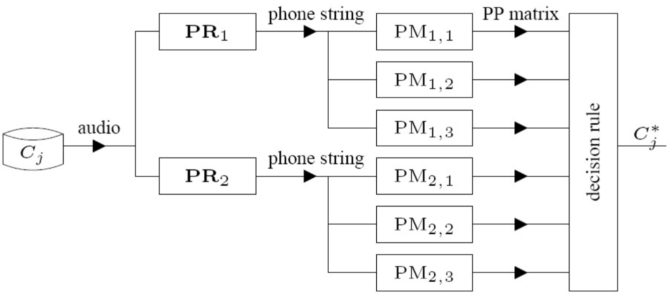


Fig. 5. MPM-pp classification block diagram

conditions. It shows that multiple languages compensate for poor performance on single engines, an effect which becomes even more prominent for short test utterances.

Table 3 compares the identification results for all distances on different test utterance lengths under matched and mismatched conditions, respectively. Under matched conditions, training and testing data are from the same distance. Under mismatched conditions, we do not know the test segment distance; we make use of all $p = 8$ sets of $PM_{i,j}$ phonetic bigram models, where p is the number of distances, and modify our decision rule to estimate $C_j^* = \min_j (\min_k \sum_i \mathrm{PM}_{i,j,k})$, where i is the index over phone recognizers, j is the index over speaker phonetic models, and $1 \leq k \leq p$. The results indicate that MPM-pp performs similar under matched and mismatched conditions. This compares quite favorable to the traditional Gaussian Mixture Model approach, which significantly degrades under mismatching conditions [66]. By applying higher-level information derived from phonetics rather than solely from acoustics, we believe to better cover speaker idiosyncrasies and accent-specific pronunciations. Since this information is provided from complementary phone recognizers, we anticipate greater robustness, which is confirmed by our results.

Table 2. MPM-pp SID rate on varying test lengths at Dis 0

Language	60 sec	40 sec	10 sec	5 sec	3 sec
CH	100	100	56.7	40.0	26.7
DE	80.0	76.7	50.0	33.3	26.7
FR	70.0	56.7	46.7	16.7	13.3
JA	30.0	30.0	36.7	26.7	16.7
KR	40.0	33.3	30.0	26.7	36.7
PO	76.7	66.7	33.3	20.0	10.0
SP	70.0	56.7	30.0	20.0	16.7
TU	53.3	50.0	30.0	16.7	20.0
Fusion	**96.7**	**96.7**	**96.7**	**93.3**	**80.0**

Table 3. MPM-pp classification accuracy on varying test lengths under matched (left-hand) and mismatched (right-hand) conditions

	Matched Conditions				Mismatched Conditions			
Test Length	60s	40s	10s	5s	60s	40s	10s	5s
Dis 0	96.7	96.7	96.7	93.3	96.7	96.7	96.7	90.0
Dis L	96.7	96.7	86.7	70.0	96.7	100	90.0	66.7
Dis 1	90.0	90.0	76.6	70.0	93.3	93.3	80.0	70.0
Dis 2	96.7	96.7	93.3	83.3	96.7	96.7	86.7	80.0
Dis 4	96.7	93.3	80.0	76.7	96.7	96.7	93.3	80.0
Dis 5	93.3	93.3	90.0	76.7	93.3	93.3	86.7	70.0
Dis 6	83.3	86.7	83.3	80.0	93.3	86.7	83.3	60.0
Dis 8	93.3	93.3	86.7	66.7	93.3	93.3	86.7	70.0

5.3 Classification of Gender

The NIST 1999 speaker recognition evaluation set [67] with a total of 309 female and 230 male speakers was applied to gender identification experiments [56]. For each speaker, two minutes of telephone speech were used for training and one minute of unknown channel type for testing. Experiments were conducted on the MPM-pp approach. In addition, a different decision rule, MPM-ds was investigated. For the MPM-ds approach the perplexity was replaced by a decoding score, i.e. the negative log probability distance score. For decoding, the equal-probability phonetic bigram models were replaced by language-specific models, resulting from training bigram phonetic models for each of the phone recognizers and each gender category. For classification, each phone recognizer applied the language-specific model. While the MPM-pp approach requires to only decode with m recognizers, the MPM-ds approach requires to run $m \times n$ recognition processes, where m refers to the number of phone recognizers and n to the number of classes to be discriminated. Furthermore, the MPM-ds approach heavily depends on reliable probability estimates from the phonetic models. However, the amount of data available for gender classification was assumed to be sufficient for this task. For testing, 200 test trials from 100 men and 100 women were randomly chosen. Table 4 compares the results of the MPM-pp with the MPM-ds decision rule. Both approaches achieved a 94.0% gender classification accuracy, which indicates that comparable results can be achieved when enough data for training is available. Earlier experiments on speaker identification showed that MPM-pp clearly outperforms MPM-ds, most likely due to the lack of training data for a reliable estimate of phonetic models [56].

5.4 Classification of Accent

In the following experiments we used the MPM-pp approach to differentiate between native and non-native speakers of English. Native speakers of Japanese with varying English proficiency levels make up the non-native speaker set. Each

Table 4. Comparison between MPM-pp and MPM-ds on gender classification

	CH	DE	FR	JA	KR	PO	SP	TU	ALL
MPM-pp	88.5	89.5	89.0	86.5	87.5	89.0	92.0	90.0	94.0
MPM-ds	89.5	88.5	91.0	89.0	88.0	91.5	92.0	89.0	94.0

Table 5. Number of speakers, utterances, and audio length for native and non-native classes

	n_{spk}		n_{utt}		τ_{utt}	
	native	non-native	native	non-native	native	non-native
training	3	7	318	680	23.1 min	83.9 min
testing	2	5	93	210	7.1 min	33.8 min

speaker read several news articles, training and testing sets are disjoint with respect to articles as well as speakers. The acquisition of the database is described in detail in [68]. The data used for the experiments are summarized in Table 5.

In two sets of experiments, we first employ 6 of the above described GlobalPhone phone recognizers $\text{PR}_\text{i} \in \{\text{DE}, \text{FR}, \text{JA}, \text{KR}, \text{PO}, \text{SP}\}$ [69] and then augment these by a seventh language $\{\text{CH}\}$ to study differences resulting from the added language [70]. During classification of non-native versus native speakers, the 7×2 phonetic bigram models produce a perplexity matrix for the test utterance to which we apply the lowest average perplexity decision rule. On our evaluation set of 303 utterances for 2–way classification between native and non–native utterances, the classification accuracy improves from 93.7% using models in 6 languages to 97.7% using models in 7 languages. An examination of the average perplexity of each class of phonetic bigram models over all test utterances reveals the improved separability of the classes, as shown in Table 6. The average perplexity of non-native models on non-native data is lower than the perplexity of native models on that data, and the discrepancy between these numbers grows after adding training data decoded in an additional language.

Table 6. Average perplexities for native and non-native classes using 6 versus 7 phone recognizers

Phonetic model	6 languages		7 languages	
	non- native	native	non-native	native
non-native	29.1	31.7	28.9	34.1
native	32.5	28.5	32.8	31.1

5.5 Classification of Proficiency Level

We apply the MPM-pp approach to classify utterances from non-native speakers according to assigned speaker proficiency classes using the same data as in the accent classification task. The original non-native data had been labeled with the proficiency of each speaker on the basis of a standardized evaluation procedure conducted by trained proficiency raters [68]. All speakers received a floating point grade between 0 and 4, with a grade of 4 reserved for native speakers. The distribution of non-native training speaker proficiencies showed that they fall into roughly three groups. We created three corresponding classes for the attempt to classify non-native speakers according to their proficiency. Class 1 represents the lowest proficiency speakers, class 2 contains intermediate speakers, and class 3 contains the high proficiency speakers. The phonetic bigram models are trained as before, with models in 7 languages and 3 proficiency classes. Profiles of the testing and training data for these experiments are shown in Table 7.

Table 7. Number of speakers, utterances, audio length, and average speaker proficiency score per proficiency class (C-1 to C-3)

	n_{spk}			n_{utt}			τ_{utt} (min)			ave. prof		
	C-1	C-2	C-3	C-1	C-2	C-3	C-1	C-2	C-3	C-1	C-2	C-3
training	3	12	4	146	564	373	23.9	82.5	40.4	1.33	2.00	2.89
testing	1	5	1	78	477	124	13.8	59.0	13.5	1.33	2.00	2.89

Similar to the experiments in accent identification, we compared the application of 6 versus 7 phone recognizers. As the confusion matrix in Table 8 indicates, the addition of one language leaves to small improvement over our results using models in 6 languages. It reveals that the phonetic bigram models trained in Chinese cause the system to correctly identify more of the class 2 utterances at the expense of some class 3 utterances, which are identified as class 2 by the new system. Our results indicate that discriminating among proficiency levels is a more difficult problem than discriminating between native and non-native speakers. The 2–way classification between class 1 and class 3 gives 84% accuracy, but classification accuracy in the 3–way proficiency classification approach achieves 59% in the 6-language experiment and 61% using the additional seventh phone recognizer.

5.6 Classification of Language

In this section, we apply the MPM-pp framework to the problem of multi-classification of four languages: Japanese (JA), Russian (RU), Spanish (SP) and Turkish (TU). We elected to use a small number of phone recognizers in languages other than the four classification languages in order to duplicate the circumstances common to our identification experiments, and to demonstrate a degree of language independence which holds even in the language identification

Table 8. Confusion matrix for 3-way proficiency classification using 6 versus 7 phone recognizers

Phonetic model	6 languages			7 languages		
	C-1	C-2	C-3	C-1	C-2	C-3
C-1	8	3	19	8	5	17
C-2	8	41	61	6	53	51
C-3	2	12	99	1	20	92

domain. Phone recognizers in Chinese (CH), German (DE) and French (FR), with phone vocabulary sizes of 145, 47 and 42, respectively, were borrowed from the GlobalPhone project. The data for this classification experiment, were also borrowed from the GlobalPhone project but not used in training the phone recognizers. It was divided up as shown in Table 9. Data set 1 was used for training the phonetic models, while data set 4 was completely held-out during training and used to evaluate the end-to-end performance of the complete classifier. Data sets 2 and 3 were used as development sets while experimenting with different decision strategies.

Table 9. Number of speakers, utterances, and audio length per language

	Set	JA	RU	SP	TU
n_{spk}	1	20	20	20	20
	2	5	10	9	10
	3	3	5	5	5
	4	3	5	4	5
$\sum n_{\text{utt}}$	all	2294	4923	2724	2924
$\sum \tau_{\text{utt}}$	all	6 hrs	9 hrs	8 hrs	7 hrs

For training the phonetic bigram models, utterances from set 1 in each $L_j \in \{\text{JA}, \text{RU}, \text{SP}, \text{TU}\}$ were decoded using each of the three phone recognizers $\text{PR}_i \in \{\text{CH}, \text{DE}, \text{FR}\}$. 12 separate trigram models were constructed with Kneser/Ney backoff and no explicit cut-off. The training corpora ranged in size from 140K to 250K tokens. Trigram coverage for all 12 models fell between 73% to 95%, with unigram coverage below 1%.

We first benchmarked accuracy using our lowest average perplexity decision rule. For comparison, we constructed a separate 4-class multi-classifier, using data set 2, for each of the four durations $\tau_k \in \{5\text{s}, 10\text{s}, 20\text{s}, 30\text{s}\}$; data set 3 was used for cross-validation.

Our multi-classifier combined the output of multiple binary classifiers using error-correcting output coding (ECOC). A class space of 4 language classes induces 7 unique binary partitions. For each of these, we trained an independent multilayer perceptron (MLP) with 12 input units and 1 output unit using scaled conjugate gradients on data set 2 and early stopping using the cross-validation data set 3. In preliminary tests, we found that 25 hidden units provide adequate performance and generalization when used with early stopping. The output of all 7 binary classifiers was concatenated together to form a 7-bit code, which in the flavor of ECOC, was compared to our four class codewords to yield a best class estimate. Based on total error using the best training set weights and cross-validation set weights on the cross-validation data, we additionally discarded those binary classifiers which contributed to total error; these classifiers represent difficult partitions of the data.

With phone recognizers drawn from the baseline set, classification accuracy using lowest average perplexity led to 94.01%, 97.57%, 98.96% and 99.31% accuracy on 5s, 10s, 20s and 30s data respectively, while with ECOC/MLP classification accuracy improved to 95.41%, 98.33%, 99.36% and 99.89% respectively.

5.7 Classification of Attentional State

The following experiments investigate the power of the MPM-pp approach to identify the attentional state of a speaker. More particularly, we aim to discriminate the interaction of two human beings from the interaction of one human with a robot. The data collection took place at the Interactive Systems Labs and mimics the interaction between two humans and one robot. One person, acting as the host, introduces the other person, acting as a guest, to the new household robot. Parallel recordings of audio and video focus on the host to determine if the host addresses the guest or the robot. In order to provoke a challenging scenario, the speakers were given instructions to imagine that they introduce the new household robot to the guest by explaining the various skills of the robot, for example to bring drinks, adjust the light, vacuum the house, and so on. 18 recording sessions of roughly 10 min length each were collected and manually transcribed. All utterances were tagged as command, when the robot was addressed or as conversation, when the guest was addressed. 8 sessions were used for training, 5 for development, and the remaining 5 for evaluation [41].

We compare the MPM-pp approach to a speech-based approach that applies a combination of higher-level speech features, such as sentence length (assuming that commands to a robot are shorter than conversations with another human), topic occurrence (assuming that commands are more likely to contain the word "robot"), number of imperatives (assuming that commands are rather formulated in imperative form), and perplexity calculation based on a "command" language model and a "conversation" language model (assuming that commands give lower perplexity on the former language model and conversations give lower on the latter). The results from this selection are labeled as "Feature Combi". The MPM-pp approach features the above described 12 GlobalPhone recognizers.

The results in Table 10 shows F-measure and classification accuracy. The calculation of the F-measure is based on the assumption that it is more important to detect when the robot was addressed. The results indicate that the MPM-pp approach slightly outperforms the combination of higher-level speech features, which is somewhat surprising given the amount of information that is available to the speech-feature combination. Also note, that the MPM-pp approach does not require any manually transcribed or tagged data. However, both speech-based methods are significantly outperformed by the visual estimation of the speaker's head orientation. The combination of audio and visual information leads to additional small gains [41].

Table 10. Attentional state classification with audio and visual estimation

Estimation	Precision	Recall	F-Measure	Classification
Feature Combi FC	0.19	0.91	0.31	49
MPM-pp	0.21	0.79	0.33	53.5
Head Pose (HP)	0.57	0.81	0.67	90
FC + HP	0.65	0.81	0.72	92

5.8 Language Dependencies

Implicit in our classification methodology is the assumption that phone strings originating from phone recognizers trained on different languages yield complementary information. In the following experiments we explore the influence of the variation of the phone recognizers, and investigate to what extend the performance varies with the number of languages covered.

We conducted one set of experiments to investigate whether the reason for the success of the multilingual phone string approach is related to the fact that the different languages contribute useful classification information or that it simply lies in the fact that different recognizers provide complementary information. If the latter were the case, a multi-engine approach in which phone recognizers trained on the same language but on different channel or speaking style conditions might do a comparably good job. To test this hypothesis, we had trained three different phone recognizers solely on a single language, namely English but on various different channel conditions (telephone, channel-mix, clean) and different speaking styles (highly conversational, spontaneous, planned) using data from Switchboard, Broadcast News, and Verbmobil. The experiments were carried out on matched conditions on all distances for 60 second chunks for the speaker identification task. To compare the three single-language engines to the multiple-language engines, we generated all possible language triples out of the set of 12 languages ($\binom{12}{3} = 220$ triples) and calculated the average, minimum and maximum performance over all triples. The results are given in Table 11.

The results show that the multiple-engine approach lies in all but one case within the range of the multiple-language approach. However, the average performance of the multiple-language approach always outperforms the multiple-engine approach. This indicates that most of the language triples achieve better results than the single language multiple-engines. From these results we draw the conclusion that multiple English language recognizers provide less useful information for the classification task than do multiple language phone recognizers. This is at least true for the given choice of multiple engines in the context

Table 11. Multiple languages versus single-language multiple engines [SIDrates %]

Dis	Multiple Languages	Multiple Engines
Dis 0	94.6 (80.0-100)	93.3
Dis L	93.1 (80.0-96.7)	86.7
Dis 1	89.5 (76.7-96.7)	86.7
Dis 2	93.6 (86.7-96.7)	76.7
Dis 4	90.8 (73.3-96.7)	86.7
Dis 5	92.0 (73.3-96.7)	83.3
Dis 6	89.5 (60.0-96.7)	63.3
Dis 8	87.2 (63.3-96.7)	63.3

of speaker identification. We also conducted experiments, in which the multi-engine recognizers were combined with the multilingual recognizers, but did not see further improvements [56]. The fact that the multiple engines were trained on English, i.e. the same language which is spoken in the speaker identification task, whereas the multiple languages were trained on 12 languages but English, makes the multiple-language approach even more appealing as it indicates a great potential for portability to speaker characteristic classification tasks in any language.

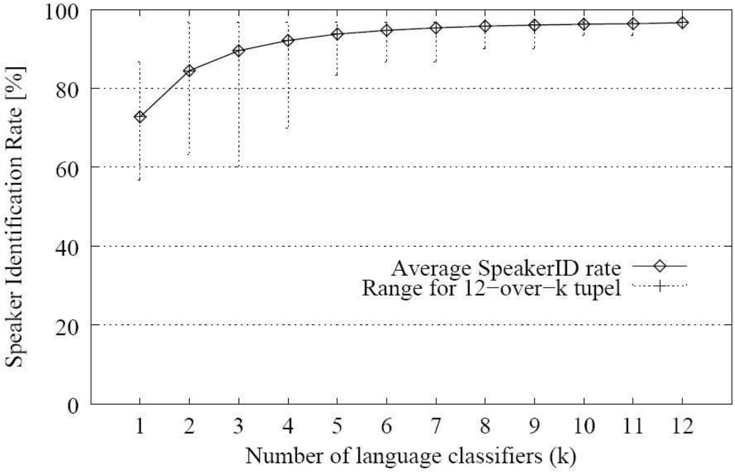


Fig. 6. Classification rate over number of phone recognizers

In the final set of experiments, we investigated the impact of the number of languages, i.e. the number of phone recognizers on speaker identification performance. Figure 6 plots the speaker identification rate over the number k of languages used in the identification process on matched conditions on 60 seconds data. The performance is given in average over the k out of 12 language k-tupel for all distances. The results indicate that the average speaker identification rate increases for all distances with the number of involved phone recognizers. For some distances a saturation effect takes place after 6 languages involved (distance 0 and 1), for others distances even adding the 12th language has a positive effect on the average performance (distance 4, 6, L). Figure 6 shows that the maximum performance of 96.7% can already be achieved using two languages. Among the total of $\binom{12}{2} = 66$ language pairs, CH-KO and CH-SP gave the best results. We were not able to derive an appropriate strategy to predict the best language tupels. Therefore, it is comforting that the increasing average indicates that the chances of finding suitable language tupels get better with the number of applied languages. While only 4.5% of all 2-tupels achieved highest performance, 35% of 4-tupels, 60% of all 6-tupels, and 88% of all 10-tupels gave optimal performance. We furthermore analyzed if the performance is related to the total number of phones used for the classification process rather than the number of different engines, but did not find evidence for such a correlation.

6 Conclusion

This chapter briefly outlined existing speech-driven applications in order to motivate why the automatic recognition of speaker characteristics is desirable. After categorizing relevant characteristics, we proposed a taxonomy, which differentiates between physiological and psychological aspects, and furthermore considers the individual speaker as well as the collective. The language-dependent characteristics language, accent, dialect, idiolect, and sociolect were described in more detail. The brief overview of classification approaches was complemented by a practical example of our implementation of a speaker characteristics identification system. This implementation applies a joint framework of multilingual phone sequences to classify various speaker characteristics from speech, such as identity, gender, language, accent and language proficiency, as well as attentional state. In this system the classification decisions were based on phonetic n-gram models trained from phone strings, performing a simple minimum perplexity rule. The good classification results validated this concept, indicating that multilingual phone strings can be successfully applied to the classification of various speaker characteristics. The evaluation on a far-field speaker identification task proved the robustness of the approach, achieving 96.7% identification rate under mismatching conditions. Gender identification gave 94% classification accuracy. We obtained 97.7% discrimination accuracy between native and non-native English speakers and 95.5% language identification rate on 5 sec chunks discriminating 4 languages. In the classification of the attentional state, the MPM-pp approach performs slightly better than a combination of higher-level speech features, achieving 53.5% classification rate. Furthermore, we compared the performances between multi-lingual and multi-engine systems and examined the impact of the number of involved languages on classification results. Our findings confirm the usefulness of language variety and indicate a language independent nature of our experiments. These encouraging results suggest that the classification of speaker characteristics using multilingual phone sequences could be ported to any language. In conclusion, we believe that the classification of speaker characteristics has advanced to a point where it can be successfully deployed into real-world applications. This would allow for more personalization, customization, and adaptation to the user and thus meet our desire for a more human-like behavior of speech-driven automated systems.

Acknowledgments. The author wishes to thank Qin Jin for providing all results on speaker and gender identification, Kornel Laskowski for his work on language identification, Alicia Tribble for performing the experiments on accent and proficiency levels, and Michael Katzenmaier for his contributions to the classification of attentional states.

References

1. Sacks, O.W.: The Man who Mistook His Wife for a Hat - and other Clinical Trials. New York (summit Books) (1985)
2. Krauss, R.M., Freyberg, R., Morsella, E.: Inferring speakers' physical attributes from their voices. Journal of Experimental Social Psychology 38, 618–625 (2002)
3. Nass, C., Brave, S.: Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. MIT Press, Cambridge (2005)
4. Sproat, R.: Review in Computational Linguist 17.65 on Nass and Brave 2005. Linguist List 17.65 (2006) http://linguistlist.org/issues/17/17-65.html
5. Nass, C., Gong, L.: Speech Interfaces from an Evolutionary Perspective: Social Psychological Research and Design Implications. Communications of the ACM 43(9), 36–43 (2000)
6. Nass, C., Lee, K.M.: Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In: CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 329–336. ACM Press, New York (2000)
7. Tokuda, K.: Hidden Markov model-based Speech Synthesis as a Tool for constructing Communicative Spoken Dialog Systems. In: Proc. 4th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, Special Session on Speech Communication: Communicative Speech Synthesis and Spoken Dialog, invited paper, Honolulu, Hawaii (2006)
8. Doddington, G.: Speaker Recognition - Identifying People by their Voices. Proceedings of the IEEE 73(11), 1651–1664 (1985)
9. Meng, H., Li, D.: Multilingual Spoken Dialog Systems. In: Multilingual Speech Processing, pp. 399–447. Elsevier, Academic Press (2006)
10. Seneff, S., Hirschman, L., Zue, V.W.: Interactive problem solving and dialogue in the ATIS domain. In: Proceedings of the Fourth DARPA Speech and Natural Language Workshop, Defense Advanced Research Projects Agency, pp. 1531–1534. Morgan Kaufmann, Pacific Grove (1991)
11. Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A.: Creating natural dialogs in the Carnegie Mellon Communicator system. In: Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH), Budapest, Hungary, pp. 1531–1534 (1999)

12. Litman, D., Forbes, K.: Recognizing Emotions from Student Speech in Tutoring Dialogues. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, Virgin Islands (2003)
13. Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L.: JUPITER: A telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing 8(1) (2000)
14. Hazen, T., Jones, D., Park, A., Kukolich, L., Reynolds, D.: Integration of Speaker Recognition into Conversational Spoken Dialog Systems. In: Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland (2003)
15. Muthusamy, Y.K., Barnard, E., Cole, R.A.: Reviewing Automatic Language Identification. IEEE Signal Processing Magazin (1994)
16. Gorin, A.L., Riccardi, G., Wright, J.H.: How may I help you? Speech Communication 23(1/2), 113–127 (1997)
17. Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E.: How to find trouble in communication. Speech Communication 40, 117–143 (2004)
18. Polzin, T., Waibel, A.: Emotion-sensitive Human-Computer Interfaces. In: Proc. ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast, Northern Ireland (2000)
19. Raux, A., Langner, B., Black, A.W., Eskenazi, M.: LET'S GO: Improving Spoken Language Dialog Systems for the Elderly and Non-natives. In: Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland (2003)
20. ELLS: The e-language learning system. ELLS Web-server. Retrieved December, 2006 (2004) from http://ott.educ.msu.edu/elanguage/
21. Eskenazi, M.: Issues in the Use of Speech Recognition for Foreign Language Tutors. Language Learning and Technology Journal 2(2), 62–76 (1999)
22. Barnard, E., Cloete, J.P.L, Patel, H.: Language and Technology Literacy Barriers to Accessing Government Services. In: Traunmüller, R. (ed.) EGOV 2003. LNCS, vol. 2739, pp. 37–42. Springer, Heidelberg (2003)
23. CHIL: Computers in the human interaction loop. CHIL Web-server. Retrieved December, 2006 (2006), from http://chil.server.de
24. Schultz, T., Waibel, A., Bett, M., Metze, F., Pan, Y., Ries, K., Schaaf, T., Soltau, H., Westphal, M., Yu, H., Zechner, K.: The ISL Meeting Room System. In: Proceedings of the Workshop on Hands-Free Speech Communication (HSC-2001), Kyoto, Japan (2001)
25. Waibel, A., Bett, M., Finke, M., Stiefelhagen, R.: Meeting browser: Tracking and summarizing meetings. In: Penrose, D.E.M. (ed.) Proceedings of the Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, pp. 281–286. Morgan Kaufmann, San Francisco (1998)
26. AMI: Augmented multi-party interaction. AMI Web-server. Retrieved December, 2006 (2006), from http://amiproject.org/
27. Vogel, S., Schultz, T., Waibel, A., Yamamoto, S.: Speech-to-Speech Translation. In: Multilingual Speech Processing. Elsevier, Academic Press, pp. 317–398 (2006)
28. GALE: Global autonomous language exploitation. GALE Program. Retrieved December, 2006 (2006), from http://www.darpa.mil/ipto/Programs/gale/index.htm
29. Wahlster, W. (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. LNCS (LNAI). Springer, Berlin Heidelberg New York (2000)
30. Waibel, A., Soltau, H., Schultz, T., Schaaf, T., Metze, F.: Multilingual Speech Recognition. In: The Verbmobil Book, Springer, Heidelberg (2000)

31. McNair, A., Hauptmann, A., Waibel, A., Jain, A., Saito, H., Tebelskis, J.: Janus: A Speech-To-Speech Translation System Using Connectionist And Symbolic Processing Strategies. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toronto, Canada (1991)
32. Cincarek, T., Toda, T., Saruwatari, H., Shikano, K.: Acoustic Modeling for Spoken Dialog Systems based on Unsupervised Utterance-based Selective Training. In: Proc. of the International Conference on Spoken Language Processing (ICSLP), Pittsburgh, PA (2006)
33. Kemp, T., Waibel, A.: Unsupervised Training of a Speech Recognizer using TV Broadcasts. In: Proc. of the International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, pp. 2207–2210 (1998)
34. Schultz, T., Waibel, A.: Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. Speech Communication 35(1-2), 31–51 (2001)
35. Doddington, G.: Speaker recognition based on idiolectal differences between speakers. In: Proceedings of Eurospeech (2001)
36. Goronzy, S., Tomokiyo, L.M., Barnard, E., Davel, M.: Other Challenges: Non-native Speech, Dialects, Accents, and Local Interfaces. In: Multilingual Speech Processing. Elsevier, Academic Press, pp. 273–315 (2006)
37. Jessen, M.: Speaker Classification in Forensic Phonetics and Acoustics. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007) (this issue)
38. Eriksson, E., Rodman, R., Hubal, R.C.: Emotions in Speech: Juristic Implications. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007) (this issue)
39. Reynolds, D.: Tutorial on SuperSID. In: JHU 2002 Workshop. Retrieved December, 2006 (2002) from http://www.clsp.jhu.edu/ws2002/groups/supersid/SuperSID_Tutorial.pdf
40. Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K.: The Recognition of Emotion. In: The Verbmobil Book, pp. 122–130. Springer, Heidelberg (2000)
41. Katzenmaier, M., Schultz, T., Stiefelhagen, R.: Human-Human-Robot Interaction. In: International Conference on Multimodal Interfaces, Penn State University - State College, PA (2004)
42. Kirchhoff, K.: Language Characteristics. In: Multilingual Speech Processing. Elsevier, Academic Press, pp. 5–32 (2006)
43. Goronzy, S.: Robust Adaptation to Non-Native Accents in Automatic Speech Recognition. LNCS (LNAI), vol. 2560. Springer, Heidelberg (2002)
44. Wang, Z., Schultz, T.: Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization. In: Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland, pp. 1449–1452 (2003)
45. Fischer, V., Gao, Y., Janke, E.: Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer. In: Proc. of the International Conference on Spoken Language Processing (ICSLP) (1998)
46. Sancier, M.L., Fowler, C.A.: Gestural drift in bilingual speaker of Brazilian Portuguese and English. Journal of Phonetics 25, 421–436 (1997)
47. Cohen, P., Dharanipragada, S., Gros, J., Monkowski, M., Neti, C., Roukos, S., Ward, T.: Towards a universal speech recognizer for multiple languages. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 591–598 (1997)

48. Fügen, C., Stüker, S., Soltau, H., Metze, F., Schultz, T.: Efficient handling of multilingual language models. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 441–446 (2003)
49. Navrátil, J.: Automatic Language Identification. In: Multilingual Speech Processing. Elsevier, Academic Press, pp. 233–272 (2006)
50. Reynolds, D.: An Overview of Automatic Speaker Recognition Technology. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, pp. 4072–4075 (2002)
51. Huang, X.D., Acero, A., Hon, H-W.: Spoken Language Processing. Prentice Hall PTR, New Jersey (2001)
52. Reynolds, D.: A Gaussian mixture modeling approach to text-independent using automatic acoustic segmentation. PhD thesis, Georgia Institute of Technology (1993)
53. Kohler, M.A., Andrews, W.D., Campbell, J.P., Hernander-Cordero, L.: Phonetic Refraction for Speaker Recognition. In: Proceedings of Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark (2001)
54. Jin, Q., Navratil, J., Reynolds, D., Andrews, W., Campbell, J., Abramson, J.: Cross-stream and Time Dimensions in Phonetic Speaker Recognition. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), HongKong, China (2003)
55. Campbell, J.P.: Speaker recognition: A tutorial. Proceedings of the IEEE 85, 1437–1462 (1997)
56. Jin, Q.: Robust Speaker Recognition. PhD thesis, Carnegie Mellon University, Language Technologies Institute, Pittsburgh, PA (2007)
57. Cimarusti, D., Ives, R.: Development of an automatic identification system of spoken languages: Phase 1. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Paris (1982)
58. Zissman, M.A.: Language Identification Using Phone Recognition and Phonotactic Language Modeling. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 5, pp. 3503–3506. Detroit, MI (1995)
59. Hazen, T.J., Zue, V.W.: Segment-based automatic language identification. Journal of the Acoustical Society of America 101(4), 2323–2331 (1997)
60. Navrátil, J.: Spoken language recognition - a step towards multilinguality in speech processing. IEEE Trans. Audio and Speech Processing 9(6), 678–685 (2001)
61. Parandekar, S., Kirchhoff, K.: Multi-stream language identification using data-driven dependency selection. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2003)
62. Torres-Carrasquillo, P., Reynolds, D., Deller, Jr., J.: Language identification using gaussian mixture model tokenization. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2002)
63. Eady, S.J.: Differences in the f0 patterns of speech: Tone language versus stress language. Language and Speech 25(1), 29–42 (1982)
64. Schultz, T., Rogina, I.A.W.: Lvcsr-based language identification. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, Georgia, IEEE (1996)
65. Schultz, T.: Globalphone: A multilingual text and speech database developed at karlsruhe university. In: Proc. of the International Conference on Spoken Language Processing (ICSLP), Denver, CO (2002)
66. Jin, Q., Schultz, T., Waibel, A.: Speaker Identification using Multilingual Phone Strings. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL (2002)

67. NIST: Speaker recognition evaluation plan. Retrieved December, 2006 (1999) from `http://www.itl.nist.gov/iaui/894.01/spk99/spk99plan.html`
68. Tomokiyo-Mayfield, L.: Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in LVCSR. PhD thesis, CMU-LTI-01-168, Language Technologies Institute, Carnegie Mellon, Pittsburgh, PA (2001)
69. Schultz, T., Jin, Q., Laskowski, K., Tribble, A., Waibel, A.: Speaker, accent, and language identification using multilingual phone strings. In: Proceedings of the Human Language Technologies Conference (HLT), San Diego, Morgan Kaufman, San Francisco (2002)
70. Schultz, T., Jin, Q., Laskowski, K., Tribble, A., Waibel, A.: Improvements in non-verbal cue identification using multilingual phone strings. In: Proceedings of the 40nd Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, The Association for Computational Linguistics (2002)

Foreign Accent

Ulrike Gut

Freiburg University
ulrike.gut@anglistik.uni-freiburg.de

Abstract. This paper investigates how speakers can be classified into native and non-native speakers of a language on the basis of acoustic and perceptually relevant features in their speech. It describes some of the most salient acoustic properties of foreign accent, based on a comparative corpus analysis of native and non-native German and English. These properties include the durational features vowel reduction, consonant cluster reduction and overall speech rate as well as the intonational variables pitch range and pitch movement. The paper further presents an experiment demonstrating that perceptual judgments of foreign accent correlate primarily with the speakers' speech rate.

Keywords: foreign accent, acoustic properties, perceptual judgments and acoustic correlates.

1 Introduction

Speakers are traditionally classified into native and non-native speakers of a language although, at closer inspection, the division line between the two classes is far from clear-cut. "Native" speakers of a language are usually exposed to the language from birth on, acquire it fully and use it throughout their lives. "Non-native" speakers of a language usually come into contact with it at a later stage, for example in formal classroom teaching or by immigration to a foreign country. They often do not acquire the language fully and continue to use other languages in their daily lives. Speech produced by the latter group typically shows properties of a "foreign accent". As yet, among linguists, no exact, comprehensive and universally accepted definition of foreign accent exists. However, there is a broad consensus that the term refers to the deviations in pronunciation of non-native speech compared to the norms of native speech (e.g. Scovel 1969:38). Foreign accent can be measured in two ways: by eliciting global judgments and quality ratings of samples of non-native speech from judges or by carrying out instrumental-acoustic measurements of various phonetic aspects of non-native speech and by comparing them to native speech.

This article examines both the acoustic and perceptual correlates of foreign-accented German and English. In the first part, instrumental-phonetic analyses of the acoustic correlates of foreign accent will be presented and the various segmental and prosodic features of non-native speech that may contribute to a foreign accent are discussed. The second part of the article is concerned with the perceptual correlates of foreign accent. The results of an experiment investigating the correlation between perceptual accent ratings and acoustic properties of non-native speech will be presented.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 75–87, 2007.

2 Acoustic Correlates of Foreign Accent

Foreign accent has been divided into phonological and phonetic accent, the former comprising phonological deviations such as phoneme substitutions, as for example in the pronunciation of *the* as [də], and the latter referring to incorrect pronunciations of otherwise correct phonological representations (Markham 1997). In addition, foreign accent can be divided into segmental deviations, i.e. phoneme substitutions or incorrect pronunciations of individual vowels and consonants, and prosodic deviations such as deviant speech rhythm, intonation and stress patterns. The majority of descriptions of the correlates of foreign accent are based on auditory analyses and manual transcriptions of deviations and often lack in systematization and representativeness. Systematic instrumental analyses of the phonetic properties of non-native speech have shown a number of acoustic deviations in foreign-accented speech. For example, it was found that non-native English produced by Japanese, Spanish-speaking, Jordanian and Brazilian learners differs from native speech in terms of the voice onset time (VOT) of plosives (Riney & Takagi 1999, Flege & Munro 1994, Flege, Frieda, Wally & Randazza 1998, Port & Mitleb 1983, Major 1987a). Likewise, the realization of consonant clusters by Brazilian learners of English is suggested to contribute to their foreign accent (Major 1987b). English produced by native speakers of Polish, French, Tunisian Arabic and Spanish is characterized by a lack of vowel reduction and the non-realization of weak vowels in unstressed syllables (Scheuer 2002, Wenk 1985, Ghazali & Bouchhioua 2003, Mairs 1989, Flege & Bohn 1989). Furthermore, German learners of English produce different vowel qualities for the phonemes /e/ and /æ/ than English native speakers do (Barry 1989), English learners of Thai produce deviant tones (Wayland 1997) and Austrian learners of English show differences from native speakers in the realization of falling pitch movements (Grosser 1997).

The majority of studies concerned with the phonetic correlates of foreign accent carried out so far are restricted to the investigation of a particular combination of native language and target language such as Japanese-accented English. The purpose of the present study, in contrast, is to determine the general properties of foreign accent. The following questions are raised: Is it possible to classify speakers into native and non-native on the basis of some acoustic features of their speech? Which acoustic features distinguish non-native speech from native speech irrespective of the speakers' first language? Which of these acoustic features correlate with human auditory judgments of the strength of foreign accent? The focus of the present study lies on the acoustic characteristics of a foreign accent in both German and English. In particular, three acoustic features of non-native speech will be investigated: general durational features such as speech rate, reduction processes in both vowels and consonant clusters and features of pitch including pitch range and pitch movement.

For this, a large-scale corpus-based study of the acoustic properties of non-native speech was carried out. It is based on the LeaP corpus, which consists of 359 recordings of non-native and native speech in both German and English comprising 73.941 words and a total amount of recording time of more than 12 hours (Milde & Gut 2002, Pitsch, Gut & Milde 2003). It contains four different speaking styles: free speech in an interview situation (length between 10 and 30 minutes), reading of a passage (length of about two minutes), retellings of a story (length between two and 10 minutes) and the reading of nonsense word lists (30 to 32 words). During the collection of the corpus data it was aimed to record a representative range of non-native speakers in terms of age, sex, native language/s, level of competence, length of exposure to the target language, age at first exposure to the target language and non-linguistic factors such as motivation to learn the language, musicality and so forth. The non-native English in the corpus was produced by 46 speakers with 17 different native languages, whose age at the time of the recording ranges from 21 to 60. 32 of them are female and 14 are male. Their average age at first contact with English is 12.1 years, ranging from one year to 20 years of age. The age of the 55 non-native speakers of German at the time of recording ranges from 18 to 54 years. 35 of them are female and 20 are male. Altogether, they have 24 different native languages. The average age at first contact with German is 16.68 years, ranging from three years to 33 years of age. The corpus further contains eight recordings with native speakers of (British) English and 10 recordings with native speakers of Standard German.

2.1 Durational Features of Foreign Accent: Speech Rate

The object of the first set of acoustic analyses was to explore differences between native and non-native speech in terms of general durational features. These features include the overall articulation rate as well as the duration of various linguistically meaningful units such as utterances and syllables. Utterances were defined as sequences of words between two pauses of a minimum length; the division of syllables was based on standard phonological criteria (e.g. Giegerich 1992). Syllables were further divided into stressed and unstressed since the difference between these two types of syllables is correlated with significant differences in duration in both native English and native German (e.g. Hoequist 1983, Campbell 1989, Gut 2003). The story retellings and the reading passages in the LeaP corpus were analyzed with the following quantitative measurements:

- **artrate:** articulation rate (total number of syllables divided by total duration of speech)
- **mlu:** mean length of utterance (in syllables)
- **mls:** mean length of stressed syllables
- **mlr:** mean length of reduced syllables (unstressed syllables with reduced or deleted vowel)

A total of 40.274 syllables produced by the non-native speakers of German, 3.261 syllables produced by the native speakers of German, 30.871 syllables produced by the non-native speakers of English and 2.492 syllables produced by the English native speakers were analyzed.

Table 1 shows that non-native English differs significantly from native English in all aspects of general speech rate. Non-native retellings, on average, show a slower articulation rate and a shorter mean length of utterance than story retellings by native speakers. Moreover, the mean length of syllables, both stressed and unstressed, is significantly longer in non-native speech. When reading the story, the non-native speakers produce a slower articulation rate as well as a shorter mean length of utterance and longer syllables of both types.

Table 1. Mean values of artrate, mlu, mls and mlr for the non-native and the native speakers of English in the retellings and reading passage style. (*** equals significance at p<0.001, ** equals significance at p<0.01, * equals significance at p<0.05).

		artrate	mlu	mls	mlr
retellings	non-native English	2.3	3.8	280.7	155.4
	native English	4.1	7.5	209.3	90.2
		***	***	***	***
reading passage style	non-native English	3.25	5.9	258.6	140.4
	native English	4.1	8.9	212.25	101.3
		*	**	*	*

A comparison of non-native German with native German gives similar results (Table 2). On average, native story retellings have a longer mean length of utterance, shorter stressed syllables and a higher articulation rate than their non-native counterparts. The readings of the story by the non-native speakers differ from the native readings in three acoustic variables: non-native readings have a slower articulation rate and have, on average, longer syllables. No significant difference was found between the non-native and native mean length of utterance in reading passage style.

2.2 Durational Features of Foreign Accent: Reduction

The second line of investigation was concerned with reduction processes in native and non-native speech. In both German and English, vowel reduction and vowel deletion occur regularly in specific contexts. Reduced vowels in German and English are shorter than full vowels and change their quality (e.g. Delattre 1981, Gut 2006). For example, reduction is illustrated in the production of the schwa /ə/ as the first vowel in the

Table 2. Mean values of artrate, mlu, mls and mlr for the non-native and the native speakers of German in the retellings and reading passage style. (**=significant at p<0.01, *=significant at p<0.05).

		artrate	mlu	mls	mlr
retellings	non-native German	2.4	4.4	254.9	189.2
	native German	3.3	7.2	212.7	159.7
		*	**	*	n.s.
reading passage style	non-native German	3.3	6.5	232.7	178.4
	native German	4.1	7.9	198.7	137.2
		*	n.s.	**	**

English word *alike* [əlaɪk] or the second vowel in the German word *diesem* [dizəm]. Vowel deletion often occurs in the realization of the second syllable in the German word *laufen* as [fn] and in the second syllable of the English word *nation* [neɪʃn]. Likewise, in both languages word-final consonant clusters, i.e. sequences of two or more consonants, are regularly reduced in connected speech. This means that for example in the words *jumped* and *hast* one or more consonants of the cluster are deleted so that they are realized as [jʌmt] and [has] (e.g. Neu 1980, Kohler 1995). In the LeaP corpus, the following measurements of reduction processes were taken:

- **percentage reduced/deleted vowels (prv):** percentage of all syllables with reduced or deleted vowel of all syllables
- **ratio full/red:** mean durational ratio of all syllable pairs in which a syllable with a full vowel is followed by a syllable with a reduced or a deleted vowel
- **2consclus:** retention rate (i.e. no deletion) of all word-final consonant clusters in words with phonologically underlying two-consonant clusters
- **3consclus:** retention rate of all word-final three-consonant clusters and four-consonant clusters
- **content words:** retention rate of all word-final two-, three- and four-consonant clusters in content words (nouns, verbs, adjectives and adverbs)
- **function words:** retention rate of all word-final two-, three- and four-consonant clusters in function words (prepositions, conjunctions and auxiliary verbs)

A total of 40.274 syllables produced by the non-native speakers of German, 3.261 syllables produced by the native speakers of German, 30.871 syllables produced by the non-native speakers of English and 2.492 syllables produced by the English native speakers were analyzed in terms of vowel reduction. In addition, a total of 3.965 words with underlying word-final clusters produced by the non-native speakers of English and a total of 229 such words produced by the native English speakers were analyzed. 4.045 potential word-final coda clusters were analyzed in the speech of the non-native speakers of German. The native German speakers produced a total of 232 words with underlying word-final consonant clusters.

Table 3 illustrates various significant differences in vowel reduction and consonant-cluster reduction between the non-native and the native speakers of English. The non-native speakers produce, on average, fewer syllables with reduced and deleted vowels and a smaller durational difference between neighboring syllables with a full vowel and a reduced or deleted vowel. Non-native and native speakers of English do not differ in the retention rate of two-consonant clusters. Conversely, the native speakers reduce three-consonant clusters significantly more frequently than the non-native speakers. Word-final clusters in content words are retained more often than in function words in both types of speech, but the retention rate of clusters in function words is significantly higher in non-native English than in native English.

Table 4 illustrates the that there are fewer differences in vowel and consonant cluster reduction between non-native German and native German. The overall percentage of syllables with reduced and deleted vowels does not differ between non-native German and native German. In contrast, in non-native German, the durational difference between adjacent syllables with full vowels on the one hand and reduced or deleted

Table 3. Percentage of syllables with reduced and deleted vowel of all syllables, mean durational ratio of adjacent syllable pairs with the first syllable containing a full and the second a reduced or deleted vowel (ratio full/red), overall retention rate of word-final two-consonant and three-consonant clusters and retention rate of word-final clusters in content words and function words produced by the non-native and the native speakers of English. (***=significant at p<0.001; **=significant at p<0.01).

	prv	**ratio full/red**	**2consclus**	**3consclus**	**content wors**	**function words**
non-native English	24.01	1.98:1	80.2	37.12	70.8	44.2
native English	30.65	2.45:1	82.5	4.77	73.3	20.5
	**	**	n.s.	***	n.s.	***

vowels on the other is lower. For word-final consonant clusters in German, the overall retention rate is not significantly different between the two speaker groups, neither in two- or three-consonant clusters nor in content words and function words.

Table 4. Percentage of syllables with reduced and deleted vowel of all syllables, mean durational ratio of adjacent syllable pairs with the first syllable containing a full and the second a reduced or deleted vowel (ratio full/red), overall retention rate of word-final two-consonant and three-consonant clusters and retention rate of word-final clusters in content words and function words produced by the non-native and the native speakers of German. (***=significant at p<0.001).

	prv	**ratio full/red**	**2consclus**	**3consclus**	**content wors**	**function words**
non-native German	28.66	1.49:1	65.1	41.4	65.9	59.5
native German	29.2	1.76:1	74.8	70	82.8	66.6
	n.s.	***	n.s.	n.s.	n.s.	n.s.

2.3 Pitch Range and Pitch Movement in Foreign-Accented Speech

The third acoustic feature investigated as a possible correlate of foreign accent was pitch. The height of pitch changes continuously across an utterance, but the linguistically most important pitch movement is the utterance-final pitch movement, often referred to as the nucleus. In both English and German, nuclear pitch movements can have the form of falls or rises or combinations of the two (e.g. Grabe 1998). Another linguistically relevant aspect of pitch is the pitch range, which expresses the difference between the maximum and the minimum pitch height in an utterance or sequence of utterances (e.g. Patterson 2000). Two different measurements were taken for the retellings and story readings in the LeaP corpus:

- **pitch range:** average difference between the highest and lowest pitch in the entire recording (in semitones)
- **falls:** average extent of pitch movement in falling nuclear tones in semitones
- **rise:** average extent of pitch movement in rising nuclear tones in semitones

In total, 910 falling and 803 rising nuclear tones were produced by the non-native speakers of English and 86 falls and 30 rises were produced by the native English speakers. The non-native speakers of German produced a total of 1.208 falling and 1.379 rising pitch movements, however, many of them were realized as steps up or down and not as continuous pitch movements. The native speakers produced 112 falling and 61 rising pitch movements, also including steps up and down.

Distinct differences in pitch range exist between native and non-native speakers in both languages. Table 5 illustrates that, although for both speaker groups the average pitch range is smaller in the retellings than in the readings, the average pitch range in native English is greater than that in non-native English in both speaking styles.

Table 5. Mean pitch range in the reading passages and the retellings and average extent of falling and rising nuclear pitch movements in non-native and native English. (***=significant at p<0.001; **=significant at p<0.01).

	pitch range reading	**pitch range retelling**	**fall**	**rise**
non-native English	12	10.3	3.64	4.129
native English	17	12.7	7.81	3.8
	**	***	**	n.s.

Table 5 further illustrates that the nuclear falls in non-native English, on average, are significantly smaller than the nuclear falls produced by the native speakers of English. On average, native speakers' falling pitch movements extend over 7.81 semitones, which is more than twice as much as in the falls produced by the non-native speakers. In contrast to non-native English, in native English, nuclear rises, on average, are much smaller than falling nuclear pitch movements.

Native German also has a wider average pitch range than non-native German in both reading passage style and the retellings. Similarly, in native German, falls have a more pronounced slope than in non-native German. They extend over an average of 5.67 semitones in native German, but only 3.8 semitones in non-native German. The slope of rises in native German is, on average, smaller than that of falls, which is a further difference from non-native German.

Table 6. Mean pitch range in the reading passages and the retellings and average extent of falling and rising nuclear pitch movements in non-native and native German. (*=significant at p<0.05).

	pitch range reading	**pitch range retelling**	**fall**	**rise**
non-native German	12.7	13.12	3.8	4.98
native German	15.3	16.7	5.67	4.19
	*	*	*	n.s.

3 Perceptual Correlates of Foreign Accent

Studies in which the degree of foreign accent is rated by judges differ greatly in terms of the procedures used to elicit and evaluate non-native speech. For example, raters are presented with different scales comprising a varying number of equal-appearing intervals, often labeled as ranging from "very strong foreign accent" to "no accent, native-like" and the type of non-native speech judged by the raters varies from readings of single sentences to samples of spontaneous speech. In addition, the number and professional background of judges in foreign accent rating tasks varies considerably. Nevertheless, a number of studies have shown that native speakers as raters of foreign accent agree to an acceptable degree in their judgments (Cunningham-Andersson & Engstrand 1989, Thompson 1991, Munro & Derwing 1999, Piske, MacKay & Flege 2001, Moyer 1999).

A small number of studies has been concerned with the relationship between foreign accent ratings and specific linguistic parameters of non-native speech. Consonantal features that have been identified to correlate with perceived foreign accent are the voice onset time (VOT) of plosives in non-native English produced by Japanese and by Brazilian speakers (Riney & Takagi 1999, Major 1987a) and the realization of consonant clusters by Brazilian speakers of English (Major 1987b). Scheuer (2002) reports that the non-realization of reduced vowels in unstressed syllables and other vocalic errors correlate most strongly with negative evaluations of Polish speakers' foreign accent ratings. Cunningham-Andersson & Engstrand (1989) list 25 different phonological and phonetic errors that contribute to the impression of a foreign accent in Swedish. Tajima, Port & Dalby (1997) report that the intelligibility of Chinese-accented English sentences was improved by changing the durational patterns of segments to native values. Finally, Anderson-Hsieh, Johnson & Koehler (1992) found that accent ratings correlate with syllable-errors and phoneme substitutions as well as the rated quality of the overall prosody.

The present paper investigates the relationship between foreign accent ratings and those acoustic properties of non-native speech identified as relevant for speaker classification in the previous section. For each speaker in the LeaP corpus, an accent rating was obtained. Seven native speakers of German, four female and three male, with a mean age of 23.8 years and without a professional background in language teaching or assessment rated speech samples by the 55 non-native speakers in the German sub-corpus. The material consisted of an extract from the interview of about 30 seconds' length. The raters were informed that they were to rate the quality of the foreign accent without reference to the speaker's morphosyntactic abilities or possible idiosyncrasies in the use of vocabulary. Prior to the experiment, the raters were provided with three anchor recordings representing a speaker with a very strong accent, a native-like speaker and one with an average foreign accent each. The raters were given a 9-point scale ranging from "very strong accent" to "native-like". The experiment was web-based and gave the raters the opportunity to listen to each of the recordings as often and as long as they wanted.

For the English sub-corpus, only recordings with those non-native speakers aiming at a British English pronunciation, as established in the interviews, were included. They were rated by four male native speakers of British English (mean age 34.5 years) without a professional background in language teaching or assessment, following the same procedure as for the German experiment.

The following acoustic correlates of foreign accent were selected in the free speech recordings and correlated with the foreign accent ratings:

- **mls:** mean length of stressed syllables
- **mlr:** mean length of reduced syllables (unstressed syllables with reduced or deleted vowel)
- **ratio full/red:** mean durational ratio of all syllable pairs in which a syllable with a full vowel is followed by a syllable with a reduced or a deleted vowel
- **3consclus:** retention rate of all word-final three-consonant clusters and four-consonant clusters
- **artrate:** articulation rate (total number of syllables divided by total duration of speech)
- **pitch range:** average difference between the highest and lowest pitch in the entire recording (in semitones)

Table 7 illustrates which of the acoustic properties of non-native speech correlate with the mean accent ratings. It can be seen that only those properties of non-native speech that have to do with speed of delivery correlate with the mean accent ratings for the non-native speakers of German: the mean length of stressed syllables, the mean length of reduced syllables and the articulation rate. Pitch range, cluster reduction and vowel reduction measured in the ratio between adjacent full-vowelled and reduced syllables do not correlate significantly with ratings of foreign accent. None of the acoustic

Table 7. Correlation of acoustic features with the mean accent rating for the non-native speakers of German and the non-native speakers of English

	mls	**mlr**	**ratio full/red**	**3consclus**	**artrate**	**pitch range**
German	.46**	.38*	-.12	-.15	.-38*	-.18
English	.31	.27	-.3	.32	-.28	-.18

measurements listed in Table 7 correlate with the accent ratings received by the non-native speakers of English. It seems that raters base their judgments on other acoustic cues than those listed in Table 7, although these were found to constitute areas of systematic divergence between native and non-native speech.

4 Summary and Conclusion

The objectives of the present paper were to examine whether and how speakers can be classified into native and non-native speakers on the basis of the acoustic features of their speech. In particular, it was investigated which acoustic features distinguish non-native from native speech irrespective of the speakers' first language and which of these acoustic features correlate with human auditory judgments of the strength of foreign accent. A comparative corpus analysis of native and non-native English and German was carried out that focused on the acoustic properties of the general durational features in speech rate, vowel reduction and consonant cluster reduction and on the intonational parameters pitch range and pitch movement.

The results show that non-native speech varies systematically from native speech with respect to general durational properties. In both English and German, native speakers produce a significantly higher articulation rate and longer mean length of utterance than non-native speakers. Overall, both stressed and unstressed syllables are longer in non-native speech, which makes it slower than native speech. It was further found that the non-native speakers' speech rate varies with speaking style. In reading passage style articulation rate is significantly faster than in the story retellings and the mean length of utterance is longer. This constitutes another area of difference between non-native and native speech, as differences in speech rate between speaking styles are far less pronounced in native speech.

The second prominent difference between native and non-native speech lies in the realization of vowel reduction and deletion. In particular, this concerns the lack of durational difference between syllable pairs in which a syllable with a full vowel precedes a syllable with a reduced or deleted vowel. Only in about a third of the recordings contained in the LeaP corpus, the durational difference between those two types of syllables equals that of native speech. Lack of vowel reduction is especially evident in non-native English. Whereas the non-native speakers of German produce the same overall amount of reduced or deleted vowels than the German native speakers, non-native speakers of English do not succeed in a relatively sufficient reduction or deletion of vowels.

The third line of research concerned the reduction of consonant clusters in non-native speech. On the whole, word-final consonant clusters in non-native English are likelier to be retained, i.e. to be produced faithfully, than to be simplified by reduction. The greatest difference between word-final cluster reduction in native and non-native English lies in the reduction of three-consonant clusters, which are nearly always reduced in native speech but produced faithfully in about a third of all cases by the non-native speakers. Furthermore, in native English, the reduction rate of consonant clusters in function words is much greater than in non-native English. Non-native German does not differ from native German in terms of word-final cluster reduction.

The fourth acoustic feature analyzed in non-native speech was pitch range. Pitch range in non-native speech is, on the whole, narrower than in native speech. However, the analysis of native pitch range in the LeaP corpus showed distinct differences with speaking style in English. Reading passage style is characterized by a pitch range that is on average five semitones wider than that of the retellings. This is mirrored in non-native English where pitch range, on average, is also wider in reading passage style than in the semi-spontaneous speech in the retellings. No such variation of pitch range with speaking style was found for either native or non-native German. Another significant difference between native and non-native speech lies in the phonetic realization of utterance-final falling tones, which are shorter in non-native speech. A comparison of nuclear falls and rises shows that the non-native speakers' falls tend to be shorter than their rises, that is the pitch movement stretches over fewer semitones. In native speech, in contrast, the pitch movement of falls is distinctly greater than that of rises.

The second aim of the present paper was to find acoustic correlates of human foreign accent ratings. Of those acoustic features of non-native speech that had proven to vary systematically between native and non-native speakers in the previous analyses, however, only the general durational properties such as articulation rate and mean length of syllables correlated with native speaker ratings of the degree of foreign accent. Vowel reduction, consonant cluster reduction and pitch range did not seem to influence the accent ratings given by native speaker judges. This finding replicates results described by Neumeyer, Franco, Weintraub & Price (1996). They investigated a number of acoustic properties such as segmental accuracy and timing scores of non-native French and their correlation with native speakers' pronunciation ratings and found the only reliable relationship between durational properties and accent ratings. Yet, the present paper did not include an analysis of segmental deviances in non-native speech. It is likely that apart from durational values other acoustic cues guide the decision of native judges of foreign accent, for example phonemic substitutions and other segmental processes. This was shown by Cunningham-Andersson & Engstrand (1989), who isolated various phonetic and phonological features in Swedish that were reliably identified as "foreign accented" by native speaker judges and by Moyer (2004), whose native speaker judges listed a number of phoneme substitution errors as criteria for their ratings of accent in German.

In conclusion, the present paper showed that there are a number of general acoustic features of non-native speech that differ significantly from native speech. The most valid of them are features of speech rate as demonstrated in the correlation with human judgments of foreign accent. However, before these findings can be applied directly for an (automatic) speaker classification one needs to consider that the method of a quantitative corpus analysis has the drawback that it cannot do justice to the speech of individual non-native speakers. Not every non-native speaker has a foreign accent as many studies have shown: some speakers who acquire a language as late as in their twenties are indistinguishable from native speakers even in strict experimental conditions (e.g. Bongaerts et al. 1997, Moyer 1999). Qualitative analyses of individual speakers' speech properties thus need to complement quantitative corpus analyses. The group values presented here, for example, disguise that of the 46 non-native speakers of English, 12 produce a pitch range similar to that of the native speakers and 14 produce falls and rises with a slope equal to that of the native speakers. Likewise, 12 of the German non-native speakers produce falls and rises that are phonetically identical to those of the native speakers.

References

Anderson-Hsieh, J., Johnson, R., Koehler, K.: The relationship between native speaker judgments on nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. Language Learning 4(2), 529–555 (1992)

Barry, W.: Perception and production of English vowels by German learners: instrumental-phonetic support in language teaching. Phonetica 46, 155–168 (1989)

Bongaerts, T., van Summeren, C., Planken, B., Schils, E.: Age and ultimate attainment in the pronunciation of a foreign language. Studies in Second Language Acquisition 19, 447–465 (1997)

Campbell, N.: Syllable duration determination. In: Proceedings of the 1st European Conference on Speech Communication and Technology (Eurospeech '89), vol. 2, pp. 698–701, Paris, France (1989)

Cunningham-Andersson, U., Engstrand, O.: Perceived strength and identity of foreign accent in Swedish. Phonetica 46, 138–154 (1989)

Delattre, P.: An acoustic and articulatory study of vowel reduction in four languages. In: Delattre, P. (ed.) Studies in Comparative Phonetics, pp. 63–93. Groos, Heidelberg (1981)

Flege, J., Bohn, O.-S.: An instrumental study of vowel reduction and stress placement in Spanish-accented English. Studies in Second Language Acquisition 11, 35–62 (1989)

Flege, J., Frieda, E., Walley, A., Randazza, L.: Lexical factors and segmental accuracy in second language speech production. Studies in Second Language Acquisition 20, 155–187 (1998)

Flege, J., Munro, M.: The word unit in second language speech production and perception. Studies in Second Language Acquisition 16, 381–411 (1994)

Ghazali, S., Bouchhioua, N.: The learning of English prosodic structures by speakers of Tunisian Arabic: word stress and weak forms. In: Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS '03), pp. 961–964, Barcelona (2003)

Giegerich, H.: English Phonology. Cambridge University Press, Cambridge (1992)

Grabe, E.: A Comparison of English and German Intonational Phonology. Series in Linguistics. MPI (1998)

Grosser, W.: On the acquisition of tonal and accentual features of English by Austrian learners. In: James, A., Leather, J. (eds.) Second Language Speech - Structure and Process, pp. 211–228. Mouton de Gruyter, Berlin (1997)

Gut, U.: Non-native speech rhythm in German. In: Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS '03), pp. 2437–2440, Barcelona (2003)

Gut, U.: Unstressed vowels in non-native German. In: Hoffmann, R., Mixdorff, H. (eds.) Proceedings of the 3rd International Conference on Speech Prosody (Speech Prosody '06), Dresden, Germany (2006)

Hoequist, C.: Syllable duration in stress-, syllable- and mora-times languages. Phonetica 40, 203–237 (1983)

Kohler, K.: Einführung in die Phonetik des Deutschen. Schmid, Berlin (1995)

Mairs, J.: Stress assignment in interlanguage phonology: an analysis of the stress system of Spanish speakers learning English. In: Gass, S., Schachter, J. (eds.) Linguistic Perspectives on Second Language Acquisition, pp. 260–283. Cambridge University Press, Cambridge (1989)

Major, R.: English voiceless stop production by speakers of Brazilian Portuguese. Journal of Phonetics 15, 197–202 (1987a)

Major, R.: Phonological similarity, markedness, and rate of L2 acquisition. Studies in Second Language Acquisition 9, 63–82 (1987b)

Markham, D.: Phonetic Imitation, Accent, and the Learner. Lund University Press, Lund (1997)

Milde, J.-T., Gut, U.: A prosodic corpus of non-native speech. In: Bel, B., Marlien, I. (eds.) Proceedings of the International Conference on Speech Prosody (Speech Prosody '02), pp. 503–506, Aix-en-Provence, France (2002a)

Moyer, A.: Ultimate attainment in L2 phonology. Studies in Second Language Acqusition 21, 81–108 (1999)

Moyer, A.: Age, Accent and Experience in Second Language Acquisition: An Integrated Approach to Critical Period Inquiry. Multilingual Matters. Clevedon (2004)

Munro, M., Derwing, T.: Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning 49, 285–310 (1999)

Neu, H.: Ranking of constraints on /t,d/ deletion in American English. In: Labov, W. (ed.) Locating Language in Time and Space, pp. 37–54. Academic Press, New York (1980)

Neumeyer, L., Franco, H., Weintraub, M., Price, P.: Automatic text-independent pronunciation scoring of foreign language student speech. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICLSP '96), pp. 1457–1460, Philadelphia, PA (1996)

Patterson, D.: A Linguistic Approach to Pitch Range Modelling. PhD thesis, University of Edinburgh (2000)

Piske, T., MacKay, I., Flege, J.: Factors affecting degree of foreign accent in an L2: a review. Journal of Phonetics 29, 191–215 (2001)

Pitsch, K., Gut, U., Milde, J.-T.: Multimodale bilinguale Korpora. In: Seewald, U. (ed.) Sprachtechnologie für multilinguale Kommunikation, pp. 406–421. Gardez! Verlag, St-Augustin (2003)

Port, R., Mitleb, F.: Segmental features and implementation in acquisition of English by Arabic speakers. Journal of Phonetics 11, 219–229 (1983)

Riney, T., Takagi, N.: Global foreign accent and voice onset time among Japanese EFL speakers. Language Learning 49, 275–302 (1999)

Scheuer, S.: What makes foreign accent sound foreign? In: James, A., Leather, J., (eds.) Proceedings of New Sounds 2000, pp. 306–314. University of Klagenfurt, Klagenfurt, Austria (2002)

Scovel, T.: Foreign accents, language acquisition, and cerebral dominance. Language Learning 20, 245–253 (1969)

Tajima, K., Port, R., Dalby, J.: Effects of temporal correction on intelligibility of foreign-accented English. Journal of Phonetics 25, 1–24 (1997)

Thompson, I.: Foreign accent revisited: The English pronunciation of Russian immigrants. Language Learning 41, 177–204 (1991)

Wayland, R.: Non-native production of Thai: acoustic measurements and accentedness ratings. Applied Linguistics 18, 345–373 (1997)

Wenk, B.: Speech Rhythms in Second Language Acquisition. Language and Speech 28, 157–175 (1985)

Acoustic Analysis of Adult Speaker Age

Susanne Schötz

Lund University, Dept. of Linguistics and Phonetics,
Centre for Languages and Literature
susanne.schotz@ling.lu.se
http://www.sol.lu.se

Abstract. Information about the age of the speaker is always present in speech. It is used as perceptual cues to age by human listeners, and can be measured acoustically and used by automatic age estimators. This chapter offers an introduction to the phonetic study of speaker age, with focus on what is known about the acoustic features which vary with age. The age-related acoustic variation in temporal as well as in laryngeally and supralaryngeally conditioned aspects of speech has been well documented. For example, features related to speech rate, sound pressure level (SPL) and fundamental frequency (F_0) have been studied extensively, and appear to be important correlates of speaker age. However, the relationships among the correlates appear to be rather complex, and are influenced by several factors. For instance, differences have been reported between correlates of female and male age, between speakers of good and poor physiological condition, between chronological age and perceived age, and also between different speech sample types (e.g. sustained vowels, read or spontaneous speech). More research is thus needed in order to build reliable automatic classifiers of speaker age.

Keywords: Speaker age, Phonetics, Acoustic analysis, Acoustic correlates.

1 Introduction

Every human being goes through the process of ageing. This is a very complex process, which affects us in numerous ways, including the way we speak. Our voices and speech patterns change from early childhood to old age. Although most changes occur in childhood and puberty, age-related variation can be observed throughout our adult lives into old age. Consequently, our age is reflected in our speech, and speaker age can be – and has been – studied using several methodological approaches, mainly acoustic analysis and perception experiments.

This chapter offers an introduction to the phonetic study of speaker age, with focus on acoustic variation. First, a summary is given of the age-related changes in the speech production mechanism, followed by short reviews of the study of speaker age from a perceptual and machine recognition perspective. The main part of the chapter comprises an overview of several known acoustic correlates of adult speaker age, including an overview of factors influencing these correlates.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 88–107, 2007.

2 Ageing of the Speech Production Mechanism

From young adulthood to old age, the speech production mechanism undergoes numerous anatomical and physiological changes, which have not all been fully explored. For instance, there are substantial gender differences in the extent and timing of the ageing process [1,2]. Moreover, the physiological differences between individuals seem to grow with advancing age [3]. It is also important, but sometimes difficult, to distinguish among age-related, disease-related and environment-related changes in speech. Linville [4,2,5] has provided excellent reviews of the numerous changes occurring in speech as we grow older. This section is mainly based on her work.

2.1 Respiratory System

Changes in the respiratory system affect speech breathing as well as the voice. The respiratory system reaches its full size after puberty but continues to change throughout adulthood to old age. Changes include decreased lung capacity (mainly due to loss of elasticity in lung tissue), stiffening of the thorax and weakening of respiratory muscles.

2.2 Larynx

The age-related changes of the larynx after it has reached its full size in puberty are numerous, and they affect mainly fundamental frequency and voice quality. Ossification of cartilages occurs later and is less extensive in females (fourth decade) than in males (third decade), while calcification probably occurs later than ossification in both females and males (cf. [6,7,8,9]).

Muscle atrophy occurs in all intrinsic laryngeal muscles. As research has focused on the vocal folds, we do not know to which extent other intrinsic muscles are affected. Whether there are any gender differences is also still unclear. The changes in the complex structure of the vocal folds with increased speaker age are substantial. Besides general degeneration and atrophy, the folds shorten in males (particularly after age 70). Also, the epithelium (the thin outer protective layer of tissue) thickens progressively in females, especially after age 70, while it thickens in males up to age 70 but then grows thinner again. The mucous glands reduce their secretions, leading to less hydrated vocal folds, particularly in males. There also seems to be some evidence of laryngeal nerve degeneration, as well as some changes in the blood supply to the laryngeal muscles.

2.3 Supralaryngeal System

Changes in the supralaryngeal system may also affect speech. The craniofacial skeleton grows continuously by about 3–5% from young adulthood to old age. Muscle atrophy occurs in the facial, mastication and pharyngeal muscles. A slight lowering of the larynx in the neck increases the length of the vocal tract. Extensive degenerative changes occur in the temporomandibular joint, including a gradual reduction in size and reductions in blood supply. In the oral cavity, the mucosa grow thinner and lose elasticity, which is most apparent after age 70, and the mucosal surface roughens. Changes in the pharynx and soft palate include thinning of the epithelium, muscle atrophy and decreased sensation. The tongue surface becomes thinner and fissured, while the tongue muscles suffer from atrophy and fatty infiltration, beginning in the second or third decade.

2.4 Neuromuscular Control

The effects of ageing on motor function can be observed in both the peripheral and the central nervous system. They may affect speech rate, co-ordination of articulators and breath support as well as the regulation of fundamental frequency (F_0). Peripheral changes include a type of "dying back" neuropathy, where the distal ends of the nerve fibres are affected earlier. Also, the number of motor units declines and conduction velocity slows down slightly.

Central changes include a decline in brain weight from age 20 to 90 by about 10% as well as a decrease in brain size. There are reports of decreases in the number of nerve cells in the cortex as well as age-related changes in these cells, which may slow down motor movements. In addition, dopamine levels in the brain may decline by up to 50%, leading to slower sensorimotor processes.

2.5 Female and Male Ageing

In addition to what has already been mentioned, a few more words deserve to be said about the differences between female and male ageing. These are often related to the timing and extent of age-related changes throughout life. One obvious difference is the dramatic changes occurring in males at puberty; another is that females experience greater changes around menopause. Nevertheless, the age-related changes in adults are generally greater in men than in women as regards (1) the extent of laryngeal structure change, (2) fine-motor control of laryngeal abductory and adductory movements, (3) tongue movements and (4) speech rate. It has also been noted that the mucous membranes in the larynx are more sensitive in females than in males and that females may thus be more vulnerable to age-related changes in this respect (P. Kitzing, personal communication, 31 January 2006). On the other hand, men and women display similar age-related changes in speech breathing.

3 Perception and Automatic Recognition of Speaker Age

Human listeners are able to judge speaker age at levels considerably better than chance. A large number of perception tests have been carried out with a various types of subjects, speech material and testing conditions. In recent years, a few studies on machine perception (or automatic recognition) of speaker age have emerged as well. This section briefly summarises human perception of age, and also describes a number of experiments on automatic recognition of speaker age.

3.1 Human Perception of Speaker Age

Most people are able to estimate an individual's age from speech samples alone at accuracy levels significantly better than chance [10,11,12,2], perhaps because of constant confrontation with this task throughout our lives, e.g. when listening to someone on the telephone or radio [13]. However, we are still unable to tell exactly how well listeners are able to judge speaker age. The numerous perception studies of speaker age have varied considerably in method and speech material, as well as in speaker and listener characteristics, and the results are often difficult to compare. Listeners' choice of cues and the accuracy obtained seem to depend on the type and length of the speech samples [14]. Moreover, the relationship of the perceptual cues used by listeners in age estimation with the acoustic correlates of chronological as well as perceived age has still not been fully established. In fact, the cues used by listeners to estimate speaker age do not always correspond to age-related changes which can be measured acoustically [4].

From a large number of studies concerning perception of speaker age, we have learned that human listeners are fairly good at estimating the age of an unknown (and unseen) speaker. Perceptual cues to speaker age include variation in pitch, speech rate, voice quality, articulation and phrasing. Moreover, it is likely that listeners use different acoustic cues and listening strategies when estimating the age of female and male speakers. For instance, F_0 seems somewhat more important for the age perception of female speakers than of male ones [15]. In addition, stimulus duration (i.e. longer speech samples, regardless of speech type) seems to be important when judging female speakers, while stimulus type (i.e. spontaneous speech, regardless of duration) seems to be more important in the case of male speakers [15].

Human perception of age is influenced by numerous phonetic as well as non-phonetic factors, e.g. the physiological state of the speaker, the age of the listener and the speech sample type. These factors have to be regarded in machine perception of age as well.

3.2 Automatic Recognition of Speaker Age

Automatic recognition of age can be used to improve human–machine communication. If user age could be identified automatically, spoken dialogue systems could adapt their communication behaviour. For instance, the system could use more youthful language when talking to a teenager. It could also suggest age-adapted information, such as tourist attractions or directions.

As the number of children and elderly people who use computers in their daily lives increases, age-adapted speech recognition is becoming more important. Still, research on automatic age recognition is relatively scarce [16]. One explanation is that it certainly is not an easy task. Age cues are present in every phonetic dimension, and they are hard to separate from other speaker variation characteristics, such as physiological condition and dialect. This section summarises the relatively few attempts to build automatic age estimators.

Minematsu et al. [17,18] built automatic classifiers of perceived age (PA, judged by 12 students) using linear discriminant analysis (LDA) and artificial neural networks (ANN) with mel frequency cepstral coefficients (MFCC), ΔMFCC and amplitude derivatives (ΔPower) as features. Eighty-six speakers (43 judged as elderly and 43 as non-elderly) were modelled using Gaussian mixture models (GMM) and normal distribution (ND). Elderly speakers were correctly identified in 90.9% of cases using LDA. The classifier was then improved by adding the features speech rate and local perturbation of power. This increased the identification rate to 95.3%.

Shafran et al. [16] used hidden Markov model (HMM) based classifiers with cepstral and F_0 features to recognise gender, age, dialect and emotion from a corpus consisting of 1,854 phone calls (65% female, 35% male callers) to a customer care system. The corpus contained a total of 5,147 utterances with an average length of 15 words divided into five age groups: (< 25, ≈ 25, 26–50, ≈ 50 and > 50). A trivial classifier assigning the most probable class label to all test points (33.3%) served as baseline. Results for age were 68.4% correct classifications using only cepstral features, and 70.2% correct using cepstral as well as F_0 features.

Minematsu et al. [19] conducted a study with male spekers (123 aged 6–12, 141 aged 20–60 and 143 aged 60–90). Thirty students in their early twenties estimated direct speaker age from single sentences. Each speaker was then modelled with GMM using MFCC, ΔMFCC and ΔPower as features. The two methods used for the machine estimations showed almost the same correlation between human judgements and machine estimation: the first method modelled PA as discrete labels (0.89), while the second one was based on the normal distributions of PA (0.88).

Müller et al. [20] compared six of the most common machine learning approaches for classification tasks – decision trees[1] (DT), ANN, k-nearest neighbour (kNN), naïve Bayes (NB) and support vector machines (SVM) – in a study of automatic classification of age group using jitter and shimmer as features. 393 speakers (about 10,000 utterances from 347 speakers over 60 years, about 5,000 utterances from 46 speakers under 60 years; gender distribution: 162 females, 231 males), were used in the study. All six methods performed significantly better than the baselines, which were simple classifiers always predicting the more frequently occurring class (elderly: 88%, male: 59%). ANN performed best with 96.57% correct age group estimations.

Müller et al. also used Bayesian networks (BN) to integrate a gender classifier with two age classifiers by first separately calculating the probability of a given speaker being female or male as well as being elderly or non-elderly, and then combining the results to obtain the most probable age and gender classification. This approach reduced errors likely to occur in a sequential classifier (gender first, then age), where failure to determine the correct gender strongly affects the performance of a gender-specific age classifier.

[1] C4.5 decision tree induction [21].

Müller [22,23] further developed his approach for age and gender classification under the name Agender, with target applications such as mobile shopping and pedestrian navigation systems. Classification models were trained using the same five machine learning techniques as in [20], i.e. DT, ANN, kNN, NB and SVM, as well as an additional method: GMM. Features were extended to include jitter, shimmer, F_0, HNR (harmonics-to-noise ratio), speech rate (syllables per second), and pause duration and frequency. The number of speakers was increased to a total of 507 female and 657 male speakers, divided into four age classes for each gender. The majority of the speakers were children and seniors. The best accuracy for the four age classes was obtained with ANN (63.5%)

The author [15] carried out two studies with classification and regression trees (CART) to learn more about which acoustic-phonetic features are important in automatic age recognition. The first study used 50 features (e.g. measures of F_0, duration and formant frequencies) from the phoneme segments of 2,048 versions of one Swedish word (*rasa* [ˈʁɑ̀ːsa], ‘collapse’), produced by 214 females and 214 males. The best CART for age group was 72% correct judgements, and the best correlation between direct chronological and estimated age was 0.45. Estimation accuracy was compared with that of human listeners. Although humans and CARTs used similar cues, the human listeners (mean error ± 8.89 years) were better judges of age than the CART estimators (± 14.45 years).

The second study used 748 speakers and 78 features to construct separate estimators of direct age for female, male and all speakers. CARTs were built for 390 single features, 13 feature groups (consisting of all features for one phonetic quality, e.g. F_1, B_1 and L_1) and five larger feature groups of all prosodic, all resonance, all inverse filtered, all spectral and all features. Results showed that F_0 and duration were the most important single features. Of 13 feature groups, F_0 and duration performed best for female speakers, while the formant groups of F2 and F3 were best for the male speakers. For the larger groups, the CART using all features was the best for female speakers, while the group with all prosodic features performed better for the male speakers. The best estimator of the second experiment (mean error ± 14.07 years) performed only marginally better than the one from the first study.

To sum up this section, automatic age estimation attempts have used MFCC as well as acoustic-phonetic features. The number and age range of speakers have varied among studies, as has the type of speech samples, the method used and the accuracy desired. In order to build reliable automatic age estimators, more knowledge is needed about how different acoustic features vary with speaker age for both genders as well as for different speech sample types and lengths.

4 Acoustic Correlates of Adult Speaker Age

A large number of acoustic features vary with speaker age. This variation is most clearly observable in children, but information about adult speaker age can also be – and has been studied from an acoustic-phonetic perspective. Acoustic variation has been found in temporal as well as in laryngeally and supralaryngeally conditioned aspects of speech. Moreover, the relationships among the numerous acoustic correlates of speaker age appear to be rather complex, and are influenced by several factors, of which several are further described in Section 5.

There are several comprehensive overviews of acoustic correlates of age. For instance, [24] has summarised research up till 1987, and [4,2] has provided excellent reviews of known acoustic aspects of the ageing voice. Based on these sources as well as on several other studies, this section gives an overview of the acoustic features usually related to speaker age. Furthermore, in an attempt to clarify which features have been found to be important age correlates, some of the reported acoustic variation with increased age is summarised in Table 1. Variation with chronological age (CA) as well as with perceived age (PA) in women and men is described.

4.1 General Variation

Old women and men alike demonstrate a general higher intra-subject as well as inter-subject variation of acoustic features when compared with young speakers. For example, increased variation has been found in some F_0 measures, as well as in speech rate (e.g. phoneme duration and VOT), vocal sound pressure level (SPL), jitter, shimmer and HNR [25,26,2]. More age-related differences have been found for male than female speakers [27], and higher correlations of acoustic features with PA than with CA have generally been observed [28]. Moreover, correlations seem to vary with speech sample type [28].

4.2 Speech Rate

Temporal – static as well as dynamic – aspects of speech are strongly affected by the age of the speaker. The speech rate is linked to segment (syllable, phoneme, sub-phoneme, etc.) duration, to the number of speech segments per time unit and also to pause duration and frequency. A large number of studies have found a 20–25% decrease with older CA in speaking and reading rates. Increases have been found in consonant, vowel and sub-phonemic (prevoicing, plosive closure and release, vowel transition) durations as well as in pause duration and frequency [29,30,24,31,32,33,34,35,2,28,15,36]. Women often demonstrate a smaller decrease in speech rate with older CA than men, or none at all [15,37]. This feature also appears to show a larger inter-speaker variation for female speakers [15]. Slower speech rates, a larger number of breaths and longer pause durations have been related to old male and female PA [28].

The results for the sub-phonemic segment voice onset time (VOT) are rather confusing. Some studies have found elderly (CA) women and men to exhibit shorter overall VOTs than younger people [29,31,38]. However, increased VOT with older male CA has also been observed [26]. Other researchers have reported only subtle differences and increased variation with advancing age [39,2,15]. It has also been suggested that age-related differences in VOT is related to phonetic context and perhaps even languages [40,15].

Table 1. Some reported acoustic variation with increased chronological age (CA) and perceptual age (PA) in female and male adult speakers (*decr.*: decrease, *dur.*: duration, *flat.*: flatter, *freq.*: frequency, *incr.*: increase, *no*: no change, *sp.*: spectral, *steep.*: steeper). Please refer to the text for details (adapted from [15]).

Group	*Feature*	*Variation with increasing adult age*			
		Female		*Male*	
		CA	*PA*	*CA*	*PA*
general	variation	incr.		incr.	
	overall changes	few	more	many	more
speech rate	syllables/second	decr. or no		decr.	decr.
	utterance dur.	incr.	incr.	incr.	incr.
	phoneme dur.	incr.		incr.	
	VOT	incr., decr. or no		incr., decr. or no	
	pause freq.&dur.	incr.	incr.	incr.	incr.
sound pressure level (SPL)	mean SPL	no		incr. or decr.	
	max. SPL range	decr.		decr.	
	amplitude SD	incr. or no	incr. or no	incr. or no	incr.
F_0	mean F_0	first no or decr., then decr., incr. or no	decr.	first decr., then incr.	first decr., then incr.
	F_0 range	first incr., then decr.	incr. or no	first incr., then decr. or no	
	F_0 SD	incr. or no	incr. or no	incr., decr. or no	incr.
tremor	vocal tremor	incr. or no	incr.	no	
jitter & shimmer	jitter	incr. or no	incr. or no	incr. or no	
	shimmer	incr. or no	incr. or no	incr. or decr.	
sp. noise	HNR	decr. or no		varying or no	
	NHR	incr. or no	incr. or no	incr. or varying	
sp. energy distribution	sp. tilt	flat. or no		steep., flat. or no	
	sp. tilt (LTAS)	steep. or varying		flat. or varying	
	sp. emphasis	no or varying		no or varying	
	sp. balance	no		no	
resonance	F_1	decr. or no	decr.	decr. or no	decr.
	F_2	incr., decr. or no	decr.	incr., decr. or no	decr.
	F_3–F_4	decr. or no.		decr. or no	
	F_1–F_3(LTAS)	decr.	no	decr.	decr.

4.3 Sound Pressure Level (SPL)

Conversational speech SPL (also called intensity level) appears to remain stable or decrease slightly with increased CA, but has also been reported to increase for men after age 70, even for speakers without hearing loss [41,24,42,2,15]. The habitual SPL range in vowels is likely to increase with advancing female and male CA, and may be an important correlate of speaker age [43,15,36]. However, the maximum vowel SPL range seems to decrease in both women and men, while minimum SPL levels increase for women (to the author's knowledge, no studies have been made concerning men) with advancing CA [31,2].

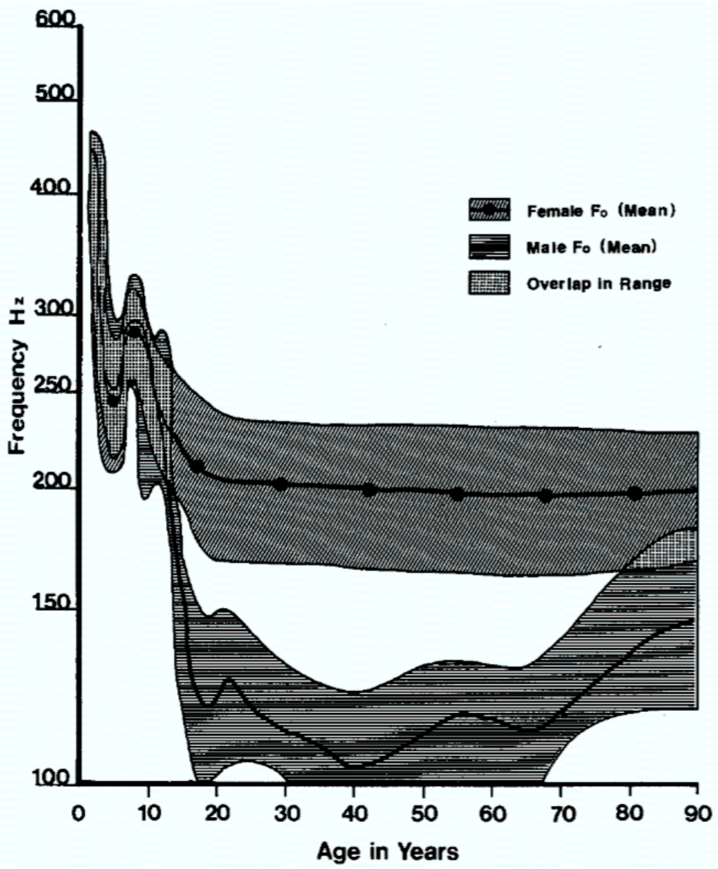


Fig. 1. Speaking F_0 and its standard deviation as a function of speaker age (1–90 years) for female and male speakers (source: [44])

4.4 Fundamental Frequency (F_0)

F_0 patterns in speech related to CA are different for women and men, as shown in Figure 1. Female F_0 has been found to remain fairly constant until menopause, when a drop (of about 10–15 Hz) usually occurs. F_0 then remains stable into old age, but may also rise or lower slightly [45,46,27,47,4,43]. Observations of decreasing F_0 from age 20 to 50 in females have also reported [15,36]. A lower F_0 is also associated with older female PA [2,28]. In males, F_0 lowers slightly (by about 10 Hz) from young adulthood to middle CA, but then rises considerably (about 35 Hz) with old CA [48,46,2,15,36]. Higher F_0 has been reported to be a cue to old male PA [33,2]. However, there are also studies which have failed to find correlations between mean F_0 and CA in men [3]. Moreover, the way changes in F_0 relate to perceptual cues is not in line with the above findings. For instance, listeners have reported lower male F_0 to be a cue to older age [2].

Maximum phonational frequency range – i.e. the complete range of frequencies which a speaker can produce, from the lowest (without creak) to the highest tone (including falsetto) – expands in the lower end following menopause in females, but is restricted in both the upper and lower ends later in life [49,2]. Contradictory findings suggest that men either undergo similar changes in F_0 range as women [50,2], or that old and young males do not differ in F_0 range unless physiological condition and state of health are taken into account [3,51]. A larger habitual F_0 range has been observed for the vowel /a/ in both women and men of old CA [43]. Relatively stable habitual F_0 range values for both genders until about the age of 60, followed by an increase (females) or decrease (males) have also been observed [15].

4.5 Variation in F_0 and Amplitude

Fundamental frequency and amplitude instability and variation are related to various voice qualities. Jitter and shimmer (see p. 97) are often connected with harshness, hoarseness or vocal roughness, while increases in the more gross F_0 variation, as measured in standard deviation (F_0 SD), may cause vocal tremor or a "wobbling" voice quality [24,52,2].

Higher F_0 SD (with greater variation for females) has been found in both men and women with advancing CA and PA [24,49,33,2,43], but sometimes only a minor correlation has been reported, or none at all [3,28,15]. Substantial increases in fundamental amplitude standard deviation (Amp SD) have been demonstrated in older men and women, and have been associated with both CA and PA [53,43]. However, relatively stable Amp SD values with advancing age have been reported as well [15], and Brückl and Sendlmeier [28] found a strong positive correlation with female CA and PA only in spontaneous speech but almost none in sustained vowels or read speech.

Jitter and shimmer are defined as period-to-period fluctuations in vocal fold frequency and amplitude, as shown in Figure 2, and they are considered to be correlates of rough or hoarse voice quality. These features have often been analysed in acoustic studies of age using a number of measures with varying results. Although sometimes no correlation with age has been found for jitter [51,32,54,55,15,36], other researchers have reported increased jitter levels for older female and male CA (but not PA) [49,53,26,43,22]. However, higher and more variable jitter values seem to be more related to physiological health than to age [3,53,2,28].

Higher shimmer levels have been found for older female CA and PA as well as for older male CA (independently of health) [3,51,53,26,43,22]. However, stable (females) or decreasing shimmer levels after age 40 (males) have also been observed [15,36]. Other studies have found shimmer to correlate strongly with CA and PA only in spontaneous speech samples (but not in read speech or in prolonged vowels) [28]. Other studies have observed correlations of shimmer and CA in sustained vowels, but only when 80-year-olds were compared with younger age groups [55].

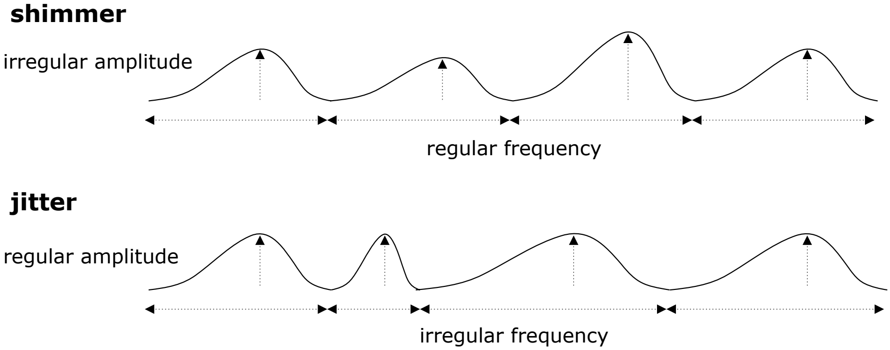


Fig. 2. Irregularities (microvariations) in vocal fold movements can be measured as shimmer (variation in amplitude) and jitter (variation in frequency) (after [22])

Linville [2] concludes that it is impossible to draw any firm conclusions as to the effect of ageing on jitter and shimmer since several factors, including sound pressure level, mean F_0, analysis system differences and individual health and fitness variables, appear to have a strong effect on these measures, especially in women. Moreover, the large number of measures used for these features and the differences in speech material used in various studies also appear to contribute to the problem with comparison of results.

4.6 Other Voice Measures

Spectral tilt (ST), spectral emphasis (SE) and spectral balance (SB) are all measures of the relative energy levels in different frequency bands of the spectrum [56,57]. ST usually represents the slope – i.e. the difference between the energy levels of two different frequency bands – of the source (inverse filtered) spectrum in dB per octave. SE is a measure of the relative energy levels in the higher frequency bands, while SB is often measured in four contiguous frequency bands. The three measures have sometimes been defined differently [58,59,57,60].

ST has been observed either to flatten (i.e. the energy in the frequency band 4–5 kHz increased with female and male CA) in some vowels, or to remain relatively stable until age 60 (females) or 80 (males), where an increase follows [15]. A longitudinal study found a steeper spectral tilt in old men compared with the same men when young [61]. SE and SB have been found not to correlate significantly with CA [59,15].

The age-related variation of the energy distribution in long-term average spectra (LTAS) has also been studied to some extent. An LTAS is an averaged spectrum of all voiced sounds across a relatively long speech sample. Elderly women have been observed to have higher spectral levels at 320, 6080, 6240, 6400, 6560 and 6720 Hz but lower levels at 3040 and 3200 Hz than young women, and a tendency for older women to have higher levels at 160 Hz has been found as well [62]. Somewhat higher female LTAS levels with advancing age have been observed for 160 (but only from age group 40 to 70), 320 and 2240–2560 Hz, while slightly lower levels with increased age were found at about 5920–7200 Hz [15]. Differences in spectral amplitude have been found between old and young men, too. Old males have demonstrated higher LTAS levels at 160 Hz and lower levels at 1600 Hz than young males [62]. Moderately higher LTAS amplitudes with advancing male age have also been found at 160 (but only for age group 40 to 70), 320 and 1760–2080 Hz [15]. A strong spectral attenuation of high frequencies has also been observed in LTAS at older CA and PA in males, but not in females [63].

Spectral noise is defined as the unmodulated aperiodic energy in vowel spectra [2]. It is considered an acoustic correlate of breathy and harsh or hoarse voice quality [64,65], and has been analysed using various methods. Visual analysis of spectral noise in spectrograms has shown that this feature is much more strongly correlated with physiological condition than with CA [66].

The harmonics-to-noise ratio (HNR) is a measure that quantifies the amount of additive noise in the voice signal, and it can be calculated in several ways [67,68]. The ratio reflects the dominance of the periodic level over the aperiodic one, as quantified in dB. HNR has sometimes been reported to decrease with older female CA [54], or to increase with younger male CA [69], while other researchers have failed to find strong correlations with CA in females [69] or both genders [70,15,36]. No studies exist (to the author's knowledge) of HNR in relation to PA.

Other measures of spectral noise used in acoustic studies of speaker age include the parameters VTI, SPI and NHR of the commercial voice quality analysis software Multi-Dimensional Voice Program (MDVP, see e.g. [71]). Voice turbulence index (VTI) is a measure of the relative energy level in high-frequency noise. It is calculated as the average ratio of the inharmonic spectral energy in the 2.85.8 kHz range to the harmonic spectral energy in the 0.074.5 kHz range. Soft phonation index (SPI) measures the relative energy in low-frequency noise, calculated as the average ratio of the lower (0.071.6 kHz) to the higher (1.64.5 kHz) frequency harmonic energy. The noise-to-harmonics ratio (NHR) is the average ratio between noise in the frequency band 1.5–4.5 kHz and the harmonic energy in the frequency band 0.07–4.5 kHz; it is sometimes referred to as a low-frequency harmonics-to-noise ratio [72]. Increased values for all three features in women and men of older CA have been reported [43]. Other researchers have failed to find strong correlations of these features with female and male CA [73,15], though weak (NHR but not VTI) and strong (SPI) positive correlations with female PA [28] or have also been observed.

Vocal tremor can be measured using the MDVP parameters FTRI (intensity of the strongest frequency modulation) and ATRI (intensity of the strongest amplitude modulation). FTRI (but not ATRI) has been found to increase with both female CA and PA in vowels, but not in read or spontaneous speech [28].

4.7 Resonance Measures

Research has revealed that age-related changes in the supralaryngeal structures provide acoustic cues to adult speaker age [10,13,74]. However, there are relatively few studies of the age-related changes in the vocal tract resonance features.

Formant frequencies in vowels have been reported to lower with female and male CA and PA owing to increased vocal tract length [75,76]. There also seems to be a trend towards vowel centralisation (or reduction) for old CA [77,78]. It appears that some old speakers centralise more than others, suggesting an increase in formant frequency variation across speakers of old CA [2]. Moreover, different results have been observed for different vowels. F_1 has been found to decrease with older female age in [yː], and to drop substantially with age for both genders at about age 40 in [ɛː], while other vowels ([a], [ɑː] and [uː]) did not vary much with age in either gender [15,36]. In the same study, F_2 was found to increase with advancing age in [ɑː] and [ɛː] for both genders. In [a] and [uː], F_2 tended to decrease slightly, interrupted by increases and peaks at age group 40 in both genders. A fairly stable F_2 was observed in [yː]. F_3, F_4 and F_5 have been found to show somewhat different patterns depending on gender and vowel quality. Often (but far from always) decreases were observed from age class 20 to 30, followed by little change or a very slight increase, with an occasional rise or fall after age 80 for one or both genders [15]. For PA, formant information seems to lose its significance when F_0 information is present [52].

Energy peaks in long-term average spectra (LTAS), corresponding to the average formant frequencies across all vowels in a speech sample, have been studied by Linville and Rens [79]. They found a significant lowering of peaks 1, 2 and 3 (corresponding to F_1–F_3) with old female CA, and a significantly lower peak 1 (F_1) in old male CA. Moreover, the age-related lowering of peaks was greater in females than in males.

To sum up this section, previous research has found numerous acoustic correlates of chronological and perceptual speaker age. Some features, such as measures of F_0 and speech rate, have been found to be more important than others and have thus been investigated to a larger extent. In addition, there are also a number of factors which may also influence acoustic analysis of speaker age. Some of these factors are described in the following section.

5 Factors Which May Influence Acoustic Analysis of Speaker Age

Several factors (besides age) may affect the analysis outcome in acoustic studies of speaker age. These are often related to the material and the methods used, and may contribute to the divergent and sometimes even contradictory results found in different studies. Differences have been reported between correlates of female and male age, between speakers of good and poor physiological condition, between chronological age (the age of a speaker as measured in time from birth) and perceived age (the mean age of a speaker as estimated by a group of listeners), and also between different speech sample types (e.g. sustained vowels and read or spontaneous speech). This section offers a brief overview of some of the factors which may influence analysis results.

5.1 Speaker-Related Factors

Speaker-related factors include physical (anatomical and physiological) attributes such as gender, race, weight, health and physiological condition. Women and men differ in several vocal characteristics. Some can be explained by anatomical differences while others, such as the paralinguistic use of breathy voice quality, appear to be learned behaviours [80]. Differences in body physiology, vocal training and medical condition may also affect the age-related variation in speech [3,81,82,83], including effects of medication [61] and cigarette smoking [84]. For instance, smokers generally exhibit lower F_0 than non-smokers [85], while professional sopranos and tenors have a higher F_0 than age-matched non-singers [44]. Furthermore, age-related differences in habitual F_0 seem less prominent or even absent in singers and other voice professionals [82].

Cultural, social and psychological factors, including speaker language, dialect, sociolect, emotional state and attitude, may influence and even mask age-related acoustic variation. For instance, there are language-related, dialectal and attitudinal differences in habitual F_0, HNR and shimmer levels [47,86]. Moreover, consideration must also be given to the fact that voice settings are more or less subject to swings in fashion [7], and that the pronunciation of a language is constantly changing. Young individuals often wish to speak differently from their parents [87]. One example is the increased use of the more open allophones [æː] and [œː] of the /ε/ and /ø/ phonemes in Swedish [88]. Another example is the growing use of the glottal stop in British English [87].

5.2 Speech-Material-Related Factors

Speech-material-related factors include the number and age distribution of the speakers and the duration and speech type (and number of speech types) of the speech samples analysed. Fewer speakers will yield less reliable results, as will an unbalanced (for age) speech corpus. Valid measurements of some features are obviously obtained more easily from certain speech types. For instance, formant frequencies are more reliably measured in sustained vowels than in connected speech, and calculations of the average number of syllables per second are more reliable in longer speech samples. Moreover, studies which have used more than one speech type have sometimes found contradictory results for different speech types. One example is Brückl and Sendlmeier [28], who found that vocal tremor correlated with age in sustained vowels, but not in read or spontaneous speech.

5.3 Methodological Factors

Methodological factors, such as differences in recording and analysis equipment and techniques, may strongly influence the outcome of acoustic analyses. One example concerns the vocal effort made by speakers to adapt to the distance to a listener or a microphone, which may affect speech rate, voice quality, measures of F_0 and even some formant frequencies [89]. Different measurement techniques could also be one reason why, for instance, it has not yet been possible to draw any firm conclusions as to the effect of ageing on jitter and shimmer [2].

Another major methodological factor in acoustic studies of speaker age is whether the findings are related to chronological or perceived age. In automatic age recognition applications, the goal is in many cases to identify speakers' actual CA, and not the mean PA as estimated by a group of listeners. However, if only CA is considered in an acoustic study, no knowledge about the relative importance of the correlates to listeners will be gained [11]. On the other hand, when the acoustic correlates of PA are examined, the age judgements of a group of listeners – often quite small – will have to be trusted. Since PA is a subjective measure, results may not be reliable, as listener characteristics (gender, age, etc.) affect the age estimates. Thus, the purpose of each study or application will have to determine whether CA or PA is chosen as the frame of reference.

A connected question is whether we should use archival recordings in combination with recent ones of the same speakers (longitudinal studies) or speech samples from different speakers recorded close in time (cross-sectional studies). Although it may be tempting to use longitudinal data because of the invariant speaker-specific parameters, several aspects which may affect the results should be regarded. Differences in recording equipment and technical sound quality may yield unreliable results. Moreover, voice communication habits may change over time, one example being that Australian women aged 18–25 years recorded in 1993 had significantly lower F_0 levels than women of the same age recorded in 1945 [90]. Another example concerns VOT and F_0 SD. Several cross-sectional studies have reported that VOT decreased and F_0 SD increased in males with advancing age. However, in a longitudinal study of male speakers recorded twice over a period of 30 years, [91] found the opposite results.

In spite of the numerous factors which may affect acoustic analysis, different studies have agreed on several acoustic correlates of speaker age. However, many experiments have varied in the number and choice of speakers and acoustic features, as well as in speech material and method. Some studies have reduced the effect of certain factors by controlling variables or by using a large material.

In summary, speaker age is a very complex characteristic of speech. It leaves traces in all acoustic-phonetic dimensions and it is influenced by numerous other factors, such as physiological condition. Studying it is by no means a trivial task. The studies carried out so far have varied greatly in the type of speech material used (read, spontaneous, prolonged vowels etc.) as well as in analysis method (number and kind of of features investigated). More research is needed with a larger and more systematically varied material and methods to fully explore the age-related acoustic variation in speech and to identify optimal combinations of features for automatic recognition of speaker age.

References

1. Beck, J.M.: Organic variation of the vocal apparatus. In: Hardcastle, W.J., Laver, J. (eds.) The Handbook of Phonetic Sciences, pp. 256–297. Blackwell Publ., Oxford (1997)
2. Linville, S.E.: Vocal Aging. Singular Thomson Learning, San Diego, CA (2001)
3. Ramig, L.A., Ringel, R.L: Effects of physiological aging on selected acoustic characteristics of voice. Journal of Speech and Hearing Research 26, 22–30 (1983)
4. Linville, S.E.: The aging voice. In: Kent, R.D., Ball, M.J. (eds.) Voice Quality Measurement. Singular Thomson Learning, San Diego, CA, pp. 359–376 (2000)
5. Linville, S.E.: The aging voice. The American Speech-Language-Hearing Association (ASHA) Leader 12, 21 (October 19, 2004)
6. Jurik, A.: Ossification and calcification of the laryngeal skeleton. Acta Radiol Diagn. 25, 17–22 (1984)
7. Lindblad, P.: Rösten. Studentlitteratur, Lund (1992)
8. Dedivitis, R.A.: Abrahão, M., Simães, M.J., Mora, O.A., Cervantes, O.W.: Aging histological changes in the cartilages of the cricoarytenoid joint. Acta Cir Bras [serial online] 19 Retrieved 16 August 2006 (2004) from http://www.scielo.br/acb
9. Mupparapu, M., Vuppalapati, A.: Ossification of laryngeal cartilages on lateral cephalometric radiographs. The Angle Orthodontist 75, 196–201 (2005)
10. Ptacek, P.H., Sander, E.K.: Age recognition from voice. Journal of Speech and Hearing Research 9, 273–277 (1966)
11. Ryan, W.J., Burk, K.W.: Perceptual and acoustic correlates of aging in the speech of males. Journal of Communication Disorders 7, 181–192 (1974)
12. Huntley, R., Hollien, H., Shipp, T.: Influences of listener characteristics on perceived age estimations. Journal of Voice 1, 49–52 (1987)
13. Shipp, T., Hollien, H.: Perception of the aging male voice. Journal of Speech and Hearing Research 12, 703–710 (1969)
14. Ramig, L.: Aging speech: Physiological and sociological aspects. Language and Communication 6, 25–34 (1986)
15. Schötz, S.: Perception, Analysis and Synthesis of Speaker Age. PhD thesis, Travaux de l'Institut de linguistique de Lund 47. Lund: Dept. of Linguistics and Phonetics, Lund University (2006)
16. Shafran, I., Riley, M., Mohri, M.: Voice signatures. In: Proc. of The 8th IEEE Automatic Speech Recognition and Understanding Workshop, St. Thomas, U.S. Virgin Islands (2003)
17. Minematsu, N., Sekiguchi, M., Hirose, K.: Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In: Proc. of ICASSP 2002, Orlando, FL, pp. 137–140 (2002)
18. Minematsu, N., Sekiguchi, M., Hirose, K.: Performance improvement in estimating subjective agedness with prosodic features. In: Proc. of Speech Prosody 2002, Aix-en- Provence, pp. 507–510 (2002)
19. Minematsu, N., Yamauchi, K., Hirose, K.: Automatic estimation of perceptual age using speaker modeling techniques. In: Proc. of Eurospeech, Geneva, pp. 3005–3008 (2003)
20. Müller, C., Wittig, F., Baus, J.: Exploiting speech for recognizing elderly users to respond to their special needs. In: Proc. of Eurospeech 2003, Geneva, pp. 1305–1308 (2003)
21. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kauffman, San Mateo (1993)

22. Müller, C.: Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht. PhD thesis, Computer Science Institute, Saarland University (2005)
23. Müller, C.: Automatic recognition of speakers' age and gender on the basis of empirical studies. In: Proc. of Interspeech 2006, Pittsburgh, PA (2006)
24. Hollien, H.: Old voices: What do we really know about them? Journal of Voice 1, 2–13 (1987)
25. Morris, R.J., Brown, W.S.: Age-related differences in speech variability among women. Journal of Communication Disorders 27, 49–64 (1994)
26. Decoster, W.: Akoestische kenmerken van de ouder wordene stem. PhD thesis, Leuwen: Leuwen University Press (Summary in English) (1998)
27. Higgins, M.B., Saxman, J.H.: A comparison of selected phonatory behaviours of healthy aged and young adults. Journal of Speech and Hearing Research 13, 1000–1010 (1991)
28. Brückl, M., Sendlmeier, W.: Aging female voices: An acoustic and perceptive analysis. In: Proc. of VOQUAL'03, Geneva, pp. 163–168 (2003)
29. Benjamin, B.: Phonological performance in gerontological speech. Journal of Psycholinguistic Research 11, 159–167 (1982)
30. Oyer, E., Deal, L.: Temporal aspects of speech and the aging process. Folia Phoniatrica (Basel) 37, 109–112 (1985)
31. Morris, R.J., Brown, W.S.: Age-related voice measures among adult women. Journal of Voice 1, 38–43 (1987)
32. Brown, W.S., Morris, R.J., Michel, J.F.: Vocal jitter in young adult and aged female voices. Journal of Voice 3, 113–119 (1989)
33. Shipp, T., Qi, Y., Huntley, R., Hollien, H.: Acoustic and temporal correlates of perceived age. Journal of Voice 6, 211–216 (1992)
34. Amerman, J.D., Parnell, M.M.: Speech timing strategies in elderly adults. Journal of Voice 20, 65–67 (1992)
35. Slawinski, E.B.: Acoustic correlates of [b] and [w] produced by normal young to elderly adults. Journal of the Acoustical Society of America 95(4), 2221–2230 (1994)
36. Schötz, S., Müller, C.: A study of acoustic correlates of speaker age. In: Speaker Classification II. LNCS(LNAI), vol. 4441, Springer, Heidelberg (2007)
37. Hoit, J., Hixon, K., Altman, M., Morgan, W.: Speech breathing in women. Journal of Speech and Hearing Research 32, 353–365 (1989)
38. Stölten, K., Engstrand, O.: Effects of sex and age in the Arjeplog dialect: A listening test and measurements of preaspiration and vot. In: Proc. of Fonetik 2002, vol. 44, TMH-QPSR, pp. 29–32 (2002)
39. Petrosino, L., Colcord, R.D., Kurcz, K.B., Yonker, R.J.: Voice onset time of velar stop productions in aged speakers. Journal of Perceptual and Motor Skills 76, 83–88 (1993)
40. Neiman, G., Kluch, R., Shuey, E.: Voice onset time in young and 70-year-old women. Journal of Speech and Hearing Research 26, 118–123 (1983)
41. Ryan, W.J.: Acoustic aspects of the aging voice. Journal of Gerontology 27, 256–268 (1972)
42. Morris, R.J., Brown, W.S.: Age-related differences in speech intensity among adult females. Folia Phoniatrica (Basel) 46, 64–69 (1994)
43. Xue, S.A., Deliyski, D.: Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. Educational Gerontology 21, 159–168 (2001)

44. Brown, W.S., Morris, R.J., Hollien, H., Howell, E.: Speaking fundamental frequency characteristics as a function of age and professional singing. Journal of Voice 5, 310–315 (1991)
45. Hollien, H., Shipp, T.: Speaking fundamenal frequency and chronological age in males. Journal of Speech and Hearing Research 15, 155–159 (1972)
46. Kitzing, P.: Glottografisk frekvensindikering: En undersökningsmetod för mätning av röstläge och röstomfång samt framställning av röstfrekvensdistributionen. PhD thesis, Lund University, Malmö (1979)
47. Traunmüller, H., Eriksson, A.: The frequency range of the voice fundamental in the speech of male and female adults. [manuscript]. Retrieved 2 January 2006 (1995) from http://www.ling.su.se/staff/hartmut/aktupub.htm
48. Mysak, E.: Pitch and duration characteristics of older males. Journal of Speech and Hearing Research 2, 46–54 (1959)
49. Linville, S.E.: Acoustic-perceptual studies of aging voice in women. Journal of Voice 1, 44–48 (1987)
50. Ptacek, P.H., Sander, E.K., Maloney, W.H., Jackson, C.C.R.: Phonatory and related changes with advanced age. Journal of Speech and Hearing Research 9, 350–360 (1966)
51. Ringel, R.L., Chodzko-Zajko, W.J.: Vocal indices of biological age. Journal of Voice 1, 31–37 (1987)
52. Linville, S.E.: The sound of senescence. Journal of Voice 10, 190–200 (1996)
53. Orlikoff, R.: The relationship of age and cardiovascular health to certain acoustic characteristics of male voices. Journal of Speech and Hearing Research 33, 450–457 (1990)
54. Ferrand, C.T: Harmonics-to-noise ratio: An index of vocal aging. Journal of Voice 16, 480–487 (2002)
55. Shuey, E., Herr-McCauley, J., Prohaska, C., Martin, K.: Perturbation measures and chronologic age. Presented at the annual convention of the American Speech-Language-Hearing Association (ASHA), November 13–15, Chicago, IL (2003)
56. Campbell, N.: Loudness, spectral tilt, and perceived prominence in dialogues. In: Proc. of ICPhS 95. vol. 3, pp. 676–679. Stockholm (1995)
57. Heldner, M.: Spectral emphasis as an additional source of information in accent detection. In: Bacchiani, M., Hirschberg, J., Litman, D., Ostendorf, M. (eds.) Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, ISCA, Red Bank, NJ, pp. 57–60 (2001)
58. Sluijter, A.M.C., van Heuven, V.J.: Spectral balance as an acoustic correlate of linguistic stress. Journal of the Acoustical Society of America 100, 2471–2485 (1996)
59. Traunmüller, H.: Perception of speaker sex, age, and vocal effort. In: PHONUM, Reports in Phonetics 4. Umeå University, pp. 183–186 (1997)
60. O'Leidhin, E., Murphy, P.: Analysis of Spectral Measures for Voiced Speech with Varying Noise and Pertubation Levels. Proc. of ICASSP 1, 869–872 (2005)
61. Decoster, W., Debruyne, F.: Changes in spectral measures and voice onset time with age: A cross-sectional and a longitudinal study. Folia Phoniatrica et Logopaedica 49, 269–280 (1997)
62. Linville, S.E.: Source characteristics of aged voice assessed from long-term average spectra. Journal of Voice 16, 472–479 (2002)
63. Winkler, R., Brückl, M., Sendlmeier, W.: The aging voice: an acoustic, electroglottographic and perceptive analysis of male and female voices. In: Proc. of ICPhS 03, Barcelona, pp. 2869–2872 (2003)

64. McAllister, A., Sundberg, J., Hibi, S.: Acoustic measurements and perceptual evaluation of hoarseness in children's voices. Logopedics Phoniatrics Vocology 23, 27–38 (1998)
65. Kreiman, J., Gerratt, B.R.: Perception of aperiodicity in pathological voice. Journal of the Acoustical Society of America 117, 2201–2211 (2005)
66. Ramig, L.A.: Effects of physiological aging on vowel spectral noise. Journal of Gerontology 38, 223–225 (1983)
67. Krom, G.d.: A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. Journal of Speech and Hearing Research 36, 224–266 (1993)
68. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proc. of the Institute of Phonetic Sciences, vol. 17, pp. 97–110 (1993)
69. Wang, C.C., Huang, H.T.: Voice acoustic analysis of normal Taiwanese adults. J. Chin. Med. Assoc. 67, 179–184 (2004)
70. Schötz, S.: Prosodic cues in human and machine estimation of female and male speaker age. In: Bruce, G., Horne, M. (eds.) Nordic Prosody. Proc. of the IXth Conference, Lund 2004. Frankfurt am Main: P. Lang, pp. 215–223 (2006)
71. Deliyski, D., Gress, C.: Intersystem reliability of MDVP for DOS and Windows 95/98. Paper presented at the 1998 Annual Convention of American Speech-Language-Hearing Association, San Antonio, Texas (1998)
72. Pereira Jotz, G., Cervantes, O., Abrahao, M., Settanni, F.A.P., de Angelis, E.C.: Noise-to-harmonics ratio as an acoustic measure of voice disorders in boys. Journal of Voice 16, 28–31 (2002)
73. Shuey, E., Herr-McCauley, J., Anders, M.: Indices of turbulence in aging voice. Presented at the annual convention of the American Speech-Language-Hearing Association (ASHA), November 17-20, Philadelphia, PA (2004)
74. Jacques, R., Rastatter, M.: Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners. Folia Phoniatrica (Basel) 42, 118–124 (1990)
75. Endres, W., Bambach, W., Flösser, G.: Voice spectrograms as a function of age, voice disguise, and voice imitation. Journal of the Acoustical Society of America 49, 1842–1848 (1971)
76. Linville, S.E., Fisher, H.: Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. Journal of the Acoustical Society of America 78, 40–48 (1985)
77. Rastatter, M., Jacques, R.: Formant frequency structure of the aging male and female vocal tract. Folia Phoniatrica (Basel) 42, 118–124 (1990)
78. Rastatter, M., McGuire, R., Kalinowski, J., Stuart, A.: Formant frequency characteristics of elderly speakers in contextual speech. Folia Phoniatrica et Logopaedica 49, 1–8 (1997)
79. Linville, S.E., Rens, J.: Vocal tract resonance analysis of aging voice, using long-term average spectra. Journal of Voice 15, 323–330 (2001)
80. Klatt, D., Klatt, L.: Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America 87, 820–857 (1990)
81. Orlikoff, R.: Heartbeat-related fundamental frequency and amplitude variation in healthy young and elderly male voices. Journal of Voice 4, 322–328 (1990)
82. Sataloff, R.T., Rosen, D.C., Hawksha, M., Spiegel, J.R.: The three ages of voice: the aging adult voice. Journal of Voice 11, 156–160 (1997)
83. González, J.: Formant frequencies and body size of speaker: a weak relationship in adult humans. Journal of Phonetics 32, 277–287 (2004)

84. Braun, A., Rietveld, T.: The influence of smoking habits on perceived age. In: Proc. of ICPhS 95. vol. 2, pp. 294–297. Stockholm (1995)
85. González, J., Carpi, A.: Early effect of smoking on voice: A multidimensional study. Medical Science Monitor 10, 649–656 (2004)
86. Wagner, A., Braun, A.: Is voice quality language-dependent? Acoustic analyses based on speakers of three different languages. In: Proc. of ICPhS 03, Barcelona, pp. 651–654 (2003)
87. Roach, P.: Phonetics. Oxford University Press, Oxford (2001)
88. Andersson, L.G.: Språket, Vetenskapsradion [radio programme]. Article retrieved 24 August 2006 (2006) from http://www.sr.se
89. Traunmüller, H., Eriksson, A.: Acoustic effects of variation in vocal effort by men, women, and children. Journal of the Acoustical Society of America 107, 3438–3451 (2000)
90. Pemberton, C., McCormack, P., Russell, A.: Have women's voices lowered across time? A cross sectional study of Australian women's voices. Journal of Voice 12, 208–213 (1998)
91. Decoster, W., Debruyne, F.: Longitudinal voice changes: facts and interpretation. Journal of Voice 14, 184–193 (2000)

Speech Under Stress: Analysis, Modeling and Recognition

John H.L. Hansen and Sanjay Patil

Center for Robust Speech Systems, University of Texas at Dallas, Richardson,
TX-75080 USA
john.hansen@utdallas.edu

Abstract. In this chapter, we consider a range of issues associated with analysis, modeling, and recognition of speech under stress. We start by defining stress, what could be perceived as stress, and how it affects the speech production system. In the discussion that follows, we explore how individuals differ in their perception of stress, and hence understand the cues associated with perceiving stress. Having considered the domains of stress, areas for speech analysis under stress, we shift to the development of algorithms to estimate, classify or distinguish different stress conditions. We will then conclude with revealing what might be in store for understanding stress, and the development of techniques to overcome the effects of stress for speech recognition and human-computer interactive systems.

Keywords: stress classification, pitch contours, Teager energy operator, robustness in speech recognition, Lombard effect, hidden Markov models, speech technology.

1 Introduction

Speech production involves a sequence of complex coordinated articulator movements, airflow from the respiratory system, and timing of the vocal system physiology. While speech is produced by changes in the articulator positioning, some utterances produced will not be similar in all respects for a speaker. This is because in many situations, the subject is under some type of emotional stress which will impact the utterance causing a deviation in the articulator movements. In human communications, listeners can handle or process these subtle changes far better than the automatic human-machine interface. We have yet to fully comprehend the aspects associated with stress and its effect on human speech production, perception and its impact on automatic speech systems. Thus, speech is a complex signal in a way that encodes information about the speaker, his/her state, acoustic environment, the person's intention, their language background, accent and dialect aspects, and further para-linguistic knowledge.

The main focus of this chapter is to define stress and then move on to show its impact on the speech production system, and thus on the speech systems used

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 108–137, 2007.

for speech recognition and speaker recognition. Specifically when considering robustness in speech recognition systems, the disparity between the training and test utterance significantly impacts performance. An attempt to understand the effect of stress on the human production system will certainly improve the performance of speech recognition systems as well as help in synthesizing speech to simulate emotions.

Before considering analysis and system development, it would be useful to define the "stress" elements of speech. Defining "stress" is a difficult problem because it represents a continuum and is not necessarily a binary decision. In general, a single definition cannot encompass all circumstances. Most definitions might be considered somewhat vague for practical uses. In spite of this, our definition will emphasize aspects from the science of linguistics – *emphasis given to a syllable*. Hence, we use the phrase, "speech under stress" to imply that the subject is speaking under some form of pressure which results in an alteration of the speech production process. The speech occurring under a condition which is devoid of stress, devoid of pressure is termed as "speech under neutral condition". Hence, stress is a psychological state that is a response to a perceived threat or task demand and is normally accompanied by specific emotions (e.g., fear, anger, anxiety, etc). These changes can affect speech behavior, even against an individual's will. Thus, any deviation in speech with respect to the neutral style, whether it is speaking style, word selection, word usage, sentence duration is termed as speech under stress. Therefore, speech under stressful conditions refers to speech spoken under some environmental factor or emotional state which perturbs speech production from a natural, conversational framework. There are many situations where the physical and mental conditions of a speaker can change. Some would include police/fire/ambulance personnel responding to emergency situations, military personnel in either peacekeeping or other military operations. Air traffic controllers represent another group who rely on voice communications in time sensitive stressful conditions. Stress is induced by high cognitive workload, sleep deprivation, frustration over contradictory information, emotion such as fear, pain, psychological tension, and other modern day multitasking conditions. Other areas with greater levels of emotions occur include:

1. Forensics – deception detection systems, analysis of 911 phone calls that can include threats [1,2].
2. Safety and Security – air traffic controllers and pilots in noisy high stress environments, deep sea divers, NASA-space explorations, power system operators, [1,3,4,5], military persons facing examination panel [6,7], law enforcement training [8].
3. Psychology – emotional state of patients [9,10].

There have been a number of studies on workload or cognitive task stress and efficiency in noisy environments [3,11].

As shown in Figure 1, speech production and the speaker are affected by various components which include stress caused by cognitive load or physical load, Lombard effect due to the noisy environment, accent change and language

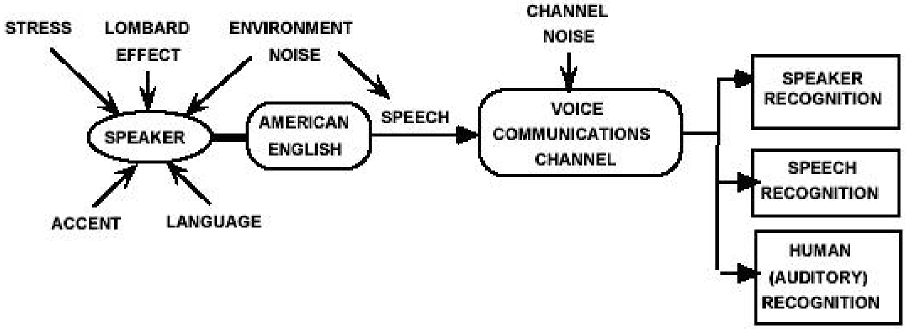


Fig. 1. Effect of stress and other components on speech and speech system

variability. These diverse factors or conditions degrade automatic speech system performance as well as human speech perception. We note that in adverse noisy, stressful situations where speech technology such as speech recognition, speaker verification, or dialog systems are used, addressing noise is not sufficient to overcome performance losses. In noisy stressful scenarios, even if noise could be completely eliminated, the production variability brought on by stress, including Lombard effect has a much more pronounced impact on speech system performance (as will be shown in this chapter).

As voice technology continues to mature, it becomes increasingly important to understand how stress and emotion influence speech production in actual environments. From a communications standpoint, it is clear that there are three distinct domains of speech under stress that include:

(i) physical speech production
(ii) hearing and human perception, including assessing if a subject is under stress, and,
(iii) speech system and technology – feature variation from the acoustic signal for speech and speaker recognition.

2 Domains of Speech Under Stress

2.1 Domain A: Production

Stress is a psychological state that is a response to a perceived threat or task demand and is accompanied by specific emotions (e.g., fear, anxiety, anger). The verbal indicators of stress could be identifying speech markers of stress (e.g., stuttering, repetition, and tongue-slip). Verbal markers of stress range from highly visible to invisible markers as perceived by the listener and that these markers are continuously monitored both consciously and subconsciously by the speaker and thus prone to correction [9].

Respiration is frequently a sensitive indicator in certain emotional situations. When an individual experiences a stressful situation, his respiration rate increases. This presumably will increase subglottal pressure during speech, which is known to increase fundamental frequency (or pitch) during voiced section [8]. An increased respiration rate also leads to shorter duration of speech between breaths which would affect the temporal pattern (articulation rate). The dryness of the mouth found during situations of excitement, fear, anger, etc., can also effect speech production (e.g., muscle activity of larynx and condition of vocal cords). Muscle activity of the larynx and vibrating vocal cords directly affect the volume velocity through the glottis, which in turn affects fundamental frequency. Other muscles (for example those controlling the tongue, lips, jaw, etc.) shape the resonant cavities of the vocal system and therefore do not have a direct influence on fundamental frequency, though they do contribute to changes in speech production under stress.

It is logical to postulate that if an individual is in a situation where the speed of task completion is critical (e.g., pilot flying an aircraft, air traffic controller), overall duration of utterances would also change under stressed conditions. It has also been suggested that under noisy conditions (Lombard effect), speakers vary their speech characteristics so that portions rich in information are emphasized, and those less important to intelligibility de-emphasized [3,12,13,14,15]. The control of vocal intensity is based on adjustments of laryngeal and subglottal variables. Speakers usually vary their intensity in typical speech production to affect suitable speech intelligibility for human communication [16].

Table 1. Quantifiable / Subjective cues in speech under stress. * The connection between the "observation/feature" and whether it is measurable implies that the stress component can be easily relayed by the speaker and perceived by the listener

OBSERVATION / FEATURE	MEASURABLE
Stuttering, repetition, tongue-slip, pauses between utterances, speed of word production	Quantifiable*
Energy, intensity, pitch (fundamental frequency)	Both quantifiable and subjective
Formant locations / structure (vocal tract), glottal structure (spectral slope), duration	Mostly subjective, but can be measured

Stress has a continuum of observability from the standpoint of the speaker and listener [19]. Changes in speech production which are clearly observable from the speaker, such as a dramatic increase in pitch, are equally observable to the listener. If a speaker wishes to conceal his/her stress, other production changes which are less observable may be altered instead (e.g., on roller coaster rides it is socially acceptable to scream, while in formal speech a speaker may try to maintain pitch and intensity, but adjust less observable markers such as glottal spectral slope). From a communication standpoint, stress could impact the physiological properties of production (i.e., a pilot or person on a roller coaster in high-G force physical movement), environmental factors such as background

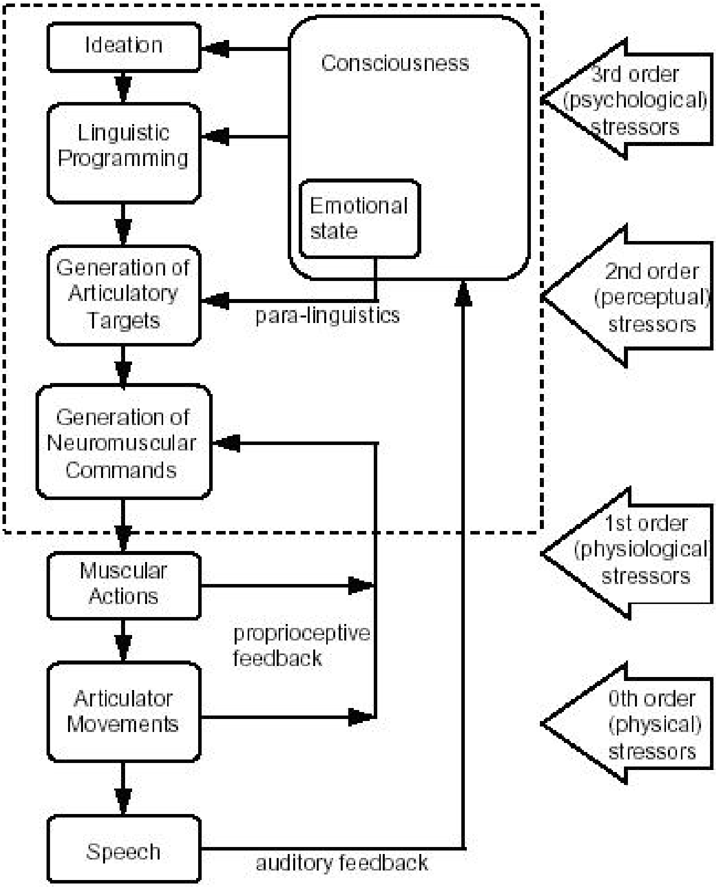

Ideation
Consciousness
3rd order (psychological) stressors
Linguistic Programming
Emotional state
Generation of Articulatory Targets
para-linguistics
2nd order (perceptual) stressors
Generation of Neuromuscular Commands
1st order (physiological) stressors
Muscular Actions
proprioceptive feedback
Articulator Movements
0th order (physical) stressors
Speech
auditory feedback

Fig. 2. Effect of stressors on speech production process [17,18]

noise (e.g., Lombard effect), or cognitive factors (e.g., person on the witness stand in a court trial) which could impact word selection.

Therefore, the speech production system can be affected by different stressors which play different roles in the formulation of speech production from word selection, grammar and sentence structure, and physical phoneme/word production. Figure 2 highlights the levels where speech communication/production occurs and their corresponding stress order [17,18]:

1. physical stressors – changes in the vocal apparatus caused due to vibration, movement or G-force, such that it directly affects the articulators.
2. Unconscious physiological stressors – stress causing changes in breathing rate or muscle tension. This might be caused by chemical effects, sleep deprivation or fatigue.

3. Conscious physiological stressors – stress causing increase in vocal effort, increase in voice so as to make oneself heard. This might be due to a noisy environment, experience, or emotion.
4. Internal stress feedback stressors – stress causing changes in vocal effort, mostly caused under the situation which might be interpreted as a threat to one's existence or some perceived conflict/threat.

2.2 Domain B: Perception

Research in speech quality and intelligibility has shown that consonant presence plays a major factor in a listener's ability to perceive the speaker's information content. Therefore, under stressed conditions, a speaker presumably may adjust consonant structure including increasing duration or intensity to give/emphasize additional acoustic cues to the listener. The most important part is that while lexical stress clearly influences duration, the listener will perceive the same utterance with different prosodic content across different stress conditions.

Listeners can identify the "stress markers" in the speaker's message even if these are not obvious. The listener will perceive the signal not merely based on the acoustic signal but using para-linguistics obtained in the context of the conversation, as well as based on his experiences [20]. Therefore, it is important that the speech is perceived in an appropriate manner and that a speaker should insert the appropriate cues within the signal as well as having the listener perceive the utterance in the proper way.

2.3 Domain C: Speech Systems

Speech systems including automatic speech recognition, automatic speaker recognition, speech synthesis, and speech coding / communication systems are all impacted by speech under stress [3,21,22,23]. In the presence of background noise, the speaker will alter his speech in order to communicate more effectively across a noisy environment (this is the Lombard effect). The effect of ambient noise has been suggested to be different for a male speaker versus female speaker [14,15]. In some cases, the situational stress or workload task stress will alter the speech, such as the case if a speaker is experiencing anger, fear, or a pilot flying an aircraft.

As described in Figure 1, the speech signal will be altered by cognitive stress, environmental noise, microphone mismatch which include the speaker, environment, and speech technology employed. These factors all impact the training conditions and therefore will deteriorate the speech systems' performance. If the system is trained with speech from one domain and another one is used for testing, the difference in frequency response causes degradation in the speech system performance. This frequency mismatch can be due to speech production changes caused by stress, microphone mismatch, or communication channel mismatch. Generally speaking, microphone or channel mismatch can reasonably be addressed with a static frequency compensation. Speech under stress however, will require a more intricate compensation scheme over the phoneme sequence.

As part of this chapter, we will focus on speech under stress which also includes the effect of the noisy environment on speech. If we consider a task such as speech recognition, speech under stress will impact robustness. However, for a speech synthesis system, the primary focus is to produce human-like speech, although the ability to impact stress or emotions associated with speech for the synthetic voice can be helpful for some applications. For speech coding, the coding system may not preserve the stress content of the speaker and make this part of communication less effective. For the task of speaker recognition, changes in speaker traits will be difficult to identify and address if the system is trained on neutral data. Hence, a hard binary decision on the type of and extent of stress associated with the speech signal can be helpful in developing more effective speech technology that can be employed in situations where stress / cognitive load / multi-tasking / physical stress / emotion is common place.

3 Analysis

If we consider the range of potential speech characteristics, which could be analyzed for speech under stress, fundamental frequency (or pitch) has historically been the most widely studied. Probably the most extensive early study that focused on analysis was Williams and Stevens [24], while an extensive number of studies have followed since that landmark contribution (see Table 2).

Over the last twenty years, CRSS and earlier variations of our group have performed an extensive level of research on the analysis of speech under stress, algorithm development for detection of stress, speech recognition under stress, and human perception of speech under stress. These studies have concentrated on the SUSAS corpus for the majority of the research [25]. More recently, we have considered other realistic conversational corpora including CU-Move (in-vehicle route navigation dialog), FLETC corpus (police/military training scenario), and UT-SCOPE (speech under cognitive and physical stress conditions) [20,23,26]. The comprehensive feature domains focus on speech production including: fundamental frequency, intensity, duration, formant locations, spectral slope, including an extensive range of features such as traditional MFCCs features and nonlinear TEO-based features.

3.1 Analysis of Fundamental Frequency

Characteristics of fundamental frequency (f_0) include contours, mean, variability, and distribution. A subjective evaluation of more than 400 f_0 contours was conducted across all stress conditions from SUSAS [21,27]. Although f_0 contours indicate excitation differences between styles, they do not reveal significance for particular variations. Moment analysis results, shown in Table 3 include a comparison in mean, variance, standard deviation, average deviation, skewness, and kurtosis. The Student t-test results applied to a pairwise comparison with f_0 data from above show that mean values deviate significantly from neutral as well as most other styles. Speaking styles such as loud and angry showed the

Table 2. Previous studies performed on Speech under Stress

Parameter studied	Analysis
Fundamental frequency contours and its variability	Stress conditions: anger, sorrow, fear [1,2,24]
Mean articulation rate in syllables	Anger, sorrow, fear [1,2,24]
Lombard effect speech	[14,15]
Vibration space shift rate (VSSR) from speech spectrograms	Fundamental frequency [3]
Monitoring heart rate and spectral centroid of first formant	Need for further research [20]
Pitch, amplitude, timing measurements	Elevated pitch and amplitude, and increased variation [5]

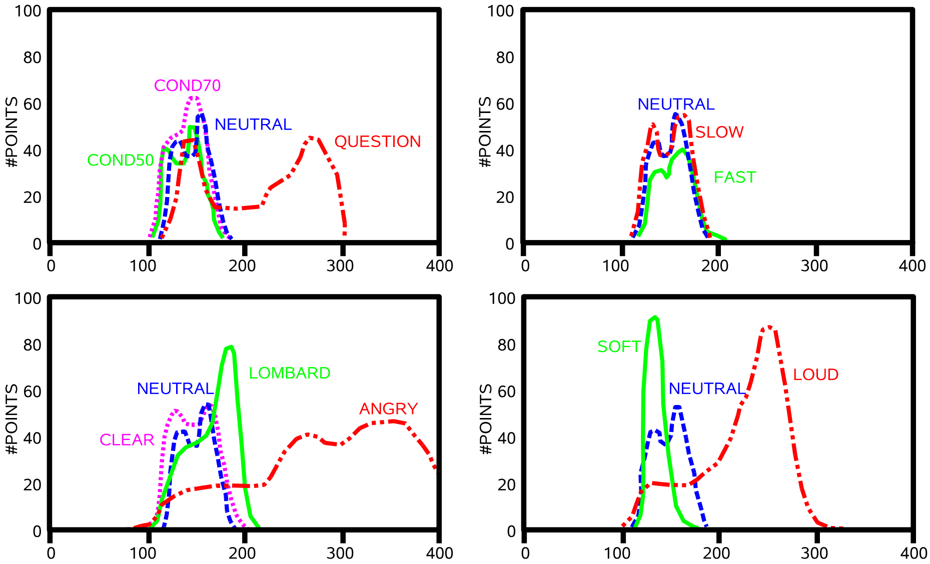


Fig. 3. Fundamental frequency (pitch) distributions across different speaking styles and stress conditions

widest deviation from neutral. Mean fundamental frequency is a good indicator over a wide variety of stress conditions. Loud, angry, and Lombard mean fundamental frequency are all significantly different from neutral as well as all other styles considered.

From Table 3 and Figure 3 and F-test statistical analysis, we can concur for most cases, f_o variance is shown to be significantly different from neutral as well as many other styles, and therefore is a good differentiating stress parameter. Pitch variance is not reliable for moderate versus high computer workload task (COND50 vs. COND70) conditions, and for slow and fast stress speaking conditions. Though the pitch distributions from Figure 3 are generally bimodal,

Table 3. Analysis of Fundamental frequency over various speaking styles and stress conditions

Stress Condition	Mean Value	Max. Value	Min. Value	Ave. Dev.	Stand. Dev.	Var.	Skew.	Kurt.
Neutral	142	182	116	13.4	15.4	239	0.22	-0.98
Slow	140	174	114	12.7	14.6	212	0.27	-1.10
Fast	149	186	121	11.9	14.0	195	0.19	-0.80
Soft	135	267	114	5.2	9.7	93	7.02	86.5
Loud	209	276	113	37.9	44.1	1944	-0.54	-0.97
Anger	283	400	96	44.3	56.3	3166	-0.38	0.44
Clear	150	211	103	19.1	22.1	489	0.23	-0.94
Cond50	140	205	111	13.7	16.2	263	0.41	-0.25
Cond70	143	216	111	13.7	16.3	266	0.34	0.01
Lombard	163	229	109	21.6	24.8	614	-0.25	-1.05

in certain stress styles (angry, question, soft, loud) the shape does deviate significantly from neutral (as measured by Kolmogorov-Smirnov pairwise test for distribution). We should note that contour shape of course plays a major role for question style. We therefore conclude that while a range of f_0 factors change in stress speaking styles, mean and variance can be effective traits for stress classification.

3.2 Analysis of Duration

In a previous study, it is shown that for stress conditions where time is of the essence, word duration, as well as subword durations such as changes in vowels versus consonants, and consonant presence, plays a major factor in a listeners' ability to perceive the speaker's information content [3,21].

As seen in Table 4, mean word duration as expected, increases for slow and decreases for fast spoken speech. The duration of consonants, semivowels, and diphthongs (to a lesser degree) remain constant in soft versus loud conditions, vowel duration decreases slightly for soft speech, and increases significantly for loud speech (as well as angry speech).

Upon more fine analysis, we observe significant changes in mean word duration for several stress conditions and proposed that it could be possible that overall word duration remains constant with shifts between consonant and vowel sections. Mean vowel and consonant duration possess similar discriminating abilities. Similarly, for variance, vowel and consonant classes continue to have reliable stress discriminating power.

Since vowels and consonants show major changes across all stress styles, several proposed discriminating features were proposed. The derived features are consonant versus vowel duration ratio (CVDR), consonant versus semivowel duration ratio (CSVDR), and vowel versus semivowel duration ratio (VSVDR) for all the stress styles. Table 5 summarizes the results which illustrates shifts in overall word duration, as well as movement between vowel, semi-vowel, and consonant classes [3,21].

Table 4. Word and Speech Class Duration over various speaking styles and stress conditions

Stress Condition	Mean Duration (msec)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	478	827	353	509	650	662	666	482	501	572
Vowel	160	294	115	147	253	271	202	148	147	198
Consonant	71	107	52	87	73	62	128	79	86	73
Semivowel	60	126	57	71	76	85	83	71	68	97
Diphthong	192	374	147	210	294	315	199	176	178	249
Stress Condition	Duration Variance (msec)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	18	49	12	16	28	41	40	16	14	24.0
Vowel	7.9	21	3.6	6.2	19	23	17	7.6	7.6	13.0
Consonant	1.8	7.1	1.1	2.9	3.7	3.3	10	2.5	3.3	2.6
Semivowel	0.7	7.1	1.0	1.3	2.9	7.8	3.2	1.7	1.4	4.3
Diphthong	3.3	14	1.0	1.1	5.6	7.0	3.3	2.4	3.4	3.5

Stress Style Key:

N – Neutral, Sl – Slow, F – Fast, So – Soft, L – Loud, A – Angry, C – Clear, C50 – computer task Cond50, C70 – computer task Cond70, Lom – Lombard effect

CVDR and CSVDR suggest that there is a shift in the percentage of time spent in vowels and semivowels towards consonants for soft, clear, and to a lesser degree the two computer task conditions (COND50 and COND70). These results indicate that the presence of stress influences word and individual phoneme duration characteristics.

3.3 Analysis of Intensity

Next, we consider analysis of intensity over stress speaking styles at the word level and phoneme levels. To focus the intensity analysis on the core portion of each phoneme, the phoneme boundary was reduced by 10% from both directions towards the phoneme mid-point, with RMS energy found for each phoneme and overall word. As expected, a marked increase resulted for loud and angry conditions, while soft, clear, and speech under the two computer task workloads had reduced word intensity. Vowel intensity remained constant for slow, clear and Lombard conditions, while consonant intensity increased for soft, angry, question, and speech under two computer task workloads (see Table 6). Word intensity possessed a good level of stress discriminating ability. However, experiments show that for several stress styles, such as angry, duration and intensity are interrelated. For detection of stress, mean RMS intensity is as successful

Table 5. Analysis of Duration over various speaking styles and stress conditions

Stress Condition	Analysis of Mean Duration (msec) and Ratios						
	Word	Vowel	Semivowel	Consonant	CVDR	CSVDR	VSVDR
Neutral	478	166	59.6	70.6	0.426	1.184	2.777
Slow	827	308	126	107	0.349	0.850	2.437
Fast	353	964	57.4	51.8	0.429	0.901	2.100
Soft	508	158	70.9	87.3	0.552	1.231	2.230
Loud	650	260	75.5	72.6	0.279	0.962	3.444
Anger	662	279	84.6	62.1	0.223	0.734	3.294
Clear	666	201	82.9	128	0.634	1.539	2.429
Cond-50	482	153	71.4	78.8	0.516	1.103	2.136
Cond-70	501	152	67.9	86.0	0.566	1.267	2.239
Lombard	572	207	97.3	73.1	0.353	0.750	2.214

as mean duration. Intensity variance across words or phoneme classes were not consistently successful for stress detection.

3.4 Glottal Pulse Shaping

The spectral based characteristics from the glottal source and vocal tract response are also impacted during speech production under stress. In this subsection, we focus on glottal source changes and in the subsequent section on vocal tract response characteristics. Glottal spectral source factors which include spectral slope, center of mass, and mean spectral level were analyzed as potential acoustic correlates of speech under stress [21].

For glottal flow spectra under all the ten stress conditions, the shape of glottal flow spectra are similar but differentiating features include spectral slope and amplitude [12]. Typically, for Lombard and angry styles, variability in spectral amplitude was observed in the 2 to 4KHz band. This generally implies a change in the shape of the glottal pulses under these conditions.

Using linear regression, the spectral tilt information was extracted across the glottal spectrum. Figure 4 summarizes that all speaking styles have a spectral tilt significantly different from neutral. Based on spectral characteristics, under certain stress conditions (loud, angry, and Lombard), glottal pulses will have steeper slopes with sharper glottal pulse corners (or irregular shapes) caused by a combination of changes in sub-glottal air pressure, vocal-fold tension, and uneven or sudden closure of the vocal folds during phonation. Alternatively, for slow and soft styles, glottal pulse shape will have gradual rise and fall times and overall smooth shapes, resulting in reduced energy in high frequency content and a steep spectral slope (-15dB/octave).

The analysis of glottal source spectrum revealed that parameters such as spectral slope and the distribution of energy to be important for relaying stress.

Table 6. Mean and Variance for Word and Speech Class Intensity over different speaking styles and stress conditions

Stress Condition	Mean Intensity (RMS)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	7663	7982	7812	7277	10561	11307	7067	7075	6934	8286
Vowel	9610	9692	9404	9326	12002	12700	9786	8857	8996	9699
Consonant	1394	1481	1425	1866	1164	1562	1287	1592	1715	1401
Semivowel	10032	9323	9983	10072	9443	11629	8272	8498	8353	8322
Diphthong	10125	9989	10460	9393	14800	14724	10394	9807	9742	10913
Stress Condition	Variance of Intensity (RMS)									
	N	Sl	F	So	L	A	C	C50	C70	Lom
Word	12.3	8.5	10.1	16.0	31.2	50.7	6.5	9.2	8.3	16.8
Vowel	93.3	76.5	93.4	63.6	116	193	109	92.2	101	80.4
Consonant	21.8	32.1	20.1	35.8	26.0	33.6	24.9	24.1	33.1	23.9
Semivowel	128	106	136	102	231	571	75.7	162	187	152
Diphthong	38.7	14.6	29.0	22.1	17.3	19.5	23.6	23.2	30.8	8.4

3.5 Vocal Tract Spectrum

To study the effect of speech under stress on the vocal tract spectrum, the mean, variance, and distribution of formant location and bandwidth across extracted phonemes were analyzed [21].

The previous research found shifts in frequency content for subjects performing a timed arithmetic task [12,28]. The results were more pronounced for front vowels than back vowels, with weaker third and fourth formants (reduced amplitudes) for the stress versus control conditions. We have found that when a speaker is under stress, typical vocal tract movement is effected, suggesting a quantifiable perturbation in articulator position.

Slow, loud, angry, and clear speaking styles show the widest shift in F1 formant locations. F2 formant frequencies generally increase among most conditions. Only slight changes occur for F3 and F4 locations across all styles. Formant bandwidth show large variations in mean for the first two formant frequencies with some changes for F3 and F4. The variance of formant location and bandwidth also showed shifts, with especially large changes in variance for loud, angry, and clear styles for F1. The changes in formant structure (location and bandwidth) are seen in Figure 5 and Figure 6 for Lombard, angry and neutral styles.

A series of Student T-tests were performed assuming both equal and unequal variance. Mean shifts in formant location (F1, F2, and F3) for loud, angry, and clear were significantly different from neutral. It was seen that styles which vary in formant location, will also increase formant variability in conveying that stress condition.

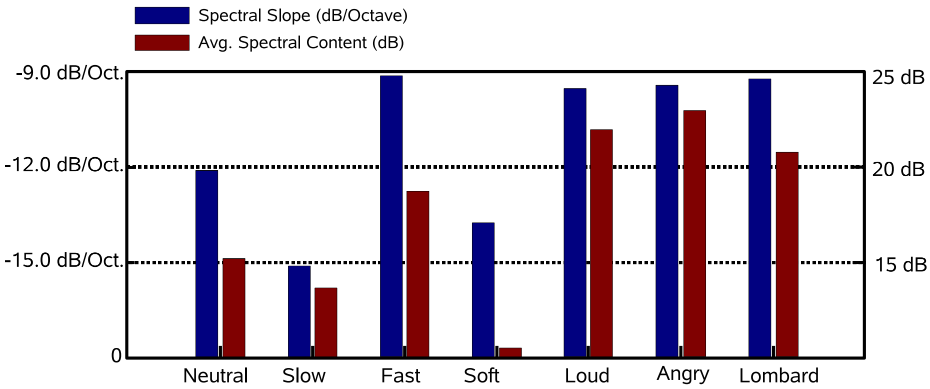


Fig. 4. Special Tilt (Glottal Pulse Shaping) Left Axis: spectral slope in dB/Octave, Right Axis: average spectral content in dB

Average formant location, average formant bandwidth and the variance of these all display varying degree of stress relayer information.

4 Applications

As discussed in the previous sections, many environmental and situational factors contribute to variation in speech production. Studies have shown that speech produced under stress causes significant loss in performance for traditional speech recognition algorithms. Stress and emotional characteristics must also be captured and modeled in order to produce more natural sounding speech coding and text-to-speech synthesis techniques. The importance in understanding how speakers vary their production systems to convey emotional or task induced stress has been shown in the previous section.

In the past, limited research has been conducted on the effect of stress on speech systems. Based on our investigations, the speech aspect that appears to provide the clearest indication of emotion or stress is fundamental frequency over time. Although important for the analysis of speech under stress, variation in pitch may not be a critical factor in attempting to reduce errors in traditional speech recognition algorithms. However, if the analysis of such parameters were to show statistically reliable indicators, it may be possible to formulate front end analysis procedures to identify periods of high stress. Recent studies demonstrate the potential for reliable stress classification via nonlinear, articulatory, and speech production features [9,24,29,30,31]. Once a period of speech under stress has been identified, a recognition system incorporating a compensation procedure specific to that form of stress could be used [32,33,34,35].

Although some variation in duration may not seriously affect speech systems, if the phonemes used for discrimination decreases in length, the probability of word misclassification can increase. Similar problems could arise for an increase

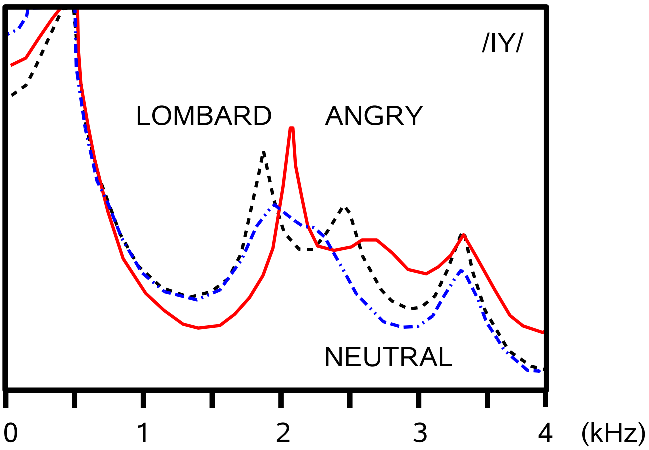


Fig. 5. Vocal tract spectrum for /IY/ phoneme

in duration in HMM modeling due to the finite number of states and numerical accuracy available in computing state transition probabilities.

Today, commercial based speech recognition systems can achieve more than 95% recognition rates for large vocabularies in restricted paradigms with relatively noise-free environments.

The issue of robustness in speech recognition can take on a broad range of problems. A speech recognizer may be robust in one environment and inappropriate for another. The main reason for this is that the performance of existing recognition systems which assume a noise-free tranquil environment (or train-test matched conditions), degrade rapidly in the presence of noise, distortion, and stress.

It is suggested that algorithms that are capable of detecting and classifying stress could be beneficial in improving automatic recognition system performance under stressful conditions. Furthermore, there are other applications for stress detection and classification. For example, a stress detector could be used to detect the physical and/or mental state of a pilot and that detection could put special procedures in place such as rerouting of communications, redirection of action, or the initiation of an emergency plan. To be able to detect and classify stress, it is necessary to understand the effect of stress on acoustical features.

There are two processing stages in a stress detection system. In the first stage, acoustical features are extracted from an input speech waveform. The second stage is focused on detection of stressed speech from neutral using one or more available methods.

A variety of methods exist for stress detection which include, but not limited to, detection-theory based methods, methods based on distance measures, and statistical modeling based techniques. A representative sample are presented in this section. These methods include:

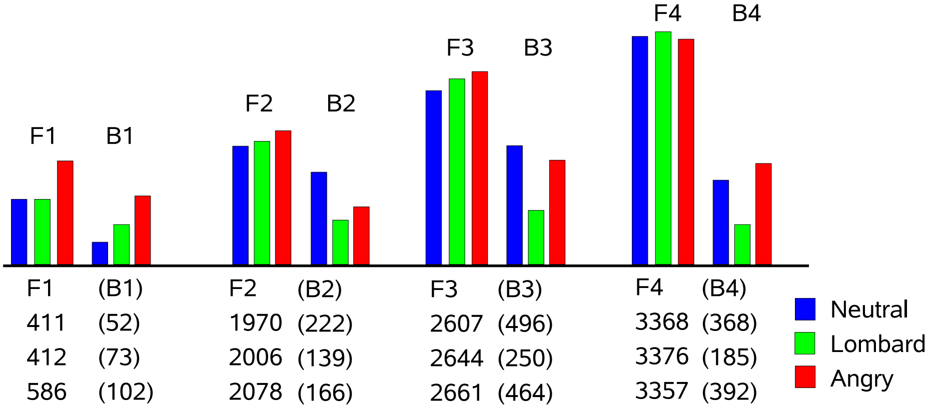


Fig. 6. Formant Frequency Location and Bandwidth value and distribution

(i) Neural Networks with linear speech model-features,
(ii) Optimum Bayesian detection used for stress classification,
(iii) TEO-based nonlinear speech features for both stress classification and stress assessment.

In the next section, we first focus on speech recognition in section 4.1 followed by stress detection methods in section 4.2.

4.1 Speech Recognition

To improve the performance of speech recognition systems in stress and noise, a number of methods have been considered including multi-style training, simulated stress token generation, training and testing in the same noise. While these methods help in matched conditions, the results degrade as test conditions drift from the base train condition. Some methods which address this drawback focus on estimation of speech features in noise, adapting speech enhancement techniques, and / or incorporating stress equalization [13,32,36]. The concept of stress equalization is based on a processing scheme which operates on a parameter sequence that is extracted from the input speech under stress. The stress equalization algorithm attempts to normalize the variation of the parameter sequence due to the presence of stress on the input speech signal.

Stress equalization techniques are a front-end processing approach to improve speech recognition under stress. The techniques can rely on maximum likelihood compensation factors to project the input stress modified features into a neutral-like space, where a neutral trained automatic speech recognition system is used. Figure 7 illustrates the impact of stress and noise on speech recognition performance [3,25]. We see that a basic speech recognition task in neutral, noise-free conditions is significantly impacted by the presence of stress (e.g., an average 31% reduction in recognition accuracy), and stress and noise combined (e.g., an average 58% reduction in recognition accuracy). Lombard, loud and angry stress styles in noise are significantly impacted. To address stress and noise, a combination two-tier approach was considered based on maximum likelihood stress compensation algorithm directly on the speech features, combined with noise suppression using Auto-LSP constrained iterative speech enhancement [3,37]. This combination scheme provided measurable levels of speech recognition improvement over noisy stressful conditions (see Figure 9, average accuracy improvement of 27%). A more rigorous stress compensation scheme developed for MFCC cepstral parameters was shown to have an even greater performance improvement in noisy Lombard effect speech [33].

Due to the extensive level of research activity in robustness for automatic speech recognition in stress and noise, it is not possible to consider even most of the advances over the past 15 years. The overview study [25] provides an effective and comprehensive step, and the interested reader is encouraged to consider the extensive bibliography at the end of this chapter. Our intension here is to provide a brief overview of the research topic in automatic speech recognition.

Another way to compensate for stress is to use a front-end artificial neural network. Figure 8 illustrates the use of an artificial neural network (ANN) for improving the performance under noisy stressful speech conditions. With a feature enhancement ANN (FE-ANN), a unique FE-ANN is created for each keyword model and further evaluated using a semi-continuous HMM recognizer followed by a likelihood ratio test for keyword detection [12,25,38]. The results show that a front-end ANN can provide consistent improvement for keyword recognition under Lombard effect.

A more rigorous method to address stress was based on morphological constrained feature enhancement with an adaptive cepstral stress compensation technique would be a third alternative studied for speech recognition systems [33]. Figure 10 shows the improvement achieved with MCE with adaptive mel-Cepstral Compensation. It should be noted that some features which are robust for speech recognition in noise, may not be as successful in stress and those successful in stress may not be as successful in noise. For example, linear predictive (LP) based MFCC features are more effective for speech recognition under stress versus FFT based MFCC, FFT based MFCCs are more successful for speech recognition in noise but performance decreases for speech under high stress conditions (angry, loud, etc.) [36]. The studies here suggested that effective speech features, compensation methods, and alternative training methods, can all lead to improved speech recognition performance in speech under stress.

4.2 Stress Detection

Since the range of speech under stress can include several broad types, the domains for stress detection in partitioned into the following four categories:

1. Speech under deception
2. Lombard effect detection

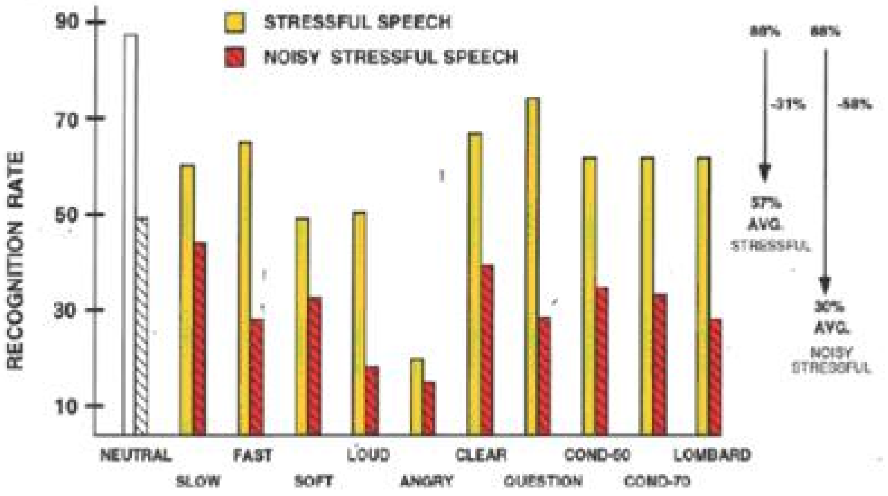


Fig. 7. Application of stress equalization for ASR - VQ-HMM ASR system

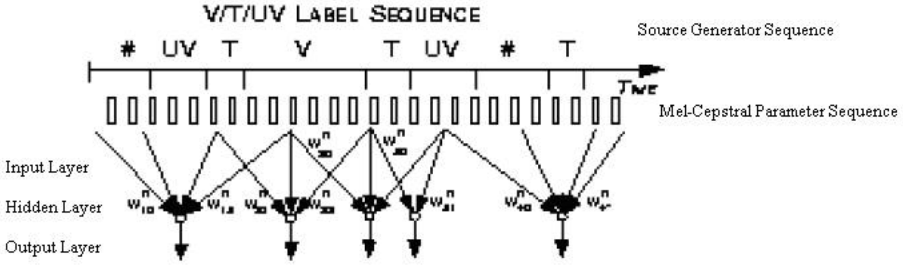


Fig. 8. FE-ANN for robustness of Speech Recognition Systems

3. Cognitive Stress detection
4. Physical Stress detection

Certain systems use voice stress analyzers based on microtremors, which have been shown to not be good indicators of stress [39,40].

4.3 Detection-Theory-Based Framework for Stress Classification

A Flexible framework for stress detection can be easily established using detection theory. For such a scheme, there are two hypotheses termed H_0 and H_1. Under H_0, the speech is neutral; while under H_1, the speech is stressed. Given an input speech feature vector, x, two conditional probability density functions (PDF), $p(x|H_0)$ and $p(x|H_1)$, must be estimated. With these PDFs, the likelihood ratio, λ, is defined as follows,

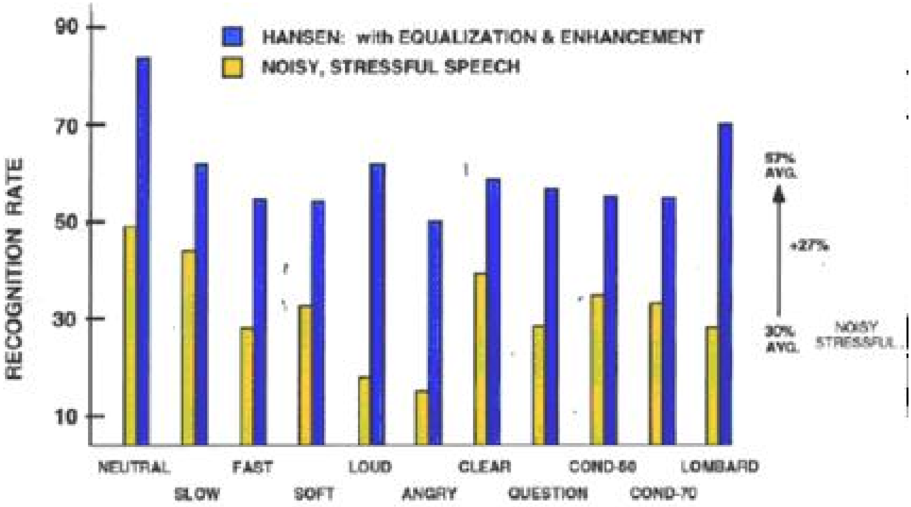


Fig. 9. Robust Recognition under noisy and stressful speech [3,27]

$$\lambda = \frac{p(x|H_1)}{p(x|H_0)} \qquad (1)$$

The decision of whether the input speech is neutral or stressed is made by comparing the likelihood or log likelihood ratio with a predefined threshold, β. If the ratio is larger than β, the input speech is detected as stressed; otherwise the input speech is classified as neutral. The value of β depends on the particular criterion used for detection.

4.4 A Distance Measure for Stress Classification

The detection of stress versus neutral speech can also be achieved using a distance measure. For a given input observation speech feature vector and two prior feature distributions (one for neutral, and one for stress), two distance measurements can be obtained: the distance between the given vector and neutral speech distribution, along with the distance to the stressed speech distribution adjusted for variance. This distance measure reflects the proximity of the input sequence to the distribution of general neutral or stressed speech feature data.

Previous CRSS studies have concluded that using individual speech features for stress detection show a range in detection performance as summarized in Table 7. Acoustical features such as duration, intensity, pitch, glottal source information, and formant locations for vowels were studied for stress detection performance using isolated words from the SUSAS corpus. The two methods for detection include a traditional binary hypothesis detection-theory method, and a dual PDF distance based method. Table 8 shows the results for stress detection performance as the number of feature observations for detection is increased

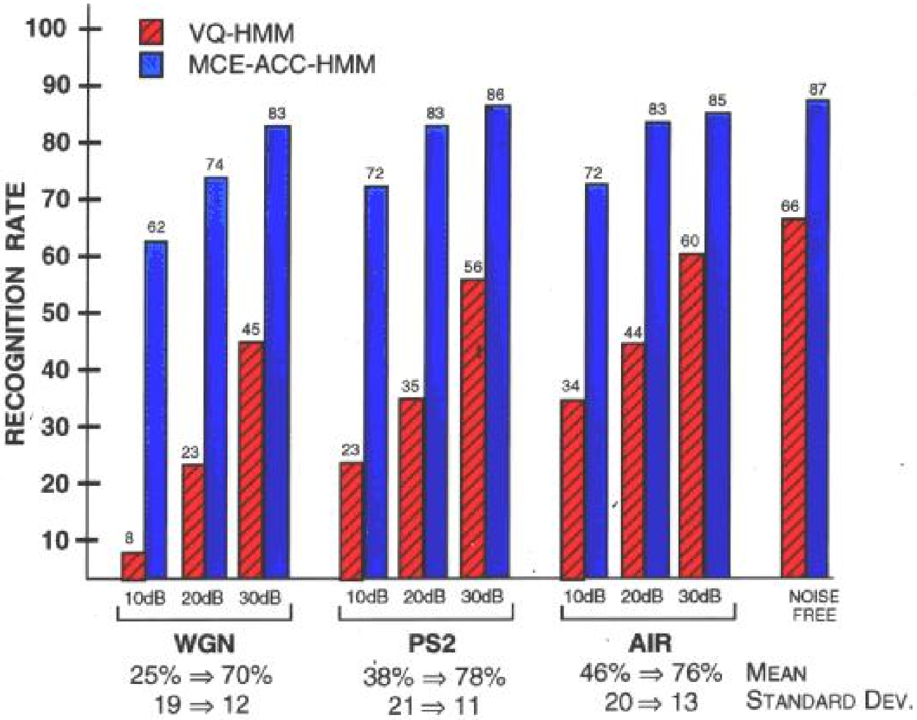


Fig. 10. MCE-HMM based Robust Speech Recognition system for stressful speech under noisy conditions

from 1 to 10 [41]. In general, given the error rate levels for the three stress classes tested, extensive experimental evaluation of stress detection at CRSS, we conclude the following:

1. Vowel duration is not a good feature for stress detection.
2. For intensity, increasing the input vector length does improve performance, especially for detecting angry and loud speech based on a detection-theory algorithm. As for the distance measure approach, increasing the input vector length does not always improve performance. The open-set test results also show that both methods perform better for detecting angry and loud speech versus detection for Lombard effect speech.
3. Compared to duration and intensity, pitch has much better performance for stress detection. Either of the methods perform similar with pitch feature.
4. The open-set results from the detection-theory-based method show that spectral slope (indicator of glottal source) is more suitable for detecting angry speech than for detecting loud speech with Lombard effect from neutral speech.
5. The features representing the vocal tract spectrum – formant location, are not suitable for stress detection.

Table 7. Stress Detection Studies using Traditional Features (Stress Conditions: Lombard, loud, angry)

Feature set	Stress/ Neutral Error Rates
Pitch	6-21 % variation
Glottal Spectral Slop	18-36 %
Intensity	18-36 %
Phone Duration	28-46 %
Formant Location	
1st Formant	38-46 %
2nd Formant	50-58 %
Feature Fusion	
Duration + Intensity + Mean Pitch	0-17 %

Table 8. Error Rates (%) of Open-set Pairwise Stress Classification using the combination of mean pitch, duration, and intensity as the feature

Vector Length	Speaking Style of Submitted Test Speech						Overall Error Rates	
	Neutral	Angry	Neural	Loud	Neutral	Lom.	Mean	Std. Dev.
1	17.68	17.03	11.67	11.97	19.85	21.21	16.5 %	3.97
5	6.15	5.00	4.62	4.62	13.08	13.08	7.76 %	4.16
10	1.67	0.00	3.03	3.03	13.64	16.67	6.34 %	6.98

The results from this section provide a representative perspective on the use of traditional speech production features for stress detection. Further studies have focused on the fusion of multiple features, and the interested reader is encouraged to explore the following references [6,7,8,9,17,30,31,41,42,43].

4.5 Neural Network Based Systems

Neural Network classifiers can also be employed for stress classification. A neural network based classification algorithm was considered for stress classification using cepstral-based features which have traditionally been employed for recognition [30]. Mel-cepstral parameters represent the spectral variations of the acoustic signal. It is suggested that such parameters are useful for stress classification since vocal tract and spectral structure vary due to stress.

Frame-based and word-level features performed in the ranged from 11-17 % for a 35 word test set which is greater than chance (9 % for eleven stress types in the SUSAS corpus). Most importantly, some stress conditions had reasonably good classification performance [30].

Another study considered the most effective feature subset for each targeted stress condition determined during a training phase emphasizing the most discriminating features (out of 27 studied) for classification of each stress style [30]. It has also been shown that a multi-dimensional HMM based system can be formulated which combined stress classification along with automatic speech recognition [44]. The resulting N-Dimensional HMM system resulted in a 73.8 % reduction in error rate as compared to the single channel stress dependent isolated word recognition system.

4.6 Stress Classification Using Nonlinear Speech Features

Next, stress classification can be considered from an alternative speech production modeling perspective. The assumption that airflow propagates as a plane wave in the vocal tract may not be the most accurate airflow model of speech production, since the flow is actually separated with concomitant vortices that are distributed throughout the vocal tract. Teager pioneered alternative approaches to speech modeling and also suggested that hearing could be viewed as the process of detecting the energy [45,46,47,48,49]. Over the past ten years, a number of studies have suggested that the so called Teager Energy Operator (TEO) can be employed to formulate new features for stress classification [6,7,8,41,42,43,50].

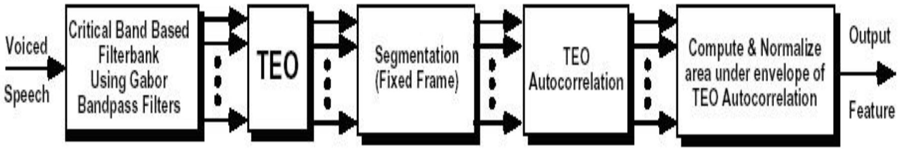

Voiced
Speech
Critical Band Based Filterbank Using Gabor Bandpass Filters
TEO
Segmentation (Fixed Frame)
TEO Autocorrelation
Compute & Normalize area under envelope of TEO Autocorrelation
Output
Feature

Fig. 11. Flow Diagram of TEO-CB-AutoEnv based feature

One of the effective nonlinear TEO-based feature developed by Zhou, Hansen and Kaiser [41,42,43,51] is the TEO operator, partitioned across a critical frequency band with an autocorrelation envelope analysis performed. The flow diagram for the TEO-CB-AutoEnv is shown in Figure 11. The theory is that the autocorrelation envelope is able to track the variability / regularity of the fine energy structure reflected in the TEO critical band partition, a trait which occurs in speech production for high stress conditions. Results from an evaluation using speech material from the SUSAS corpus is shown in Figure 12. Stress classification performance is significantly better than traditional MFCC based spectral features, or excitation based f_0 (pitch) information [41]. The performance is consistent for neutral versus emotion, speaking style, Lombard effect, and actual roller coaster ride speech under stress. The feature therefore is effective in both simulated and actual stress speech scenarios [41,42,43,51].

The same TEO-CB-AutoEnv feature has also been employed for stress detection in other scenarios. Figure 13 shows results for stress detection using data from a military examination task (SOM – Soldier of the Month Training). The results show as significant reduction over an MFCC feature based HMM baseline classifier using the new TEO-CB-AutoEnv feature [6] based on single word "no" decisions. Further experiments on this same SOM corpus have explored the impact of increased test token duration. The results from Figure 14 show

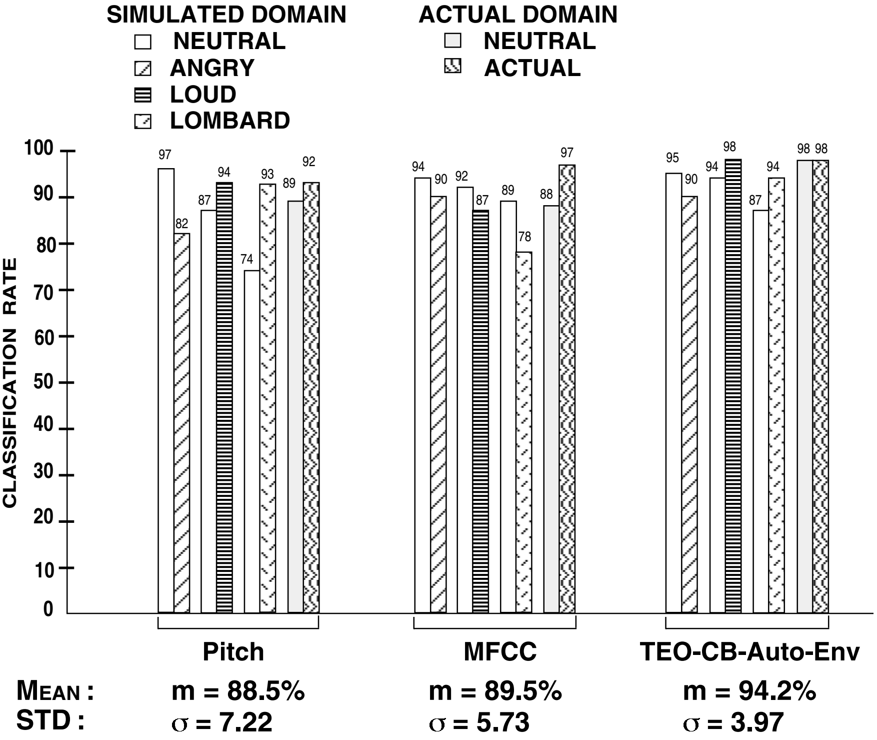


Fig. 12. Results comparing performance of TEO-based system with traditional features

that if 0-40 % of the vowel duration is removed, stress detection performance is maintained.

More recent studies employing real-life speech corpora such as US Army SOM (Soldier of the Month) or FLETC (law enforcement training scenario involving hostage rescue with weapons) have shown that TEO based features can be used for stress detection, as well as stress assessment over defined time periods [6,7,8,52,53].

4.7 Synthesis and Conversion of Speech Under Stress

As seen in our studies and elsewhere as well, the measured features that can reflect stress include changes in pitch and other excitation features, word/phoneme duration intensity, and spectral content. To activate the desired stress intonation for synthesized speech requires that the necessary variations in the actual stressed speech be represented in voice quality, pitch and duration of individual phonemes within the utterance [17,32,36,54,55,56,57,58,59]. This helps improve the naturalness of the synthetic speech. Previous approaches directed at integrating emotion in text-to-speech synthesis systems have concentrated on formulating a set of fixed rules to represent each emotion [58,59]. To represent a

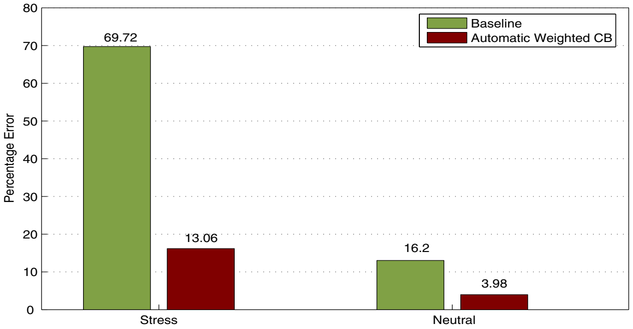


Fig. 13. Classification Error rates for Open Speaker Set (based on a single word “no” for SOM database)

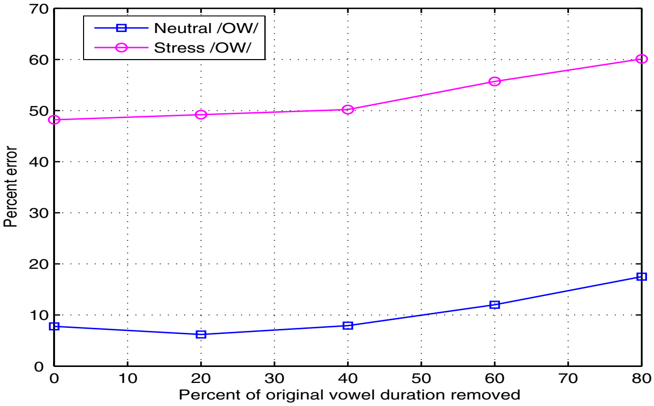


Fig. 14. Effect of vowel duration on stress classification error rates

range of variations for continuous speech, a fixed set of rules is not sufficient in general if we wish to have natural sounding speech.

It is possible to impart stress onto existing speech. One proposed method focused on converting CELP based excitation and vocal tract spectral structure from neutral to produce stress speech (Lombard, loud, angry speech styles) [32]. A subsequent study focused on HMM based modeling for voice conversion of neutral to stressed speech. The model showed it was possible to model stress perturbation techniques from one set of speakers and successfully impart these changes onto new neutral speakers [55].

Further studies have also explored the ability to impart emotion onto synthetic speech for text-to-speech applications [1,5,6,26,60,61,62,63]. These methods can be viewed as imparting a caricature or exaggerated version of the emotion/stress in order to make the emotion obvious to the listeners, and therefore generally do not always reflect true speech production changes that are more subtle. Further research is necessary to better understand speech under stress for synthesis, as well as perception of stress for synthetic speech applications.

4.8 Speech Coding System

As for speech coding, preserving the naturalness of the speech on the receiver side would help convey the emotional or stress state of the speaker. The stress perturbation algorithm for CELP coding system modified the pitch, gain, and the formant locations to convey the emotional state of the speaker [32]. The method along with hidden Markov model demonstrated conservation of speaking styles for isolated words under neutral, loud, angry and Lombard effect speaking conditions [32,55]. Future development of speech coding algorithms need to effectively capture the changes in speech production under stress. New advances in alternative excitation modeling based on GEMS or p-mike could offer improved techniques to encode speaker stress state for voice coding applications.

5 Discussions and Future Directions

As speech and language technology continues to mature, the need to effectively analyze, model, encode, detect, and classify speech under stress will increase significantly. Voice interactive systems including dialog and human-machine systems can benefit from knowledge of the speaker state. This information can help improve technology for speaker and speech recognition providing systems that are more effective in actual multi-task scenarios. The challenge, however, is to employ a framework which can provide effective analysis and modeling for improving such speech technology.

The source generator framework (SGF) proposed in [27,29,33] offers an effective means of modeling deviations from neutral to stress, and has been employed for a variety of stress equalization methods [17,25,29,33]. The basic structure is represented as:

$$(\text{speech})_{stress[X degree]}(\text{feature set}) = \varPsi[(\text{speech})_{neutral}(\text{feature set})] \qquad (2)$$

where $\varPsi$ [] is the transfer operator function which transforms neutral to stressed speech which has a certain degree of stress, say X. The above problem is two fold,

1. To define (in quantifiable sense) the degree of stress, X.
2. To define the speech production transfer operator $\varPsi[]$.

We model the transformation $\varPsi[]$ of the speech features in the neutral domain to an output stress domain. Prior formulation considered $\varPsi[]$ operators in the pitch, duration, intensity, glottal source, and vocal tract spectrum domains. It is important to recognize that if the transfer operator function is dependent only on the stress and phoneme, and generally independent of the speakers, it can be applied in more scenarios. An inverse transformation is therefore developed using this structure to compensate for the presence of stress. It is suggested that future advances in stressed speech processing could be realized using the Source Generator Framework, resulting in more effective speech and language technology with sustained performance in adverse speech/noise/environmental conditions.

References

1. Alm, C.O., Roth, D., Sproat, R.: Emotions from Text: Machine Learning for Textbased Emotion Prediction. In: Proceedings of HLT/EMNLP 05, Vancouver (2005)
2. Hollien, H.: Forensic Voice Identification. Academic Press, London (2002)
3. Hansen, J.H.L.: Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition. PhD thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta (1988)
4. Simpson, C.A.: Speech Variability Effects on Recognition Accuracy Associated With Concurrent Task Performance by Pilots. Technical report, Psycho-Linguistic Research Associates (1985)
5. Sproat, R., Olive, J.: Text-to-Speech Synthesis. In: Rabiner, L., Cox, R. (eds.) IEEE/CRC Press Handbook of Signal Processing, CRC Press, Cleveland (1997)
6. Prahallad, K., Black, A., Mosur, R.: Sub-Phonetic Modeling for Capturing Pronunciation Variation in Conversational Speech Synthesis. In: Proceedings of the 31th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), Toulouse (2006)
7. Ruzanski, E., Hansen, J.H.L., Meyerhoff, J., Saviolakis, G., Koenig, M.: Effect of phoneme characteristics on TEO Feature-based Automatic Stress Detection in Speech. In: Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), Philadelphia, vol. 1, pp. 357–360 (2005)
8. Rajasekaran, P.K., Doddington, G.R., Picone, J.W.: Recognition of Speech under Stress and in Noise. In: Proceedings of the 11th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86), Tokyo, pp. 733–736 (1986)
9. Cairns, D.A., Hansen, J.H.L.: Nonlinear Analysis and Detection of Speech under Stressed Conditions. Journal of the Acoustic Society of America 96(6), 3392–3400 (1994)
10. Dharanipragada, S., Rao, B.D.: MVDR-based Feature Extraction for Robust Speech Recognition. In: Proceedings of the 26th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01), Salt Lake City, pp. 309–312 (2001)
11. Whittmore, J., Fisher, S.: Speech during Sustained Operations. Speech Communications 20, 55–70 (1996)
12. Clary, G., Hansen, J.H.L.: A Novel Speech Recognizer for Keyword Spotting. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '02), Alberta, vol. 1, pp. 13–16 (1992)
13. Hansen, J.H.L., Bou-Ghazale, S.E.: Duration and Spectral Based Stress Token Generation for Keyword Recognition under Hidden Markov Models. IEEE Transactions on Speech & Audio Processing 3(5), 415–421 (1995)

14. Junqua, J.C.: The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognition. Journal of the Acoustic Society of America 93(1), 510–524 (1993)
15. Junqua, J.C.: The Influence of Acoustics on Speech Production: a Noise-Induced Stress Phenomenon known as the Lombard Effect. Speech Communication 20, 13–22 (1996)
16. Hicks, J.W., Hollien, H.: The Reflection of Stress in Voice-1: Understanding the Basic Correlates. In: Proceedings of the 1991 Carnahan Conference on Crime Countermeasures, pp. 189–195 (1981)
17. Hansen, J.H.L., Swail, C., South, A.J., Moore, R.K., Steeneken, H., Cupples, E.J., Anderson, T., Vloeberghs, C.R.A., Trancoso, I., Verlinde, P.: The Impact of Speech Under 'Stress' on Military Speech Technology. In: NATO RTO-TR-10, AC/323(IST)TP/5 IST/TG-01 (2000)
18. Murray, I.R., Baber, C., South, A.: Towards a Definition and Working Model of Stress and its Effects on Speech. Speech Communication 20, 3–12 (1996)
19. Goldberger, L., Breznitz, S.: Handbook of Stress: Theoretical and Clinical Aspects. Free Press, MacMilliam Pub., New York (1982)
20. Schreuder, M.J.: Prosodic Processes in Language and Music. PhD thesis, University of Groningen (2006)
21. Hansen, J.H.L.: Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition. In: IEEE Proceedings of the 15th Northeast Bioengineering Conference, Boston, pp. 31–32 (1989)
22. Paul, D.B.: A Speaker-Stress Resistant HMM Isolated Word Recognizer. In: Proceedings of the 12th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '87), Dallas, pp. 713–716 (1987)
23. Pickett, J.M.: The Sound of Speech Communication. University Park Press, Baltimore (1980)
24. Williams, C.E., Stevens, K.N.: Emotions and Speech: Some Acoustic Correlates. Journal of the Acoustic Society of America 52(4), 1238–1250 (1972)
25. Hansen, J.H.L.: Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition. Speech Communications, Special Issue on Speech Under Stress 20(2), 151–170 (1996)
26. Van Santen, J.: Prosodic modeling in Text-to-Speech Synthesis. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97), Rhodes, Greece, pp. 19–28 (1997)
27. Hansen, J.H.L.: Adaptive Source Generator Compensation and Enhancement for Speech Recognition in Noisy Stressful Environments. In: Proceedings of the 18th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93), Minn., pp. 95–98 (1993)
28. Hecker, M.H.L., Stevens, K.N., von Bismark, G., Williams, C.E.: Manifestations of Task Induced Stress in the Acoustic Speech Signal. Journal of the Acoustic Society of America 44, 993–1001 (1968)
29. Hansen, J.H.L., Cairns, D.A.: ICARUS: Source Generator based Real-Time Recognition of Speech in Noisy Stressful and Lombard Effect Environments. Speech Communications 16(4), 391–422 (1995)
30. Hansen, J.H.L., Womack, B.: Feature Analysis and Neural Network based Classification of Speech under Stress. IEEE Transactions on Speech & Audio Processing 4(4), 307–313 (1996)
31. Womack, B.D., Hansen, J.H.L.: Classification of Speech Under Stress using Target Driven Features. Speech Communication, Special Issue on Speech Under Stress 20(1), 131–150 (1996)

32. Bou-Ghazale, S.E., Hansen, J.H.L.: Stressed Speech Synthesis Based on a Modified CELP Vocoder Framework. Speech Communications: Special Issue on Speech Under Stress 20(2), 93–110 (1996)
33. Hansen, J.H.L.: Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect. IEEE Transactions on Speech & Audio Proc (SPECIAL ISSUE: Robust Speech Recognition) 2(4), 598–614 (1994)
34. Hansen, J.H.L., Bria, O.N.: Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise. In: Proceedings of the International Conference on Spoken Language Processing (ICLSP '90), Kobe, Japan, pp. 1125–1128 (1990)
35. Yapanel, U.H., Hansen, J.H.L.: A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition. In: Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03), Geneva, Switzerland, pp. 1281–1284 (2003)
36. Bou-Ghazale, S.E., Hansen, J.H.L.: A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress. IEEE Transactions on Speech & Audio Processing 8(4), 429–442 (2000)
37. Hansen, J.H.L., Clements, M.A.: Constrained Iterative Speech Enhancement with Application to Speech Recognition. IEEE Transactions on Signal Processing 39(4), 795–805 (1991)
38. Clary, G., Hansen, J.H.L.: Feature Enhancement for Multi-layer Perceptron and Semi-Continuous Hidden Markov Model Based Classifiers using Neural Networks. In: Neural and Stochastic Methods in Image and Signal Processing, Proceedings of the SPIE, vol. 1766, pp. 529–540 (1992)
39. Cestaro, V.L.: A Comparison between Decision Accuracy Rates obtained using the Polygraph Instrument and Computer Voice Stress Analyzer (CVSA) in the absence of Jeopardy. Technical report, DOD Polygraph Inst. (1995)
40. Eriksson, A., Drygajlo, A.: Forsensic Speech Science. In: Tutorial, 9th European Conference on Speech Communication and Technology (Interspeech 05 - Eurospeech) (2005)
41. Zhou, G.: Nonlinear Speech Analysis and Acoustic Model Adaptation with Applications to Stress Classification and Speech Recognition. PhD thesis, Dept. of Electrical and Computer Eng., Duke University (1999)
42. Zhou, G., Hansen, J.H.L., Kaiser, J.: Linear and Nonlinear Speech Feature Analysis for Stress Classification. In: Proceedings of the International Conference on Spoken Language Processing (ICLSP '98), Sydney, Australia, vol. 3, pp. 883–886 (1998)
43. Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator. In: Proceedings of the 23th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), Seattle, pp. 549–552 (1998)
44. Womack, B.D., Hansen, J.H.L.: N-Channel Hidden Markov Models for Combined Stress Speech Classification and Recognition. IEEE Transactions on Speech and Audio Processing 7(6), 668–677 (1999)
45. Kaiser, J.F.: Some Observations on Vocal Tract Operation from a Fluid Flow Point of View. In: Titze, I.R., Scherer, R.C. (eds.) Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control. Denver Center for the Performing Arts, Denver, pp. 358–386 (1983)
46. Teager, H.M.: Some Observations on Oral Air Flow during Phonation. IEEE Transactions Acoustic, Speech, Signal Processing 28(5), 599–601 (1980)
47. Teager, H.M., Teager, S.M.: A Phenomenological Model for Vowel Production in the Vocal Tract. In: Speech Science: Recent Advances, pp. 72–100 (1982)

48. Teager, H.M., Teager, S.: Evidence for Nonlinear Production Mechanisms in the Vocal Tract. In: NATO Advanced Study Inst. On Speech Production and Speech Modeling, Bonas, France, vol. 55, pp. 241–261. Kluwer Academic Publishers, Boston (1989)
49. Thomas, T.J.: A Finite Element Model of Fluid Flow in the Vocal Tract. Computer Speech Language 1, 131–151 (1986)
50. Hansen, J.H.L., Gavidia-Ceballos, L., Kaiser, J.F.: A Nonlinear based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment. IEEE Transactions on Biomedical Engineering 45(3), 300–313 (1998)
51. Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Nonlinear Feature Based Classification of Speech under Stress. IEEE Transactions on Speech & Audio Processing 9, 201–216 (2001)
52. Rahurkar, M., Hansen, J.H.L., Meyerhoff, J., Saviolakis, G., Koenig, M.: Frequency Band Analysis for Stress Detection Using a Teager Energy Operator Based Feature. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '02), Denver, vol. 3, pp. 2021–2024 (2002)
53. Ruzanski, E., Hansen, J.H.L., Meyerhoff, J., et al.: Stress Level Classification of Speech using Euclidean Distance Metrics in a Novel Hybrid Multi-Dimensional Feature Space. In: Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), Toulouse, vol. 1, pp. I–425–I–428 (2006)
54. Bou-Ghazale, S.E.: Analysis, Modeling, and Perturbation of Speech Under Stress with Applications to Synthesis and Recognition. PhD thesis, Robust Speech Processing Laboratory, Duke Univ. Dept. of Electrical Engineering (1996)
55. Bou-Ghazale, S.E., Hansen, J.H.L.: Stress Perturbation of Neutral Speech for Synthesis based on Hidden Markov Models. IEEE Transactions on Speech & Audio Processing 6(3), 201–216 (1998)
56. Cahn, J.: The Generation of Affect in Synthesized Speech. Journal of the American Voice I/O Society 8, 1–19 (1990)
57. Hansen, J.H.L., Clements, M.A.: Evaluation of Speech under Stress and Emotional Conditions. 82(S1), 7–8 (1987)
58. Murray, I.R., Arnott, J.L.: Implementation and Testing of a System for Producing Emotion-by-Rule in Synthetic Speech. Speech Communication 16, 369–390 (1995)
59. Murray, I.R., Arnott, J.L.: Synthesizing Emotions in Speech: is it time to get excited? In: Proceedings of the 4th International Conference on Spoken Language Processing (ICLSP '96), vol. 3, pp. 1816–1819. Philadelphia (1996)
60. Black, A.: Multilingual Speech Synthesis. In: Schultz, T., Kirchhoff, K. (eds.) Multilingual Speech Processing. Elsevier, Academic Press (2006)
61. Picard, R.W., Klein, J.: Computers that Recognize and Respond to User Emotion: Theoretical and Practical Implications. Interacting with Computers 14(2), 141–169 (2002)
62. Sproat, R. (ed.): Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer Academic Publishers, Boston (1997)
63. Van Santen, J., Kain, A., Klabbers, E.: Synthesis by Recombination of Segmental and Prosodic information. In: Proceedings of the International Conference on Speech Prosody, Japan, pp. 409–412 (2004)
64. Bachrach, A.J.: Speech and its Potential for Stress Monitoring: Monitoring Vital Signs in the Divers. Technical report, Naval Medical Research Institute (1979)
65. Chen, Y.: Cepstral Domain Talker Stress Compensation for Robust Speech Recognition. IEEE Transactions on Acoustic Speech Signal Process. 36, 433–439 (1988)

66. Darby, J.K.: Speech Evaluation in Psychiatry. Grune and Stratton, New York (1981)
67. Flack, M.: Flying Stress. Medical Research Committee, London (1918)
68. Hansen, J.H.L.: Analysis and Compensation of Noisy Stressful Speech for Environmental Robustness in Speech Recognition (invited tutorial). In: NATO-ESCA Proc. Inter. Tutorial & Research Workshop on Speech Under Stress, Lisbon, Portugal, pp. 91–98 (1995)
69. Hansen, J.H.L., Bou-Ghazale, S.E.: Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97), vol. 4, pp. 1743–1746. Rhodes, Greece (1997)
70. Hansen, J.H.L., Mammone, R., Young, S.: Editorial for the special issue: Robust Speech Recognition. IEEE transactions on Speech & Audio Processing 2(4), 549–550 (1994)
71. Hansen, J.H.L., Gavidia-Ceballos, L., Kaiser, J.F.: A Nonlinear based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment. IEEE Transactions on Biomedical Engineering 45(3), 300–313 (1998)
72. Hollien, H., Hicks, J.W.: The Reflection of Stress in Voice-2: the Special Case of Psychological Stress Evaluators. In: Proceedings of the 1991 Carnahan Conference on Crime Countermeasures, pp. 196–197 (1991)
73. House, A.S.: On Vowel Duration in English. Journal of the Acoustic Society of America 33(9), 1174–1178 (1962)
74. Kuroda, I., Fujiwara, O., Okamura, N., Utsuki, N.: Method for Determining Pilot Stress Through Analysis of Voice Communications. In: Aviation, Space, and Environmental Medicine 528–533 (1976)
75. Kaiser, J.F.: Some Useful Properties of Teager's Energy operator. In: Proceedings of the 18th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93), Minn., vol. 3, pp. 149–152 (1993)
76. Kaiser, J.F.: On a Simple Algorithm to Calculate the Energy of a Signal. In: Proceedings of the 15th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90), Albuquerque, New Mexico, pp. 381–384 (1990)
77. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S.: Approaching Automatic Recognition of Emotion from Voice: A rough Benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion, Belfast (2000)
78. Malkin, F.J., Christ, K.A.: Human Factors Engineering Assessment of Voice Technology for the Light Helicopter Family. Technical Report I-20, U. S. Armu Human Engineering Lab. (June 1985)
79. Maragos, P., Kaiser, J.F., Quatieri, T.F.: On Amplitude and Frequency Demodulation using Energy Operators. IEEE Transactions on Signal Processing 41, 1532–1550 (1993)
80. Poock, G.K., Armstrong, J.W.: Effect of Operator Mental Loading on Voice Recognition System Performance. Technical report, Naval Postgraduate School (1981)
81. Poock, G.K., Armstrong, J.W.: Effect of Task Duration on Voice Recognition System Performance. Technical report, Naval Postgraduate School (September 1981)
82. Schreuder, M., Eerten, L.v., Gilbers, D.: Music as a Method of Identifying Emotional Speech. In: Proceedings of the Workshop on Corpora for Research on Emotion and Affect (LRE '06), Genua, Italy, pp. 55–59 (2006)
83. Simonov, P.V., Frolov, M.V.: Analysis of the Human Voice as a Method of Controlling Emotional State: Achievements and Goals. Aviation, Space, and Environmental Sciences 23–25 (1977)

84. Streeter, L.A., MacDonald, N.H., Apple, W., Krauss, R.M., Galotti, K.M.: Acoustic and Perceptual Indicators of Emotional Stress. Journal of the Acoustic Society of America 73(3), 917–928 (1988)
85. Varadarajan, V., Hansen, J.H.L., Ikeno, A.: UT-SCOPE - A corpus for Speech under Cognitive/Physical Task Stress and Emotion. In: Workshop on Corpora for Research on Emotion and Affect (LREC '06), pp. 72–75 (2006)
86. Varadarajan, V., Hansen, J.H.L.: Analysis of Lombard effect under Different types and levels of Noise with Application to In-set Speaker ID systems. In: Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '06 –ICSLP), Pittsburgh (2006)
87. Womack, B., Hansen, J.H.L.: Robust Speech Recognition via Speaker Stress Classification. In: Proceedings of the 31th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), Toulouse, vol. 1, pp. 53–56 (2006)
88. Yamada, T., Hashimoto, H., Tosa, N.: Pattern Recognition of Emotion with Neutral Network. In: Proc. 21st Inter. Conf. on Industrial Electronics, Control, and Instrumentation (IECON '95), vol. 1, pp. 183–187 (1995)
89. Yapanel, U.H., Dharanipragada, S.: Perceptual MVDR-based Cepstral Coefficients for Noise Robust Speech Recognition. In: Proceedings of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong-Kong (2003)

Speaker Characteristics and Emotion Classification

Anton Batliner[1] and Richard Huber[2]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg, Martensstr. 3,
91058 Erlangen, Germany
batliner@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de
[2] Sympalog Voice Solutions GmbH, Karl-Zucker-Str. 10,
91052 Erlangen, Germany
huber@sympalog.de

Abstract. In this paper, we address the — interrelated — problems of speaker characteristics (personalization) and suboptimal performance of emotion classification in state-of-the-art modules from two different points of view: first, we focus on a specific phenomenon (irregular phonation or laryngealization) and argue that its inherent multi-functionality and speaker-dependency makes its use as feature in emotion classification less promising than one might expect. Second, we focus on a specific application of emotion recognition in a voice portal and argue that constraints on time and budget often prevent the implementation of an optimal emotion recognition module.

Keywords: emotion, automatic classification, acoustic features, speaker dependency, laryngealization, voice application, system architecture.

1 Introduction

The modelling, generation, and recognition of emotion has attracted more and more attention during the last years. Most of the time, researchers have typically dealt with prototypical, 'full-blown' emotions and with elicited, prompted, acted speech [1]. Normally, some of the 'big', full-blown emotions have been modelled and classified such as *anger, joy, despair, sadness*. Recognition rates reported are fairly high; [2] for instance report for seven emotions classification rates of up to 71.0% for speaker-independent and 92.7% for speaker-dependent modelling. Nowadays, the voice business is more and more attracted by the possibilities the recognition of user states offers for commercial systems. One main focus of interest is telephony based dialogue systems with spoken input in the broad area of customer care and customer service applications.

One of the general problems is that real life data differ, however, considerably from acted speech. It is way more difficult to collect the data, cf. Labov's well-known observer's paradox [3]: for recording, the subjects have to be observed but if they are aware of that, they are no longer fully spontaneous. Moreover,

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 138–151, 2007.

ethical issues have to be taken into account. To act does not mean the same as to behave: acting refers to a shared concept of emotion expressions — how you imagine someone should behave if they are angry, sad, etc. But in real life, neither the reference is fully clear — is the subject really angry even if we wanted to make them angry with our experimental design — nor the means of expressing specific emotions. In addition, the full range of pure emotions cannot be observed in real life encounters; instead, most of the data are not marked, i.e., neutral, and the non-neutral states are rather emotion-prone/affective in a broader meaning. Last but not least, speaker characteristics can superimpose emotion expressions or interfere with them. Specific applications need specific emotion modelling: for instance, in call center scenarios, we either look for a chance in the user's emotional state or or for a difference in the emotional state of one certain user in contrast to an average application caller.

For the time being, the speaker-independent automatic recognition of emotional user states in realistic, spontaneous speech seems to be 'fossilized' at approx. 80% class-wise computed recognition rate for a 2-class problem, and at approx. 60% for a 4-class problem, cf. [4]. Of course, higher classification performance can be obtained by fine-tuning, for instance, by pre-selection of prototypical cases, cf. [5]. We don't know of any speaker-dependent classifications for realistic, emotional spontaneous speech yet. The reason for that might simply be that it is difficult to collect enough data for one and the same subject because normally, subjects are 'burned' when they have participated in an experiment. And the reason for the low speaker-independent classification performance might be that individual speakers employ different acoustic features in a different way; moreover, features can be multi-functional, and interlabeller agreement is — for spontaneous speech — not very high.

Note that the figures given above are for carefully designed experiments, manually annotated, realistic (real-life, spontaneous) data, speaker-independent modelling, and rather good acoustic conditions. Depending on signal-to-noise ratio and degree of spontaneity, much higher or lower classification rates can be imagined: in a personalized setting (speaker-dependent modelling) with a close-talk microphone in a quiet office surrounding, if the speaker only has to produce a limited amount of commands, and if it is clear when and that they are getting angry, recognition rates well above 90% for two or even more classes can be imagined. On the other hand, in a public, noisy setting with a room microphone, free speech, and speaker-independent modelling, classification performance could drop almost to chance level. This also can happen if you switch to telephony applications where the communication channel is of restricted bandwidth; here the input quality is sometimes rather poor — just think of mobile phone calls. On the other hand, emotion recognition might not be that prone to noise as other speech processing tasks [6].

In this paper, we start with discussing automatic recognition of emotion and user states on a conceptual level. We address some basic challenges and possible reasons why the approaches until now have not been fully successful. Then we report on experiences made in a project where emotion recognition was integrated and applied in a real business environment; constraints in time and budget made it impossible, however, to implement an optimal emotion recognition module.

2 Setting the Scene

2.1 Concepts: Emotion and Speaker Characteristics

In this paper, we use the term 'emotion' in a broad sense, encompassing emotional (affective) user states such as bored, interested, stressed, despaired, perplexed as well. Other terms used for such additional states is 'interpersonal stances' [7] or 'social emotions' [8]. As for speaker characteristics, we want to focus on acoustic features because this field has been more investigated than linguistic features. We do not know of any study dealing with spontaneous, real-life speech, emotions, and in-depth description of speaker-specific traits. Thus we decided to demonstrate the possible impact of speaker-specific characteristics on emotion classification with a sort of 'gedanken experiment': how a specific phenomenon (irregular phonation, 'laryngealization', cf. below) can affect emotion recognition.

2.2 Two Different Worlds: Generation and Analysis

Synthesis of emotion uses controlled data based on acted speech, and models normally one speaker and/or the same segmental structure, focussing on forced choices in listening experiments for evaluation. Realistic emotion recognition deals with uncontrolled, i.e. spontaneous data based on many speakers, uncontrolled segmental structure and wording; as computation of features, esp. for large databases, is done automatically, extraction errors have to be accepted whose extent can only be estimated roughly.

2.3 Personalization and Data Acquisition: A Problem

Although it had been desirable to develop speaker-independent automatic dictation systems, they have been more or less speaker-dependent (speaker-adaptive) for the last decades. Only the latest versions claim to be really speaker-independent, i.e. a training phase should no longer be necessary. It might be astonishing that for such a complicated problem as emotion recognition, almost all of the studies on emotion recognition in spontaneous speech used speaker-independent modelling. We believe that two factors have been responsible for that: first, the whole speech processing community is oriented towards speaker-independence. Second — and maybe most important in our context — it is difficult to collect enough emotional data from one and the same person, cf. above. We are thus faced with a dilemma: personalization seems to be the only way out towards higher classification performance, but it is way more difficult to obtain than in the case of automatic dictation systems where only subjects are needed with enough patience to read longer stretches of text.

2.4 A Tentative Relevance Hierarchy for Speaker-Independent Emotion Recognition in Spontaneous Speech

In this subsection, we want to set up a tentative hierarchy of relevance in speaker-independent emotion recognition in spontaneous speech — as a sort of null hypothesis to be tested in further experiments. This hierarchy is based on own experimental results and on some other studies. Several caveats have to be made: most of the studies on relevant features used acted data; these are not taken into account. The next point is trivial but important: statements on relevant features can only be made on those features that were computed for the respective databases. In some studies, only few features or only features of a certain type are computed; as for other types, no statement can be made. On the other hand, if too many features are computed — nowadays, a set of basic features is often multiplied via different normalization and transformation procedures — it is often not easy to tell apart important from spurious information. And last but not least, it depends of course on the type of data — and by that, on the emotion classes annotated — the features are computed for. Hopefully, results will converge in the future.

Most relevant so far seem to be duration and Mel Frequency Cepstral Coefficient (MFCC) features, then energy and pitch variation (jitter, mean square error of regression). 'Genuine' pitch features such as F0 maximum and minimum — and by that, range — are not that important. MFCC features are 'implicit' spectral features which, however, encode linguistic information as well: they are standard features in word recognition. It is thus difficult to disentangle spectral information itself from word information. Linguistic information depends heavily on the type of data: for uniform speech such as commands, it should not be relevant. On the other hand, it is easy to imagine a full encoding with word information *(this makes me happy/sad/angry/...)*. 'Explicit spectral' information on formant band-width or voice quality and/or phonation type can sometimes tell apart specific user states but are, on the whole, less relevant than one should suppose on the basis of acted speech or perception experiments with synthetic speech.

Note that all this is tentative and based only on some few real-life, spontaneous databases. Anyway, if it proves to be true then two points are more puzzling than the other ones in the above given hierarchy, namely that F0 is not that relevant, and that voice quality and/or phonation type is not that relevant, either. We can imagine two different reasons why: first, dimensionality, second, multi-functionality. Duration and energy are *one-dimensional*: duration on the x- (time-) axis — longer or shorter — and energy on the y- (loudness/decibel) axis — higher or lower. Even if, under certain circumstances, short duration and low energy can encode prominence, at the very most, it is the other way round. (Note that we are speaking here of 'prominence' in a broad meaning, not only of prominence denoting stress/accentuation.) Therefore, we will call these two parameters one-dimensional. F0, however, behaves differently: it is not only high vs. low pitch, it is the whole configuration, i.e. specific tunes, which are prominent. And it might be the same problem for emotion encoding as for accent encoding: in the tone sequence terminology, accents can be marked by L*H or H*L, i.e. by two 'opposite' configurations, whereas almost never, accents are marked by short duration or low energy. Therefore, we will call F0 features *bi-dimensional*. In the next chapter, we will give an example for the multi-functionality of voice quality and phonation type features.

Normally, for emotion classification, acoustic features are extracted automatically by, for instance, doing forced alignment on the spoken word chain. Thus, segmentation is not perfect, and automatic extraction is error prone. Under real-life conditions, if the spoken word chain is not known, there might be more and/or different types of segmentation errors. We do not know much yet about the extent of such extraction errors; as for F0, the 'technical' errors amount at least to some few percent points, even under optimal conditions. Often, error rates are higher. (In the emotional database described and processed in [4], octave errors amount to some 6 % of all voiced parts in the words.) In addition, it is not clear yet whether extraction should be close to the signal or close to perception, esp. in the case of irregular phonation, cf. below. The impact of erroneous extraction on emotion recognition is even less clear. It might be the case that MFCCs are that good even at emotion recognition because they are a coarse but robust measure, whereas 'explicit' spectral and voice quality measurements are more error prone.

3 An Example: Laryngealizations

The normal speech register 'modal voice' comprises an F0 range from about 60 to 250 Hz for male speakers and an F0 range from about 120 to 550 Hz for female speakers. Below this register there is a special phonation type whose mechanisms of production are not totally understood yet and whose linguistic functions are not much investigated until now. There is a variety of different terms for this phenomenon which are used more or less synonymously: irregular phonation, creak, vocal fry, creaky voice, pulse register, laryngealization, etc. We use laryngealization (henceforth LA) as a cover term for all these phenomena that show up as irregular voiced stretches of speech. Normally LAs do not disturb pitch perception but are perceived as suprasegmental irritations modulated onto the pitch curve. Although LAs can be found not only in pathological speech but also in normal conversational speech, most of the time they were not objects of investigation but considered to be an irritating phenomenon that has to be discarded. In [9], five different types of LAs have been established: glottalization, damping, diplophonia, sub-harmonic, and aperiodicity. Voice quality and phonation types such as LAs are known to be utilized in the generation of emotions. We have to keep in mind, however, that the bulk of evidence so far has been obtained from acted speech or from perception experiments with synthesized speech.

Table 1 displays different functions of LAs which can be linguistic or paralinguistic. They can be caused either by higher effort or by relaxation; in the first case, they go together with *accentuation* (prominence) which is, of course, a *local* phenomenon. A typical place for relaxation is *the end of an utterance*;

Table 1. Different Functions of Laryngealizations

phenomenon	**time domain**
linguistic functions: phonotactics, grammar, ...	
accentuation	local
vowels	local
word boundaries	local
native language	local
the end of an utterance, i.e., turn-taking	local
paralinguistic functions: speaker characteristics	
speaker idiosyncrasies	local - global
speaker pathology	global
too many drinks / cigarettes	temporary
competence / power	global / temporary
social class membership	local / global / temporary
emotional states such as boredom, sadness, etc.	short-term or temporary

by that, *turn-taking* can be signalled to the dialogue partner; this is again a *local* phenomenon: [10] report that different types of LAs are used in (British and American) English conversations for holding the floor (filled pauses with glottal closure, no evidence of creaky phonation) and for yielding the floor (filled pauses with lax creaky phonation, no glottal closure). *Word boundaries* in the hiatus, i.e. word final vowel followed by word initial vowel, can be marked by LAs. Boundary marking which is, of course, *local*, with such irregular phonation is dealt with in [11] and [12]. It is well known that back *vowels* such as *[a]* tend to be more laryngealized than front vowels such as *[i]* (*local* phenomenon). A language-specific use of LAs can be either due to phonotactics, as in German, where every vowel in word-initial position is 'glottalized', or phonemes can be creaky, cf. [13]; this is a *local* phenomenon, denoting the *native language*. Normally, specific segments which are laryngealized characterize languages, cf. for vowels [14]; the Danish glottal catch (stød) [15] can be found in vowels and consonants.

[16] p. 194ff. lists different uses and functions of 'creak' phonation, amongst them the paralinguistic function 'bored resignation' in English RP, 'commiseration and complaint' in Tzeltal, and 'apology or supplication' in an Otomanguean language of Central America. Extra- and paralinguistically, LAs can be a marker of personal identity and/or social class; normally, LAs are a marker of higher class speech. [17] quote evidence that not only for human voices but for mammals in general, 'non-linear phenomena' (i.e. irregular phonation/LA) can denote individuality and status (pitch as an indicator of a large body size and/or social dominance; *"... subharmonic components might be used to mimic a low-sounding voice"*).

Note that all these characteristics which per se are **not** characteristics of single speakers can — maybe apart from the language-specific phonemes — be used more or less distinctly by different speakers. As for the paralinguistic function of LAs, speakers can simply use them throughout to a higher extent; such *speaker idiosyncrasies* are *local - global*. ‘Creaky superstars’ like Tom Waits are well-known. The reason might be unknown, or due to one or more of the following factors: *speaker pathology* (*global*), *too many drinks/cigarettes* (*temporary*), *competence/power* (*global / temporary*), or *social class membership* (*local/global/temporary*).

Emotional states such as *despair, boredom, sadness*, etc. are *temporary.* Bad news are communicated with breathy and creaky voice [18], boredom with lax creaky voice, and to a smaller extent, sadness with creaky voice [19]. [20] report for perception experiments with synthesized stimuli that disgust is conveyed with creaky voice. To display boredom or to display upper-class behaviour might coincide; the same can happen if someone who permanently uses LAs as speaker-specific trait, speaks about a sad story. On the other hand, at first sight, speakers who exhibit LAs as an idiosyncratic trait can make a sad impression without actually being sad.

The caveat has to be made that we are speaking of a sort of ‘cover phenomenon’ covering different sub-phenomena and different temporal traits: some are very short and might rather be perceived as segmental features, i.e. not as supra-segmental, prosodic features that are sort of modulated onto the speech wave. Of course, there are prototypical cases — no LA at all and laryngealized throughout — which easily can be told apart. But we simply do not know yet when people will produce which amount of LA and how an automatic classifier can model it.

It might be safer to find out non-existing/low correlations such as high pitch and fast speech with sadness. Further functions of LAs are reported in [21]. There are only a few studies dealing with the automatic detection of LAs, cf. [22,23]. We have manually corrected automatically extracted F0 values for one third of the database described in [4,5] (51 children giving commands to Sony’s pet robot Aibo). For some 6% of all voiced frames of all words, we found gross F0 errors denoting LAs; this amounts to some 14.7% words with laryngealized passages. The percentage of laryngealized words per speaker ranges from 0% to 35%; this illustrates a strong speaker dependency. At first sight, the distribution across emotional user states denotes more LAs in emotions with negative valence (*angry*, *touchy* (i.e. irritated), and *reprimanding*) than with neutral or positive valence. This could be a plausible result if we equate indicating negative valence with indicating some kind of superiority. This difference, however, disappears if we compute the distribution separately for words with the initial diphthongs [aU] and [aI] which are prone to be laryngealized more often than other vowels and diphthongs. The reason why is that in our database, some of these words – e.g. the vocative [’?aIbo] – are relatively frequent in the negative valence domain. Note that by that, we did of course not prove that LAs do not signal some emotional states, especially because in our data, emotions such as *sadness* (cf. the database processed in [24]) or *boredom* were not found. We can illustrate, however, the multi-functionality and speaker-dependency of LAs; thus it might be less likely that they are very useful as a generic feature within emotion classification. This might of course be different in a personalized setting.

4 Another Example: Pitch

Pitch is multi-functional, maybe up to the same extent as laryngealizations are. People can speak with flat F0 or with marked ups and downs — this is a personality trait. In the high days of intonation models, pitch was held responsible for the marking of word- and sentence accents, of salience, etc. During the last years, however, it has been shown that F0 is of minor importance, in relation to other parameters such as energy and duration, cf. [25,26,27] and [28]. The same might be true for emotion recognition; again, we do not know yet whether this might be due to pitch simply being less important, or to a combination of extraction errors, speaker specific traits and its bi-dimensionality which is difficult to model, esp. with sparse data.

The manual correction of the database mentioned at the end of section 3 resulted in some 6% gross F0 errors; first experiments on emotion classification with manual corrected F0 values yielded for a four-class problem some 3.5% better classification performance than with automatically extracted – i.e. sometimes erroneous – F0 values. Such a difference which is not very pronounced at first sight might, however, denote a difference between 'somehow relevant' and 'most relevant' feature types.

5 Implications from Applications

As nowadays the automatic recognition of emotion is getting more important for the voice business, several new questions are coming up. Since this single recognition process must become part of a business solution providing voice activated services to customers, we have to deal with integration and performance aspects, we have to figure out how the emotion recognition module can access data, and where the result is needed in which format; i.e. we have to care about interfaces. However, most important is that we have to know about the overall goal of the voice application in general, and we have to get an idea how this goal can be supported using emotion recognition. Last but not least do we have to find a compromise between technical and scientifical implementations on the one hand and budget and time restrictions on the other hand.

Here we will report on experiences we made in a project where emotion recognition was integrated as a component in a voice portal. A detailed description of the system, its capabilities and the results can be found in [29], [30] and [31]. Note that here we do not give details on recognition results, data sets, etc. since we want to concentrate on the fact that in integrated applications, a lot of constraints play a role and influence the emotion recognition, one would not think of in advance.

5.1 Voice Application Setup

Applying emotion recognition for business applications in the first step generates questions miles away from technological and functional aspects concerning the internals of the classifier. These questions concentrate on the business process the emotion recognition should be applied to, the other components working together, and performance and interface issues. For example, if you apply emotion recognition in a telephone based speech dialogue system, you have to care about the performance of the complete system since it has to be avoided that the caller has to wait too long for the system reaction. Thus all the processing has to be performed very fast, so that the system's response is generated within a period of at most 200 milliseconds after the caller stopped speaking.

But let us discuss these problems by means of the concrete project mentioned above. The voice application setup looks like the following: People ring up the voice application which is an information system. In addition to the usual technical components necessary for such an application – speech recognition, dialogue management and speech synthesis – new components for emotion recognition have to be deployed and integrated; constraints based on the specific architecture will be discussed below. In a first step we have to decide what we are really looking for: is it 'general anger' we want to classify in the speech signal or are we simply looking for a situation where the caller's emotional state changes for the worse? Actually, the second situation is the one we are interested in. Then, the speech dialogue system can react in a predefined manner, e.g. try to calm down the caller, transfer her to an agent or, if all agents are currently talking, give her a higher priority in the queue so that she will be transferred earlier than her position in the line would suggest. In this context the assumption usually is that the user behaves neutrally — at least in the first phase of the dialogue. Later on, either if the system makes recognition errors or gives displeasing information (e.g. a high telephone bill, a negative account balance), the caller might loose his good temper and thus change the communication style. Therefore, it is not highly important for the emotion recognition system to detect anger 'per se' in the speech signal but to be able to find those points in the spoken dialogue where the caller's emotional state changes to the worse. Thus the basic setup of the voice application has important implications for the classification task. If the task of the speech dialogue systems is different from the one presented above, it is perhaps very important to classify right from the beginning of the conversation whether the caller is angry or not.

In our example the task clearly is to detect changes in the emotional state of the user so that in case of a change for the worse, the system can apply different strategies to de-escalate. This already points towards a classification algorithm where features are used that characterize the changes in the acoustics between the current utterance and a reference utterance. This reference could either be the first user utterance in the dialogue, the preceeding utterance in the dialogue, or even perhaps an 'average' utterance computed from previous dialogues. The last procedure requires that the system is able to identify the caller, e.g. by means of analysing the calling telephone number, and to have access to a database where these references are stored. We applied such a kind of differential feature extraction algorithm for the project and compared these so called delta features with a classification algorithm using absolute features. The results reported in [29], [30] and [31] show that the delta features clearly outperform the absolute features on a data set from the described setup. The features used are prosodic features based on energy and F0 values, and duration features based on the segmentation of the speech signal into voiced and unvoiced regions. Actually, this should not be taken as proof that differential features based on prosody are generally more appropriate for emotion recognition; however, in the given case and under the constraints described above, they perform better and are the right choice for this task.

These results encourage to have a more detailed look on personalization and on those features dealt with in sections 2.4, 3 and 4; due to restrictions in time and budget, this has, however, not been possible for this specific project.

5.2 System Architecture

In this section we will have a detailed look at the constraints imposed by the chosen system architecture, the applied modules and the existing interfaces, and, especially, the features employed: as sketched above, we have a regular telephony based speech application using a speech recognizer, a dialogue management component, and a synthesis module; these three elements are standard modules also employed in other voice application environments. In our case we have to add the following components: two emotion recognition modules, one working exclusively on the recognized word chain (e.g. looking for swearwords) and the other one using only the speech signal as data source to compute its decision. Additionally, there is a decision module which takes the results of the two emotion recognizers and merges them into one classification result which is handed over to plan the necessary reaction. The speech recognition module is a standard product from the market with a predefined set of different interfaces and functions. Unfortunately, with these interfaces it is not possible to access all necessary (desired) information. It is for instance not possible to get the time alignment for the best word chain from the recognition engine, we do not have access to the features computed from the speech signal, and it is even not possible to get the incoming speech signal incrementally. Thus we have to wait until the end of the user's input before the waveform is accessible by other modules.

Looking at this system architecture, some interesting questions arise:

- Why do we use an ‘off-the-shelf’ recognizer engine with that many unwanted side effects?
- Why are there two separate emotion recognition modules, one using only acoustic, the other one only linguistic information?
- What would be a more appropriate system architecture and processing?

The application was planned to be installed and to go online either with or without emotion recognition. Basically, the operator of this information hotline wanted to have this specific automatic speech dialogue system. After in-depth discussion, they agreed with emotion recognition as additional component, because of the possible benefits. Nevertheless, they required that the resulting system should also work properly without emotion recognition and that it has to meet their internal administrative requirements. There were already other speech applications running based on the VoiceXML standard (cf. www.w3c.org/voice or www.voicexml.org for detailed information), using specific components and system architectures. Hence, the new system had to be based on this standard architecture, with the modules already in use in this company. The use of our own speech recognition module which would allow to have access to time alignment and spectral (MFCC) features was thus not possible.

If we look at the system architecture, the imposed restrictions, and the demand that the dialogue system has to react immediately after the user stopped speaking, it is obvious that the feature extraction for the emotion recognition does not have that much time. Additionally, at that time point when the emotion recognizer can start working, a recognized word chain is almost already available from the standard recognizer. Therefore it makes sense to separate the linguistic emotion classification from the acoustic processing; this is one of the reason why there are two emotion modules in the resulting system. Actually, the linguistic component is more or less 'integrated' in the recognition engine since the used grammars have to model also utterances containing swearwords and other phrases expressing emotion.

As for the acoustic emotion recognition, time constraints made it necessary to look for features and classification procedures which operate rather fast. Budget restriction made it impossible to spend additional time on the implementation of new types of features. Thus, we decided in favour of an already existing feature set based on energy, F0 values, and duration features based on the segmentation of the speech signal in voiced and unvoiced regions; as for details, cf. [31]. From the speech signal we computed one feature vector of defined length, usually a mixture of absolute and delta features. We decided to apply a rather simple Gaussian mixture model (GMM) approach for classification. For the training of the individual components of this GMM, five students manually annotated the utterances; for each utterance, a majority voting was applied. All this resulted not in an optimal but in a very good solution — given the constraints addressed above — for this specific application.

6 Concluding Remarks

In this paper, we dealt with those factors that are, in our opinion, most relevant for the — suboptimal — state of the art in emotion recognition: results obtained using acted speech and/or perception experiments with synthesized speech cannot be transferred onto real-life data; the sparse data problem prevents us from having enough training data both for speaker-independent and esp. for speaker-dependent modelling of spontaneous, real-life data; even if in theory, applications as described above could provide us with optimal classifiers and enough data for training, constraints imposed by time and budget prevent this. Of course, there are many other possible applications for emotion recognition [32,33], not only the call center scenario dealt with in this paper, which might impose other (types of) constraints on the implementation of an emotion recognition module.

A possible marking of one specific type of emotional state can be superimposed or hampered by at least these factors: several linguistic and paralinguistic functions such as given in Table 1, and by some extraction errors. Emotions are temporary phenomena and should be signalled not only locally at some specific (phonotactic) positions (cf. the linguistic functions in Table 1), and not globally as in the case of some paralinguistic functions. It might be possible to disentangle these functions on the time domain — but only with a personalized, speaker-dependent modelling. As is, the normal strategy in emotion recognition to classify speaker-independently short stretches of speech (at least syllables, sometimes words, most of the time phrases or turns/utterances) is possibly severely impaired because it is, at the time of the classification, not clear whether the marker is due to linguistic/paralinguistic factors, or to the signalling of emotions.

For many of the statements given above, there are no hard facts yet to prove or to invalidate them. Single studies will not do, converging results are the only way.

Acknowledgments

This work was partly funded by the EU in the framework of the NoE HUMAINE under grant IST-2002-507422, and by the German Federal Ministry of Education and Research (*BMBF*) in the framework of the SmartWeb project (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

References

1. Cowie, R., Cornelius, R.: Describing the emotional states that are expressed in speech. Speech Communication 40, 5–32 (2003)
2. Schuller, B., Müller, R., Lang, M., Rigoll, G.: Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. In: Proc. 9th Eurospeech - Interspeech 2005, Lisbon, PP. 805–808 (2005)
3. Labov, W.: The Study of Language in its Social Context. Studium Generale 3, 30–87 (1970)
4. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: Combining Efforts for Improving Automatic Classification of Emotional User States. In: Proceedings of IS-LTC 2006, Ljubliana, pp. 240–245 (2006)
5. Batliner, A., Steidl, S., Hacker, C., Nöth, E., Niemann, H.: Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. In: Proc. 9th Eurospeech - Interspeech 2005, Lisbon, pp. 489–492 (2005)
6. Schuller, B., Seppi, D., Batliner, A., Meier, A., Steidl, S.: Towards more Reality in the Recognition of Emotional Speech. In: Proc. of ICASSP 2007, Honolulu (to appear)
7. Scherer, K.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256 (2003)

8. Poggi, I., Pelachaud, C., de Carolis, B.: To Display or Not To Display? Towards the Architecture of a Reflexive Agent. In: Proceedings of the 2nd Workshop on Attitude, Personality and Emotions in User-adapted Interaction, User Modeling 2001, 7 pages (2001) (no pagination)
9. Batliner, A., Burger, S., Johne, B., Kießling, A.: MÜSLI: A Classification Scheme For Laryngealizations. In: House, D., Touati, P. (eds.) Proc. of an ESCA Workshop on Prosody. Lund University, Department of Linguistics, Lund, pp. 176–179 (1993)
10. Local, J., Kelly, J.: Projection and 'silences': notes on phonetic and conversational structure. Human Studies 9, 185–204 (1986)
11. Kushan, S., Slifka, J.: Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English? In: Proc. of Speech Prosody 2006, Dresden, pp. 795–798 (2006)
12. Ní Chasaide, A., Gobl, C.: Voice Quality and f_0 in Prosody: Towards a Holistic Account. In: Proc. of Speech Prosody 2004, Nara, Japan, 4 pages (2004) (no pagination)
13. Ladefoged, P., Maddieson, I.: The Sound of the World's Languages. Blackwell, Oxford (1996)
14. Gerfen, C., Baker, K.: The production and perception of laryngealized vowels in Coatzospan Mixtec. Journal of Phonetics 311–334 (2005)
15. Fischer-Jørgensen, E.: Phonetic analysis of the stød in standard Danish. Phonetica 46, 1–59 (1989)
16. Laver, J.: Principles of Phonetics. Cambridge University Press, Cambridge (1994)
17. Wilden, I., Herzel, H., Peters, G., Tembrock, G.: Subharmonics, biphonation, and deterministic chaos in mammal vocalization. Bioacoustics 9, 171–196 (1998)
18. Freese, J., Maynard, D.W.: Prosodic features of bad news and good news in conversation. Language in Society 27, 195–219 (1998)
19. Gobl, C., Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40(1-2), 189–212 (2003)
20. Drioli, C., Tisato, G., Cosi, P., Tesser, F.: Emotions and Voice Quality: Experiments with Sinusoidal Modeling. In: Proceedings of VOQUAL'03, Geneva, pp. 127–132 (2003)
21. Ishi, C., Ishiguro, H., Hagita, N.: Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction. In: Proc. of Speech Prosody 2006, Dresden, pp. 883–886 (2006)
22. Kießling, A., Kompe, R., Niemann, H., Nöth, E., Batliner, A.: Voice Source State as a Source of Information in Speech Recognition: Detection of Laryngealizations. In: Rubio Ayuso, A., López Soler, J. (eds.) Speech Recognition and Coding. New Advances and Trends. NATO ASI Series F, vol. 147, pp. 329–332. Springer, Heidelberg (1995)
23. Ishi, C., Ishiguro, H., Hagita, N.: Proposal of Acoustic Measures for Automatic Detection of Vocal Fry. In: Proc. 9th Eurospeech - Interspeech 2005, Lisbon, pp. 481–484 (2005)
24. Devillers, L., Vidrascu, L.: Real-life Emotion Recognition in Speech. In: Müller, C. (ed.) Speaker Classification II. LNCS(LNAI), vol. 4441, Springer, Heidelberg (2007)
25. Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H.: Prosodic Feature Evaluation: Brute Force or Well Designed? In: Proc. of the 14th Int. Congress of Phonetic Sciences. San Francisco, vol. 3, pp. 2315–2318 (1999)
26. Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H.: Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In: Proc. 7th Eurospeech, Aalborg, pp. 2781–2784 (2001)

27. Batliner, A., Möbius, B.: Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground? In: Barry, W., Dommelen, W. (eds.) The Integration of Phonetic Knowledge in Speech Technology, pp. 21–44. Springer, Heidelberg (2005)
28. Kochanski, G., Grabe, E., Coleman, J., Rosner, B.: Loudness predicts Prominence; Fundamental Frequency lends little. Journal of Acoustical Society of America 11, 1038–1054 (2005)
29. Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R.: An emotion-aware voice portal. In: Proc. Electronic Speech Signal Processing ESSP (2005)
30. Burkhardt, F., Stegmann, J., Ballegooy, M.V.: A voiceportal enhanced by semantic processing and affect awareness [34], pp. 582–586
31. Huber, R., Gallwitz, F., Warnke, V.: Verbesserung eines Voiceportals mit Hilfe akustischer Klassifikation von Emotion [34], pp. 577–581
32. Batliner, A., Burkhardt, F., van Ballegooy, M., Nöth, E.: A Taxonomy of Applications that Utilize Emotional Awareness. In: Proceedings of IS-LTC 2006, Ljubliana, pp. 246–250 (2006)
33. Burkhardt, F., Huber, R., Batliner, A.: Application of Speaker Classification in Human Machine Dialog Systems. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007) (this issue)
34. Cremers, A.B., Manthey, R., Martini, P., Steinhage, V. (eds.): INFORMATIK 2005 - Informatik LIVE! Band 2, Beiträge der 35. Jahrestagung der Gesellschaft für Informatik e.V (GI), Bonn, 19. bis 22 (September 2005). In: Cremers, A.B., Manthey, R., Martini, P., Steinhage, V. (eds.) GI Jahrestagung (2). LNI., vol. 68, GI (2005)

Emotions in Speech: Juristic Implications

Erik J. Eriksson[1,*], Robert D. Rodman[2,**], and Robert C. Hubal[3]

[1] Dept. Philosophy and Linguistics, Umeå University, Sweden
[2] Dept. Computer Science, North Carolina State University, USA
rodman@ncsu.edu
[3] Technology Assisted Learning Ctr., RTI International, USA

Abstract. This chapter focuses on the detection of emotion in speech and the impact that using technology to automate emotion detection would have within the legal system. The current states of the art for studies of perception and acoustics are described, and a number of implications for legal contexts are provided. We discuss, inter alia, assessment of emotion in others, witness credibility, forensic investigation, and training of law enforcement officers.

Keywords: acoustic parameters, affect, emotion, emotional categories, forensic, juristic, speech.

1 Introduction

Current natural language computational systems are able to infer much information from a person's spoken input based both on gross acoustic features and on lexical, syntactic, and semantic analyses. Information that can be inferred includes meaning, style, and certain speaker characteristics (Cole, et al., 1997 [1] Frank, et al., 2002 [2]). However, current systems are only able to draw inferences from a subset of features in the acoustic signal including intonational and stress patterns, overall loudness, peculiarities of phonation, and other distinctive properties of speech. Human listeners, on the other hand, have access to a full set of features and are able to integrate them in such a way as to acquire a wealth of information from them and apply this knowledge to identifying and classifying speakers, apprehending subtle shades of meaning, inferring implicatures and other pragmatic factors and, most relevant to the present work, perceiving affect and interpreting the speaker's emotional state.

Automated detection of emotion in speech holds considerable promise in many areas, including deception detection during interviews (Fuller, et al., 2006 [3]),

[*] Erik Eriksson's contribution to this paper was funded by a grant from the Bank of Swedish Tercentenary Foundation Dnr K2002-1121:1-4 to Umeå University for the project 'Imitated voices: A research project with applications for security and the law'.

[**] Corresponding author. Address all correspondence to: Robert Rodman Department of Computer Science, Box 8206 North Carolina State University Raleigh, NC 27695-8206, Fax: 919-515.7480, 919-515.7896.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 152–173, 2007.

separating cognitive from affective experiences in a clinical setting (Susca, 2006 [4]), and call center applications (Yacoub, et al., 2003 [5]). Here, we focus on legal settings. The entirety of legal actors (Maroney, 2006 [6]) – attorneys, defendants, executive officials, expert witnesses, judges, jurors (grand and otherwise), law enforcement officers, legislators, plaintiffs, prosecutors, regulators, suspects, victims, and witnesses - experience emotion and may display those emotions, the knowledge of which may have far-reaching ramifications for other actors. We first identify several areas where emotions in speech can affect legal judgment and decision making. Broadly, these areas involve assessment of emotion in others, emotions and memory (concerning witness credibility), emotions and culture (including effects on forensic investigation), and emotions in legal scholarship. Later, we describe some implications of emotion detection on the training and assessment of law enforcement officers, attorneys, and other actors in juristic processes. We also describe perceptual and acoustic studies supporting an optimistic outlook for the automated detection of emotion in speech.

2 Effects in Legal Contexts

2.1 Assessment of Emotion in Others

Detection of a particular emotion and the degree to which it is felt based on measurable acoustic parameters of speech could prove useful to many actors in the juristic system. Knowledge of emotional state would surely have powerful implications in the following situations and many more similar to them:

- Law enforcement officers would benefit by knowing what emotions a suspect is experiencing during interrogations, perhaps to gauge credibility or extent of knowledge.
- Attorneys could gauge credibility of a client or the client's level of knowledge or understanding by discerning the client's emotional state or state changes during discussions of the client's past behavior.
- Prosecution attorneys could profit by knowing what emotions potential jurors are experiencing during jury selection questioning so as to gauge their likely reactions to evidence.
- Defense attorneys could benefit by knowing what emotions the prosecution attorneys are experiencing as they present a case, perhaps to gauge the effects of alternate defense strategies.
- Jurors could benefit by knowing what emotions a witness is experiencing during testimony, to gauge credibility and remorse.
- Defendants could benefit by knowing what emotions a judge is experiencing when handing down a decision, perhaps to gauge the possibility of suspension of the sentence or parole.

In all these situations, the legal actors naturally exhibit emotion through their behaviors, including their gestures, facial expressions, body language, and, in particular, their speech. A device of some sort with the capability of detecting vocal patterns influenced by emotion could augment information gained through lexical, syntactic, and semantic analyses. The device need not be hidden or unobtrusive; the point would be to gather additional potentially useful information in a noninvasive manner.

To gather such information, automated emotion detection relies on exposure to a database of speech segments tagged with their emotional category (neutral, sad, angry, etc.). Each segment is also classified in accordance with its acoustic properties.

Cowie, et al. (2005) [7] present a number of such databases, generic in the sense that a general model, applicable to all speakers of the particular language, is to be constructed. Such databases may be thought to underlie *speaker independent* emotion detection. (Further detail is provided in the appendix.)

Speaker specificity of feature sets in emotion encoding was shown by Hozjan and Kačič (2006) [8]. *Speaker dependent* emotion detection may turn out to be far more effective, since one is modeling a single individual for the purpose of predicting that individual's emotional state, but it is formidable from the data collection point of view. This option is open only to experimenters who work with the same participants over a prolonged period and therefore have sufficient time to collect the amount of data needed for the modeling process. One such environment occurs in the context of vocal computer aided instruction (CAI). After some weeks of interacting with the same students, an automated CAI system, working in parallel with an intelligent tutoring system, would observe their frustrations, anxieties, achievements, and glowing successes, and from these observations build individualized databases (Burns & Capps, 1988 [9]). With such a database, a model of vocal affect/emotion individualized to the student could be produced. From that point forward, whenever the tutoring system would identify emotion in the student's voice, it could respond appropriately.

In most forensic and other legal settings, one would use a generic model. For instance, during jury selection, there are scarcely any data available regarding individual members in a jury pool. Similarly, most interrogations are relatively short, hardly enough to enable the speaker-specific learning that would need to occur for overcoming a generic model. However, in situations where witnesses testify at great length (e.g., Slobodan Miloševic at his trial in The Hague), there may be ample time to assemble and put to use a speaker dependent emotion detection system.

Finally, it remains to find conclusive results as to whether emotion detection is language dependent or independent. That is, if, say, the lowering of pitch is a mark of sadness in an English-speaking environment, would it be so in a Serbian-speaking environment, or a Zulu-speaking environment? Would the situation differ between tonal languages such as Mandarin Chinese, and the mostly non-tonal languages of Europe? These and other interesting questions are open to basic research, hence their effects on legal processes are yet unknown.

2.2 Emotions and Memory

Perception studies have shown that humans correctly recognize emotions only 60 to 70 percent of the time (Picard, 1997 [10]; Scherer, 2003 [11]). Courts often mistakenly put too much trust in eye and ear witnesses (Solan & Tiersma, 2003 [12]). Solan and Tiersma argue that courts might look further into mathematical and/or computer aided analysis of witnesses' testimonies to gauge their reliability, and this may include an assessment of the emotional state of the witness.

Witness credibility is so important that judgment of credibility of witnesses is included in the jury instructions, since eyewitness testimony is viewed as direct evidence and adds to the prosecution's circumstantial evidence. But aside from being able to accurately assess the emotions that a witness is experiencing during testimony, emotions play a crucial role in memory that needs to be better understood.

Cognitive psychologists commonly distinguish among memory formation or encoding, association, and reconstruction. All of these processes can be affected by emotion (Forgas, 2001 [13]). For instance, emotional events are thought to receive some preferential processing (Christianson, 1992 [14]; Taylor & Fragopanagos, 2005 [15]) and thus, like all stimuli that receive attentive processing, lead to more stable and perhaps more accurate memory traces. By the same token, surrounding stimuli associated with the event that are not attended to are not encoded, hence are not retrievable later. Similarly, when events are reconstructed during eyewitness testimony, salient stimuli are better recalled than less salient stimuli. Salience can be related to three factors: first, to the witness's prevailing emotion (the closer to the emotional stress of the experience, the more accurate the memory is considered to be) (Jackson, 1995 [16]; see also the encoding specificity principle in Tulving & Thomson, 1973 [17]); second, to the witness's confidence (which, however, is largely uncorrelated with the accuracy in memory of an event; Olsson, 2000 [18]); and third, to suggestions provided by others (Loftus & Ketcham, 1991 [19]; Loftus, 2003 [20]). Emotionally encoded stimuli can also alter attention; such stimuli can divert attention to themselves and away from lesser emotionally laden stimuli and thus render the emotionally encoded stimuli more salient in the context (Taylor & Fragopanagos, 2005 [15]).

There is some concern, though, that emotion or stress adversely affects eyewitness memory (Deffenbacher, et al., 2004 [21]). Understanding a witness's emotion may have interview and courtroom implications. For instance, during an interview, a defense attorney may note how an adversarial witness is becoming agitated or aroused and may justifiably claim bias in the witness's recall of events. Similarly, during questioning of a witness in the courtroom, a judge may notice the witness becoming stressed or emotionally involved by the questioning and may call for a different line of questioning or a recess to calm the witness. As stated, the credibility of a witness depends in large part on the witness's level of emotion.

2.3 Emotions and Culture

Cultural differences in emotions might impose serious problems in a forensic investigation. For instance, foreign language interpretations in police interviews have been shown to generate problems, especially if the interpreter is not properly trained, or if a police officer acts as the interpreter (Berk-Seligson, 2002 [22]). Russell (2002) [23] argued that even highly trained interpreters who are not serving dual positions such as interpreter and police officer during interviews (see Berk-Seligson, 2002 [22]), may affect interview outcomes or even verdicts in court proceedings. Russell argued that literal and correct translations of foreign languages should be emphasized. However, this might not be possible as there are numerous translation difficulties and ambiguities between languages (Wierzbicka, 1999 [24]) and between cultures (Semin, et al., 2002 [25]). Wierzbicka reported, as an example of problematic translations, the word marah, which is the closest translation of angry in Malaysian. However, the Malaysian word is incompatible with aggression and is closer in meaning to resentful than angry. So if a translator interprets the word marah as angry for a person under investigation for a crime involving aggressive behavior, the person could very well appear aggressive, even though the exact meaning of the word does not imply aggression.

Similarly, cultural differences can play a role in the expression of emotion during an investigation or during courtroom proceedings. Jackson's (1995) [16] descriptions of emotional expression, for instance, are based on the Anglo-Saxon culture which encourages hiding of emotions (Tsai & Chentsova-Dutton, 2003 [26]; Wierzbicka, 1999 [24]), a norm that is not as true in other cultures. A number of studies (e.g., Ekman, et al., 1987 [27]; Ekman & Keltner, 1997 [28]; Markus & Kitayama, 1991 [29]; Mesquita & Markus, 2004 [30]) have investigated cultural similarities and differences in facial expression and interpretation of facial expression, finding that a number of expressions represent universal emotional displays, but that cultural influences on the context of the expression or on self-concept can affect judgments of expression.

Further, during courtroom proceedings, as well as other negotiations, emotions provide means to conclude a positive result (Kopelman, et al., 2006 [31]). Kopelman, et al. showed that participants in negotiations of varying kinds (immediate outcomes, time limited ultimatum, and prolonged negotiations) displaying positive emotions throughout the discussion were more likely to arrive at a subjective interest-based agreement (e.g., a win). By making the legal actors aware of their own display of emotions, as well as letting a mediating judge be aware of these, a more neutral, and even perhaps less biased, ruling might be achieved.

2.4 Emotions in Legal Scholarship

The judicial system already acknowledges emotions as an integral component. The system itself is based on social morality norms which, in turn, are based on emotional values and views of the society (Karstedt, 2002 [32]). For instance, hate crimes are described by the culprit's attitude towards the victim and the punishment of such a crime is controlled, partly, by the culprit's emotions surrounding the event, the impact of the event on the victim(s), and the judge's perception of the social implications of the event (i.e., the need for statutory ruling) (Karstedt, 2002 [32]). Thus, emotions are inherently intertwined with the law (Vidmar, 2002 [33]).

Maroney (2006) [6], therefore, points out that emotions can and should be studied owing to their undisputed relevance to the law. A six-pronged approach is recommended for study within the law-and-emotion rubric:

1. Focus on a particular emotion such as disgust, fear, or shame, and pertinent legal considerations. For example, in the legal defense of battered women who strike back, legal recognition of the experience of the fear emotion, and the behaviors it may engender, could and should be taken into account in the course of legal proceedings.
2. Focus on causes of emotional states, or "affective forecasting". For example, a litigant who imagines winning a certain level of damages in a civil case may experience projected happiness, and that experience may induce the litigant to make important legal decisions such as rejecting a low, but perhaps appropriate, offer of settlement because it doesn't produce the projected level of happiness. The reverse, projecting sadness in the event of losing, may persuade a litigant to accept an inappropriately low offer of settlement.
3. Focus on theories of emotion - both methodological and within disciplinary categories - and how current law reflects any one particular theory, and whether current law favors one theory over another.
4. Focus on legal doctrine. Whereas the first three items focus initially on emotion, this item examines how a particular area of the law - most obviously *criminal law* - is subject to understanding emotion. Indeed, emotions such as passion are encoded in the legal system wherein a crime committed "in the heat of *passion*" may be regarded differently than a crime committed "in cold blood". But needless to say the entire panoply of legal taxonomy may be affected by the emotional state of the legal actors that are involved.
5. Focus on the theory of law, as contrasted with focusing on the theory of emotion. Here, the starting point is a particular theoretical approach to law followed by an analysis of theories of emotion from that particular point of view. For example, one might examine the emotional dimensions of "restorative justice".
6. Focus on how a particular legal actor's behavior is influenced by his or her emotions, and the emotional state of those with whom he or she interacts. This, in fact, is the primary focus of the present chapter and of most research in this area, the first five foci being as yet relatively unexplored.

Clearly, then, knowledge of emotions can be useful for the judicial system. However, means of collecting signals that carry emotional content are needed. One apparent source of such signals is the voice. It is readily available and can be collected non-invasively. Key here is to separate salient features for emotion recognition. The next section presents means of collecting, classifying, and analyzing emotion in speech.

3 Emotions in Speech

In interactions between individuals the voice is a major tool and often (e.g., during telephone conversations) the only tool of communication. Individuals not only convey the explicit meaning of their utterance via vocal quality when they are interacting with other humans, but they also present the receiver with information of a more complex nature through vocal affect, which is one of the surface manifestations of the emotions that the speaker is experiencing.

Studies of emotions in speech can be divided into two major fields, perceptual and acoustic. Perceptual studies use human listeners to assess the emotional content in a speech segment. These studies are often used for cross-culture comparisons, or to test the relation of specific acoustic cues to particular emotions (Yang & Campbell, 2001 [34]) (For example, when a change in pitch is heard, how frequently will a human listener perceive anger, or confusion, or finality?).

The other type, acoustic studies, uses speech data to extract salient features that are linked to specific emotions. Often the emotional content of speech is first perceptually evaluated and tagged before acoustic feature extraction is employed. (For example, when a human listener perceives anger, how frequently is, say, a rise in overall pitch detected acoustically?)

A more detailed description of how emotions and acoustic features are investigated may be found in the Appendix. The following sections present findings from studies in perceptual analysis and acoustic analysis. Legal implications are also presented.

3.1 Perceptual Studies

A number of perception studies have been undertaken (e.g., Breitenstein, van Lancker, & Daum, 2001 [35]; Cowie, et al., 2000 [36]; Douglas-Cowie, et al., 2000 [37]; Laukka, et al., 2005 [38]; Mullennix, et al., 2002 [39]; Scherer, et al., 2001 [40]; Thompson & Balkwill, 2006 [41]; Wurm, et al., 2001 [42]). Human classification rates have been found to lie near 70% for unknown voices (Picard, 1997 [10]; Polzin & Waibel, 2000 [43]). However, human listeners have been shown to use other effects, such as cultural or linguistic influences on the processing of emotions (Scherer, et al., 2001 [40]; Thompson & Balkwill,2006 [41]; Tickle, 2000 [45]), to assist in deriving additional information from speech.

At least four forms of speech have been used as emotional stimuli to present to human listeners. One form, speech collected from actors or amateurs, achieves its emotional content from emotions elicited by either instruction or by self-induced emoting. That is, the actors are asked to try to feel the emotion while recording speech. A second form, speech collected from actors or amateurs who are unaware of the purpose of the recording, comes from emotion elicited by a mood induction technique. That is, the emotion is induced by the nature or content of what is recorded. A third form is speech collected from real life television or radio shows. A fourth form is synthetic speech constructed by altering one or several acoustic features to (purportedly) reflect a particular emotion (Iida, et al., 2003 [46]).

Westermann, et al. (1996) [47] conducted a meta-analysis, investigating some 250 studies and comparing eleven methods of emotion induction. They found that pictorial and movie elicitation caused the strongest effect on listeners. They further found that the effect is shifted in favor of negative emotions, particularly in self-imposed methods. Finally, they found that the elicitation effect is raised if the listener is aware of the purpose of the experiment. One implication is that emotion-inducing images (e.g., a bloody stairwell) presented, say, as evidence in a courtroom may have strong effects on observers such as jurors, particularly for negative emotions (of possible benefit to the prosecution), while making jurors aware of the strong effects may help them understand their reactions (of possible benefit to the defense).

It has been argued that by recording actors eliciting emotions, full-blown and unambiguous emotions can be collected (Liscombe, et al., 2003 [48]; Scherer, et al., 2001 [40]). However, it has also been argued that acted speech is different from genuine emotions (Bachorowski & Owren, 2003 [49]; Batliner, et al., 2003 [50]) but may contain a core of truth as it often is reliably decoded by listeners (Scherer, 2003 [11]). Actors are thought to over-characterize emotions when producing them and tend to elicit emotions primarily via pitch and prosody (Batliner, et al., 2003 [50]). In fact, it has been argued that the material used in emotion research always should be collected in the real world (e.g., Cowie, et al., 2000 [36]; Cowie & Cornelius, 2003 [51]; Douglas-Cowie, et al., 2000 [37]; Picard, et al., 2001 [52]) and thus be authentic and genuine. However, this generally means that the emotions elicited will be less strong (Batliner, et al., 2003 [50]; Cowie & Cornelius, 2003 [51]; Douglas-Cowie, et al., 2000 [37]). Douglas-Cowie, et al. also point out that in using real-life material the emotions may be hidden by social or other factors or can be expressed in a degraded or mixed version. This implies that when the effects of emotional stimuli need to be made explicit to observers or recipients, particularly for real-world stimuli, the social or environmental context in which the stimuli occur should be recreated as best as possible. Witness memory and credibility may depend on these ideas.

During perceptual evaluations (also applicable to acoustic analysis), the emotional content of the utterance is related to a specific class of emotion. Cowie and Cornelius (2003) [51] argued that the size of the chosen set of emotions to identify would impact the level of categorization. That is, the more emotion classes to distinguish between, the more insecure the categorization will be. Hence, Cowie, et al. (2000) [36] and Douglas-Cowie, et al. (2000) [37] argued that the use of two (or three) continuous dimensions was a more succinct way of representing emotions. The dimensions here were valence and arousal (and power). Cowie, et al. (2000) [36] showed low inter-subject variation in mapping emotional content onto these dimensions. Laukka, et al. (2005) [38], however, argued that three dimensions are insufficient to represent differences between certain emotion classes, but increasing the number of dimensions to represent the emotional space not only makes data analysis difficult (hard to train listeners, etc.), it also limits the amount of explanation each dimension can give

(Cowie, et al., 2005 [7]; Scherer, 2003 [11]). Further, listeners often agree on the emotional content of stimuli in sets of limited number of categories (Scherer, 2003 [11]) promoting the more "classical" emotion categories (anger, happiness, etc.). The implication for legal actors, then, may be to limit their interpretations of vocal affect to these classical or universal (Ekman, et al., 1987 [27]) emotions, on the basis that other actors will tend to agree with those interpretations.

These perceptual evaluations can be done using either experts (e.g., Banse & Scherer, 1996 [53]) or laymen (e.g., Cowie, et al., 2000 [36]; Douglas-Cowie, et al., 2000 [37]; Tickle, 2000 [45]). However, a pre-processing by experts to filter out poor examples has been argued to be inappropriate as emotional elicitation may differ greatly between participants and any elicitation is an exhibition of the emotions, regardless of whether it is strong or not (Bachorowski & Owren, 2003 [49]). Hence, the expertise or experience of the observer or recipient may affect how emotional stimuli are interpreted. For example, judgments or responses to different cultural emotional stimuli may depend on knowledge of the other culture.

Indeed, perceptual studies have been undertaken to investigate effects such as cross-cultural similarities in emotion decoding (Scherer, et al., 2001 [40]; Thompson & Balkwill, 2006 [41]) or both emotion encoding and decoding (Tickle, 2000 [45]). The results of these studies suggest that there are similarities in both encoding and decoding, but differences also exist. Scherer, et al. found that individuals from nonwestern cultures, specifically Asian, were less successful in decoding the emotional content encoded by German actors, than were, for instance, Germans, French, and Americans. Note that Scherer, et al. used carefully constructed nonsense stimuli to lower the impact of language, though that may actually adversely affect individuals from cultures that demand larger contexts in which to assess emotional display (Mesquita & Markus, 2004 [30]). Tickle (2000) [45] found similar effects using English and Japanese encoders and decoders. Thompson and Balkwill suggested that there are both universal cues and culture-specific ones. They found in-group advantages to the extent of significance, suggesting that there are cultural-specific cues that other cultures overlook during decoding. The results were based on English listeners' judgments of English, German, Chinese, Tagalog, and Japanese utterances with prosodically encoded emotions.

Perceptual investigations are needed to test the salience of acoustic features to specific emotions, or to find unambiguous emotional content in speech samples. That is, either a hypothesized feature (e.g., pitch or fundamental frequency) is tested for emotional salience, in which case the feature is manipulated synthetically (Mozziconacci, 2002 [54]), or collected material needs to be emotionally labeled and confirmed (Batliner, et al., 2003 [50]; Douglas-Cowie, et al., 2000 [37]). Most of the studies presented in the next section used data perceptually evaluated before extraction of features.

Further details regarding emotion categories may be found in the Appendix.

3.2 Acoustic Studies

The main goal of acoustic studies has been to link any particular acoustic feature (or set of features) in a speech sample to the emotional state expressed (given by perceptual investigations) in that sample. Further, mathematical algorithms for classification have been used to correlate acoustic features with emotion exemplars, and therefore support a method of classification of emotions based on acoustic features.

There are two ways of investigating the emotional salience of acoustic features in speech. Mozziconacci (2002) [54] argued that the best way of finding acoustic correlates to specific emotions was to employ an analysis / re-synthesis method. A specific acoustic feature is determined a priori to be correlated with some emotion(s). The feature is then manipulated by voice synthesis while keeping other features constant. If listeners to the synthetic voice perceive the emotion in the presence of the feature, the feature can be said to correlate to some degree with the emotion(s).

The other way of investigating the emotional salience of acoustic features in speech is to use collected material and use data driven methods of extracting multitudes of features and measure the emotional salience for each of these features.

Numerous acoustic features have been investigated over the years. In one of the most inclusive studies (Batliner, et al., 2003 [50]), the task was to find acoustic correlates of user frustration. Features such as fundamental frequency (F0) and statistics for F0 (mean, standard deviation, overall range, minimum and maximum), temporal durations (length of pauses, etc.) with various reference points, speech rates, and spectral energy and tilt were all examined.

Oudeyer (2003) [55] included a plethora of features, but reduced the original number (exceeding 200) to a succinct few using a feature selection algorithm. Based on that algorithm, Oudeyer found the most salient content to be localized in the first part of the spectrum (0 - 250 Hz). However, only three of the commonly used features (mean, minimum, and maximum) for F0 were found to be among Oudeyer's top 20 features.

Features based on models of the spectrum have also been used, focusing on the first ten (Oudeyer, 2003 [55]) or sixteen (Polzin & Waibel, 2000 [43]) coefficients of a mel-frequency cepstrum and a twelve feature log frequency power coefficient vector (Nwe, et al., 2003 [56]). In sum, it appears that pitch (fundamental frequency) (e.g., Banse & Scherer, 1996 [53]; Bänziger & Scherer, 2005 [57]; Batliner, et al., 2003 [50]; Mozziconacci, 2002 [54]) or spectral information below 250 Hz (McGilloway, et al., 2000 [58]; Oudeyer, 2003 [55]) have high impact for emotion classification purposes. This is in accordance with findings by, inter alia, Williams and Stevens (1972) [59]. However, Toivanen, et al. (2004) [60] found, in conjunction with Oudeyer, that the commonly used measurements of pitch (fundamental frequency) such as mean and median did not show great significance for emotion classification. Toivanen, et al. used spoken Finnish as their language of choice and Oudeyer used French, whereas many of those that found fundamental frequency mean and median to have an impact used English as a language of choice. Therefore it can be argued that mean and median of fundamental frequency might be language specific cues.

Batliner, et al. (2003) [50] also followed McGilloway, et al. (2000) [58] in that they divided the speech into sections of interest. McGilloway, et al. called these sections "tunes" and defined them to be sections of arbitrary length between specified events (such as pauses of approximately 180 milliseconds). Hence, both Batliner, et al. and McGilloway, et al. could specify prosodic events based on the fundamental frequency curve during any particular tune and thus use these as features.

Prosodic events have also been used to separate emotional events into categories (e.g., Mozziconacci, 2002 [54]; Paeschke, et al., 1999 [61]; Polzin & Waibel, 2000 [43]; Schröder, et al., 2001 [44]). Batliner, et al. (2003) [50] found that prosodic cues alone were insufficient to achieve high classification rates, however Bänziger and Scherer (2005) [57] found successful discrimination between four emotions using "... simple F0 contours - such as F0 mean or F0 range ..." (p.265). Mozziconacci and Hermes (1999) [62] successfully correlated some intonational patterns to a subset of emotions, but found only partial correlation during a perceptual evaluation functioning as validation of the findings. Other studies have used fundamental frequency data to separate emotional content (Burkhardt & Sendlmeier, 2000 [63]; Dellaert, et al., 1996 [64]; Lee, et al., 2001 [65]; McGilloway, et al., 2000 [58]; Oudeyer, 2003 [55]; Paeschke, et al., 1999 [61]; Polzin & Waibel, 2000 [43]; Roy & Pentland, 1996 [66]). These studies are relevant because they imply that observers will rely on not just one acoustic feature of an emotion-inducing stimulus to categorize the effects. Legal actors ranging from law enforcement officers to interviewers need to learn to assess the breadth of behaviors exhibited by a subject, including the variety of vocal characteristics, before determining his or her emotional or psychological state (see Hubal, et al., 2004 [67]; Link, et al., 2006 [68]).

Voice quality (e.g., formant distributions of particular vowels, or phonation types such as creaky voice) has been studied by a few researchers (e.g., Burkhardt & Sendlmeier, 2000 [63]; Gobl & Chasaide, 2003 [69]). Burkhardt and Sendlmeier found some correlation to emotion involving voice qualities but this was not the case for Gobl and Chasaide (see also Janniro & Cestaro, 1996 [70]) who argued that the correlation between voice quality and expressed emotion is uncertain. Clearly, further research is needed in this area.

In order both to assess the multivariate description of emotions by a set of features and to use features to classify new utterances, multivariate tools and classification algorithms are needed. Oudeyer (2002, 2003) [71] [55] performed a comparison between several classification algorithms. These included several different types of neural networks, decision trees, *k*-star, kernel density, linear regression, several support vector machines, and AdaBoost. Oudeyer found that the most successful algorithm for his data was the AdaBoostM1/C4.5 method, which applies a machine learning technique to refine and stabilize the output of a decision tree method. This produced results as high as 96.1% classification rates with speaker dependent data and optimal feature sets. Other popular classification algorithms include a maximum likelihood Bayes classifier (Dellaert, et al., 1996 [64]; Polzin & Waibel, 2000 [43]), kernel regression (Dellaert, et al., 1996 [64]), k-nearest neighbor (Dellaert, et al., 1996 [64]; Toivanen, et al., 2004 [60]) and hidden Markov models (Nwe, et al., 2003 [56]).

Cues in which emotion is conveyed can also be found in higher linguistic levels. Batliner, et al. (2003) [50] used conversational artifacts, syntactic structure, and dialogue acts to find trouble in communication. They found that repetitions of statements, especially when an utterance is repeated word for word, are cues to severe problems and therefore annoyance in the observer or recipient.

Similarly, Hozjan & Kačič (2006) [8] used various durations (e.g., sentence, syllable, specific sounds) in conjunction with fundamental frequency and amplitude measurements. They found that the mean energy of segmented speech (i.e., energy means taken over segments of speech) had the highest significance for emotion classification. This measurement was closely followed by the durations of affricates, plosives, sonorants, and fricatives in that order. (See also Petrushin & Makarova, 2006 [72], for this effect in Russian.) They also found that any cue on its own was insufficient to separate any emotion pair, but combinations of different cues did. However, the combinations differed for each emotion pair and each speaker and they suggested that speakers might have personal preferences when selecting available cues to depict a specific emotion. It should be noted that their speakers had a mixed language background, which could indicate cultural or language dependent differences, although speakers of the same language showed no more similarities in cue setup than mixed-background speakers.

Schröder (2000) [73] investigated the impact of acted German affect bursts on perceived emotion. Affect bursts are short utterances produced appropriately for a specific emotion. For example, growling was used to convey threat. Schröder had actors choose whichever burst they saw fit for a specific emotion and found correlations between emotion and choice of sound. However, in a follow-up listening experiment the correlation was much weaker and confusion rates were greater.

4 Implications for Training and Assessment of Legal Actors

As was suggested, detection of emotion in others has implications for witness credibility and forensic investigation. Looking forward, automated detection of emotion using tools based on the research just described may have further implications for legal training and assessment.

4.1 Interaction Skills Training

Law enforcement officers and others in the legal system who regularly encounter suspects and witnesses need training on learning to assess those persons' emotions. As an example, there is a need for training law enforcement officers in managing encounters with the mentally ill (Engel & Silver, 2002 [74]). Frank, et al. (2002) [2] describe a system for that form of training, where law enforcement officers encounter a synthetic character (i.e., a computerized agent) and the officers must learn, using interaction skills alone, to de-escalate the situation. Along with gestural and facial expressions given by the character, emotion expressed in speech is critical and informative for these officers.

De-escalation is just one procedure that persons in the legal system perform. Other procedures include interrogation, negotiation, and crowd control, and emotion comes into play for all of these procedures. For instance, law enforcement interrogations done incorrectly can be suggestive and can lead witnesses to confident, emotionally laden, detailed mistaken memories (Loftus, 2003 [20]). All of these procedures also require, at some point, assessment of emotion as part of determination of intent. Training systems using technology similar to that of Frank, et al. that incorporate emotional characters offer great advantages in reliability, safety, and ultimately success in performance on the job.

4.2 Situated Assessment

Not only must the emotional state of individuals sometimes be assessed by an observer, but also the individual's responses to emotional stimuli must sometimes be assessed. This might be true, for instance, to gauge a defendant's behavior when emotional evidence is introduced. The closer the social and environmental context is to that which is on trial, the more realistic the response can be expected to be. That is, whereas practitioners of situated learning strive to have learners gain knowledge and acquire skills in the contexts that reflect how knowledge and skills are applied in everyday situations (e.g., Anderson, Reder, & Simon, 1996 [75]), a new line of research aims to place the individual within a simulated environment that closely mirrors the real environment, and measures the individual's assessment of the situation (e.g., Paschall, et al., 2005 [76]). The situation might measure physical behavior, but also verbal behavior (i.e., speech) exhibited by the individual. Paschall, et al. showed that a simulation is capable of differentiating between groups of participants, such as individuals diagnosed or not diagnosed with conduct disorder. The ability to detect a person's state through behavior exhibited in response to emotional stimuli holds promise for interrogation, for identifying remorse or feelings of guilt, for judging the effects of culture, and for judging credibility.

4.3 Admissibility of Machine-Detected Emotion as Evidence

Like all new technologies (e.g., fingerprints or DNA testing, at different times), admissibility as evidence may depend on the court's perception of the technology's reliability as well as its appropriateness in the particular kind of juristic process (e.g., criminal vs. civil) in question. As a precursor to the chain of judicial rulings that will undoubtedly come about in the future, a widely accepted principle of the admissibility of novel scientific evidence, called the "Frye Test" (from *Frye v. The United States in 1923*), is likely to be invoked. The criteria are that the technology would be first subjected to rigorous analysis by the scientific community during its experimental stage, and only after this community arrived at a consensus that the technique was valid would evidence of its use be admissible in court.

5 Summary

Automated detection of emotion in speech may improve legal decision making in areas that involve assessment of emotion in others, emotions and memory, emotions and culture, and training of participants in the legal process. Though current natural language systems are not yet fully able to interpret a person's emotion, ongoing perceptual and acoustic studies paint a promising picture for automated detection of the wealth of information available in the acoustic signal of speech. The advent of this technology will spur research into its effect on all aspects of the juristic system.

References

1. Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V., Varile, G., Zampolli, A. (eds.): Survey of the State of the Art in Human Language Technology. Cambridge Studies In Natural Language Processing Series. Cambridge University Press, Cambridge (1997)
2. Frank, G., Guinn, C., Hubal, R., Pope, P., Stanford, M., Lamm-Weisel, D.: JUST-TALK: An Application of Responsive Virtual Human Technology. In: Proceedings of the Interservice/Industry Training, Simulation and Education Conference, pp. 773–779. National Training Systems Association, Arlington (2002)
3. Fuller, C., Biros, D.P., Adkins, M., Burgoon, J.K., Nunamaker, J.F., Coulon, S.: Detecting Deception in Person-of-Interest Statements. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics, May 23-24, 2006, pp. 504–509, San Diego (2006)
4. Susca, M.: Connecting Stuttering Measurement and Management: II. Measures of Cognition and Affect. International Journal of Language & Communication Disorders 41(4), 365–377 (2006)
5. Yacoub, S., Simske, S., Lin, X., Burns, J.: Recognition of Emotions in Interactive Voice Response Systems. In: Proceedings of the European Conference on Speech Communication and Technology, September 1-4, 2003, pp. 729–732, Geneva, Switzerland (2003)
6. Maroney, T.A.: Law and Emotion: A Proposed Taxonomy of an Emerging Field. Law and Human Behavior 30, 119–142 (2006)
7. Cowie, R., Douglas-Cowie, E., Cox, C.: Beyond Emotion Archetypes: Databases for Emotion Modelling Using Neural Networks. Neural Networks 18, 371–388 (2005)
8. Hozjan, V., Kačič, Z.: A Rule-Based Emotion-Dependent Feature Extraction Method for Emotion Analysis from Speech. Journal of Acoustic Society of America 119(5), 3109–3120 (2006)
9. Burns, H.L., Capps, C.G.: Foundations of Intelligent Tutoring Systems: An Introduction. In: Poison, M.C., Richardson, J.J. (eds.) Foundations of Intelligent Tutoring Systems, pp. 1–19. Lawrence Erlbaum, London (1988)
10. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
11. Scherer, K.R.: Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication 40, 227–256 (2003)

12. Solan, L.M., Tiersma, P.M.: Falling on Deaf Ears. Legal Affairs (November/December 2003) Available from as of (August 30, 2004) http://www.legalaffairs.org/issues/November-December-2003/story_solan_novdec03.html
13. Forgas, J.: Handbook of Affect and Social Cognition. Lawrence Erlbaum Publishers, New York (2001)
14. Christianson, S.: Emotional Stress and Eyewitness Memory: A Critical Review. Psychological Bulletin 112, 284–309 (1992)
15. Taylor, J.G., Fragopanagos, N.: The Interaction of Attention and Emotion. Neural Networks 18(4), 353–369 (2005)
16. Jackson, B.S.: Making Sense in Law. Deborah Charles Publications, Liverpool (1995)
17. Tulving, E., Thomson, D.M.: Encoding Specificity and Retrieval Processes in Episodic Memory. Psychological Review 80, 352–373 (1973)
18. Olsson, N,: Realism of Confidence in Witness Identification of Faces and Voices. Unpublished doctoral dissertation, Uppsala University, Uppsala, Sweden (2000)
19. Loftus, E., Ketcham, K.: Witness for the Defense. St. Martin's Press, New York (1991)
20. Loftus, E.: Our Changeable Memories: Legal and Practical Implications. Nature Reviews: Neuroscience 4, 231–234 (2003)
21. Deffenbacher, K.A., Bornstein, B.H., Penrod, S.D., McGorty, K.: A Meta-Analytic Review of the Effects of High Stress on Eyewitness Memory. Law and Human Behavior 28(6), 687–706 (2004)
22. Berk-Seligson, S.: The Miranda Warnings and Linguistic Coercion: The Role of Footing in the Interrogation of a Limited-English-Speaking Murder Suspect. In: Cotterill, J. (ed.) Language in the Legal Process, pp. 127–143. Palgrave Macmillan Ltd, New York (2002)
23. Russell, S.: 'Three's a Crowd': Shifting Dynamics in the Interpreted Interview. In: Cotterill, J. (ed.) Language in the Legal Process, pp. 111–126. Palgrave Macmillan Ltd, New York (2002)
24. Wierzbicka, A.: Emotions across Languages and Cultures. Cambridge University Press, Cambridge (1999)
25. Semin, G.R., Görts, C.A., Nandram, S., Semin-Goossens, A.: Cultural Perspectives on the Linguistic Representation of Emotion and Emotion Events. Cognition & Emotion 16(1), 11–28 (2002)
26. Tsai, J.L., Chentsova-Dutton, Y.: Variation among European Americans in Emotional Facial Expression. Journal of Cross-Cultural Psychology 34(6), 650–657 (2003)
27. Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni- Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., Scherer, K.R., Tomita, M., Tzavaras, A.: Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. Journal of Personality and Social Psychology 53(4), 712–717 (1987)
28. Ekman, P., Keltner, D.: Universal Facial Expressions of Emotion: An Old Controversy and New Findings. In: Segerstrale, U., Molnár, P. (eds.) Nonverbal Communication: Where Nature Meets Culture, pp. 27–46. Lawrence Erlbaum Associates, Mahwah (1997)
29. Markus, H., Kitayama, S.: Culture and the Self: Implications for Cognition, Emotion, and Motivation. Psychological Review 98, 224–253 (1991)

30. Mesquita, B., Markus, H.R.: Culture and Emotion: Models of Agency as Sources of Cultural Variation in Emotion. In: Frijda, N.H., Manstead, A.S.R., Fisher, A. (eds.) Feelings and Emotions: The Amsterdam Symposium, pp. 341–358. Cambridge University Press, Cambridge (2004)
31. Kopelman, S., Rosette, A.S., Thompson, L.: The Three Faces of Eve: Strategic Displays of Positive, Negative, and Neutral Emotions in Negotiations. Organizational Behaviour and Human Decision Processes 99, 81–101 (2006)
32. Karstedt, S.N.E.: Emotions and Criminal Justice. Theoretical Criminology 6(3), 299–317 (2002)
33. Vidmar, N.: Case Studies of Pre- and Midtrial Prejudice in Criminal and Civil Litigation. Law and Human Behavior 26(1), 73–105 (2002)
34. Yang, L., Campbell, N.: Linking Form to Meaning: The Expression and Recognition of Emotions through Prosody. In: Proceedings of the 4th ISCA Workshop on Speech Synthesis. August 29 - September 1, 2001, Perthshire, Scotland (2001)
35. Breitenstein, C., Van Lancker, D., Daum, I.: The Contribution of Speech Rate and Pitch Variation to the Perception of Vocal Emotions in a German and an American Sample. Cognition and Emotion 15, 57–79 (2001)
36. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: FEELTRACE: An Instrument for Recording Perceived Emotions in Real Time. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 19–24. ISCA, Belfast, Ireland (2000)
37. Douglas-Cowie, E., Cowie, R., Schröder, M.: A New Emotion Database: Considerations, Sources and Scope. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 39–44. ISCA, Belfast, Ireland (2000)
38. Laukka, P., Juslin, P.N., Breslin, R.: A Dimensional Approach to Vocal Expression of Emotion. Cognition and Emotion 19(5), 633–653 (2005)
39. Mullennix, J.W., Bihon, T., Bricklemyer, J., Gaston, J., Keener, J.M.: Effects of variation in emotional tone of voice on speech perception. Language and Speech 45(3), 255–283 (2002)
40. Scherer, K.R., Banse, R., Wallbott, H.G.: Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. Journal of Cross-Cultural Psychology 32, 76–92 (2001)
41. Thompson, W.F., Balkwill, L.L.: Decoding Speech Prosody in Five Languages. Semiotica 158(1/4), 407–424 (2006)
42. Wurm, L.H., Vakoch, D.A., Strasser, M.R., Calin-Jageman, R., Ross, S.E.: Speech Perception and Vocal Expression of Emotion. Cognition and Emotion 15(6), 831–852 (2001)
43. Polzin, T., Waibel, A.: Emotion-Sensitive Human-Computer Interface. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 201–206. ISCA , Ireland, Belfast (2000)
44. Schröder, M., Cowie, R., Douglas-Cowie, M., Westerdijk, E., Gielen, S.: Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. In: Proceedings of Eurospeech, pp. 87–90. ISCA, Geneva, Switzerland (2001)
45. Tickle, A.: English and Japanese Speakers' Emotion Vocalisation and Recognition: A Comparison Highlighting Vowel Quality. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 104–109. ISCA, Belfast, Ireland (2000)
46. Iida, A., Campbell, N., Higuchi, F., Yasamura, M.: A Corpus-Based Speech Synthesis System with Emotion. Speech Communication 40, 161–187 (2003)
47. Westermann, R., Stahl, G., Hesse, F.W.: Relative Effectiveness and Validity of Mood Induction Procedures: A Meta-analysis. European Journal of Social Psychology 26, 557–580 (1996)

48. Liscombe, J., Venditti, J., Hirschberg, J.: Classifying Subject Ratings of Emotional Speech using Acoustic Features. In: Proceedings of Eurospeech, pp. 725–728. ISCA, Geneva, Switzerland (2003)
49. Bachorowski, J.A., Owren, M.J.: Production and Perception of Affect-Rated Vocal Acoustics. In: Ekman, P., Campos, J.J., Davidson, R.J., de Waal, F.B.M. (eds.) Emotions Inside Out, pp. 244–265. Annals of the New York Academy of Sciences, New York (2003)
50. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to find Trouble in Communication. Speech Communication 40, 117–143 (2003)
51. Cowie, R., Cornelius, R.R.: Describing the Emotional States that are Expressed in Speech. Speech Communication 40, 5–32 (2003)
52. Picard, R.W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1175–1191 (2001)
53. Banse, R., Scherer, K.: Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology 70(3), 614–636 (1996)
54. Mozziconacci, S.: Prosody and Emotions. In: Proceedings of Speech Prosody, pp. 1–9. ISCA, Aix-en-Provence (2002)
55. Oudeyer, P.Y.: The Production and Recognition of Emotions in Speech: Features and Algorithms. International Journal of Human-Computer Studies 59, 157–183 (2003)
56. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech Emotion Recognition using Hidden Markov Models. Speech Communication 41, 603–623 (2003)
57. Bänziger, T., Scherer, K.R.: The Role of Intonation in Emotional Expressions. Speech Communication 46(3-4), 252–267 (2005)
58. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S.: Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 207–212. ISCA Belfast, Ireland (2000)
59. Williams, C.E., Stevens, K.N.: Emotions and Speech: Some Acoustical Correlates. Journal of the Acoustical Society of America 52, 1238–1250 (1972)
60. Toivanen, J., Väyrynen, E., Seppänen, T.: Automatic Discrimination of Emotion from Spoken Finnish. Language and Speech 47(4), 383–412 (2004)
61. Paeschke, A., Kienast, M., Sendlmeier, W.F.: F0-Contours in Emotional Speech. In: Proceedings of ICPhS, pp. 929–931. Linguistics Department, San Francisco, USA, University of California, Berkeley (1999)
62. Mozziconacci, S., Hermes, D.J.: Role of Intonation Patterns in Conveying Emotion in Speech. In: Proceedings of ICPhS, pp. 2001–2004. Linguistics Department, San Francisco, USA, University of California, Berkeley (1999)
63. Burkhardt, F., Sendlmeier, W.F.: Verification of Acoustical Correlates of Emotional Speech Using Formant-Synthesis. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 151–156. ISCA, Belfast, Ireland (2000)
64. Dellaert, F., Polzin, T., Waibel, A.: Recognizing Emotion in Speech. In: Proceedings of the ICSLP, pp. 896–900. ICSA, Philadelphia (1996)
65. Lee, C.M., Narayanan, S.S., Pieraccini, R.: Recognition of Negative Emotions from the Speech Signal. In: Proceedings of Automatic Speech Recognition and Understanding, pp. 240–243 (2001)
66. Roy, D., Pentland, A.: Automatic Spoken Affect Analysis and Classification. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition, pp. 363–367. IEEE Computer Society Press, Los Alamitos (1996)

67. Hubal, R., Frank, G., Guinn, C., Dupont, R.: Integrating a Crisis Stages Model into a Simulation for Training Law Enforcement Officers to Manage Encounters with the Mentally Ill. In: Proceedings of the Workshop on Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence Spring Symposium Series, pp. 68–69. ACM Press, New York (2004)
68. Link, M.W., Armsby, P.P., Hubal, R.C., Guinn, C.I.: Accessibility and Acceptance of Responsive Virtual Human Technology as a Survey Interviewer Training Tool. Computers in Human Behavior 22(3), 412–426 (2006)
69. Gobl, C., Chasaide, A.N.: The Role of Voice Quality in Communicating Emotion, Mood and Attitude. Speech Communication 14, 189–212 (2003)
70. Janniro, M.J., Cestaro, V.L.: Effectiveness of Detection of Deception Examinations using the Computer Voice Stress Analyzer. Report No. DoDPI96-R-0005. Department of Defense Polygraph Institute, Fort McClellan (1996)
71. Oudeyer, P.Y.: Novel Useful Features and Algorithms for the Recognition of Emotions in Human Speech. Sony Computer Science Lab, Paris, France (2002) Available at Internet website: Downloaded on (2004)-03-17, `http://www3.isrl.uiuc.edu/ junwang4/langev/localcopy/pdf/oudeyerprosody2002a.pdf`
72. Petrushin, V.A., Makarova, V.: Parameters and Fricatives and Affricates in Russian Emotional Speech. In: Proceedings of Speech and Communciation (SPECOM), June 25-29, pp. 423–428, St. Petersburg (2006)
73. Schröder, M.: Experimental Study of Affect Bursts. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 132–137. ISCA, Belfast, Ireland (2000)
74. Engel, R.S., Silver, E.: Policing Mentally Disordered Suspects: A Re-examination of the Criminalization Hypothesis. Criminology 39, 225–232 (2002)
75. Anderson, J.R., Reder, L.M., Simon, H.A.: Situated Learning and Education. Educational Researcher 25(4), 5–11 (1996)
76. Paschall, M.J., Fishbein, D.H., Hubal, R.C., Eldreth, D.: Psychometric Properties of Virtual Reality Vignette Performance Measures: A Novel Approach for Assessing Adolescents' Social Competency Skills. Health Education Research: Theory and Practice 20(1), 61–70 (2005)
77. Lee, C.M., Narayanan, S.S., Pieraccini, R.: Classifying Emotions in Human-Machine Spoken Dialogs. In: Proceedings of IEEE ICME, pp. 737–740 (2002)
78. Lee, C.M., Narayanan, S.S.: Toward Detecting Emotions in Spoken Dialogs. IEEE Transactions on Speech and Audio Processing 13(2), 293–303 (2005)
79. Potapova, R., Potapova, V.: Temporal Correlates of Emotions as a Speaker-State Specific Parameters for Forensic Speaker Identification. In: Proceedings of SPECOM 2003, Moscow (2003)
80. Plutchik, R.: The Psychology and Biology of Emotion. Harper-Collins, New York (1994)
81. Jovičic, S.T., Rajkovic, M., Dordecic, M., Kašic, Z.: Perceptual and Statistical Analysis of Emotional Speech in Man-Computer Communication. In: Proceedings of Speech and Communciation (SPECOM), June 25-29, 2006, pp. 409–414, St. Petersburg (2006)
82. Rose, P.: Forensic Speaker Identification. Taylor & Francis, London (2002)
83. Kaiser, J.F.: On a Simple Algorithm to Calculate the 'Energy' of a Signal. In: Proceedings of ICASSP, pp. 381–384 (1990)
84. Slyh, R.E., Nelson, W.T., Hansen, E.G.: Analysis of Mrate, Shimmer, Jitter, and F0 Contour Features Across Stress and Speaking Style in the SUSAS Database. In: Proceedings of ICASSP, pp. 2091–2094 (1999)

A Appendix

Research in emotion detection from speech acoustics is a four-pronged investigation. The questions surround emotional categories, acoustic parameters, classifiers, and databases. In this Appendix we dig deeper into the emotion detection literature to illustrate these points.

A.1 Emotional Categories

Hundreds of emotion categories have been identified and discussed in the literature. From a practical stance, one needs to choose a small subset of these that suit a particular application. In an instructional context, for example, one might focus on the emotions of *confidence*, *confusion*, and *frustration.* In a judicial context, *anxiety*, *hostility*, or *uncertainty* may be the emotions of interest. In contexts where detecting and distinguishing as few as three emotions may be overly difficult, merely attempting to identify negative emotions from non-negative ones may have to suffice (Lee, et al., 2001, 2002, 2005 [65] [77] [78]).

To give an idea of how wide ranging researchers' views on categories of emotions are, we offer in Table 1 a non-comprehensive, alphabetized list of emotions that have appeared in the literature. One difficulty in compiling this list is that emotions are discussed in terms of both nouns ("happiness") and adjectives ("happy"). Adjectival descriptions have been translated into nominal ones for consistency. We believe the description should be consistent, but feel that the difference between whether one is experiencing the emotion of happiness versus experiencing a happy emotion, in describing a pervading feeling of joy, is of less concern as it relates to juristic implications.

The list in Table 1 is not only incomplete, but also the items are not mutually exclusive, in that the emotions overlap, several emotions may be experienced at the same time, and the emotions are "fuzzy" in the sense that one imagines them to be experienced to a greater or lesser degree, and not to be merely present or absent in all cases. The latter situation is addressed in Gobl and Chasaide (2003) [69] , who present eight sliding scales of emotions (shown in Table 2 in their original, adjectival format). (Potapova & Potapova, 2003 [79], present a similar scheme with their "scaleable subtypes" of emotions, e.g., fear is replaced by consternation - dread - terror.)

Emotions have also been represented on planes of varying dimensionality. For instance, Cowie, et al. (2000) [36] and Douglas-Cowie, et al. (2000) [37] presented their work with two or three dimensions (*activation*, *valence*, and *power*). Listeners were asked to rate stimuli along these scales. (A similar scheme is to plot Plutchik's (1994) [80] circle, as is discussed in Jovicic, et al., 2006 [81]). Laukka, et al. (2005) [38] extended Cowie's set of dimensions with a fourth, *intensity*, to accommodate further separation of emotions.

Correlating these dimensions with acoustic features, however, can be difficult. One approach is described by Jovičic, et al. (2006) [81], who suggest a three-level hierarchy of emotions within their multidimensional framework: primary, secondary, and tertiary. Primary emotions are fundamental and easiest to detect

Table 1. List of Emotions

anger, anxiety, bemusement, bliss, boredom, certainty, complacency, confidence, confusion, contempt, contentedness, delight, depression, despair, disgust, excitement, exhilaration, fear, friendship, frustration, fury, happiness, hostility, impatience, interest, neutrality, outrage, pleasure, politeness, relaxation, sadness, serenity, shame, stressfulness, surprise, terror, timidity, volatility.

Table 2. Sliding Scale of Emotions (from Gobl & Chasaide, 2003 [69])

relaxed-stressed, content-angry, friendly-hostile, sad-happy, bored-interested, intimate-formal, timid-confident, afraid-unafraid.

acoustically, for example *fear*. A secondary fear emotion would further subdivide into, say, *anxiety*, *terror*, *phobic*, distinctions that are more difficult to detect reliably from the speech signal. A tertiary fear emotion would presumably identify even finer distinctions, say *mildly anxious* to *severely anxious*, and these would be detected by “micro prosodic features” (p. 413).

A.2 Acoustic Parameters

There are many acoustic properties of the speech signal discussed in the literature, which reflect the panoply of vocal affects to be linked to the emotional states. We show a non-exhaustive list in Table 3, again noting that the properties are not always independent (orthogonal) amongst themselves, and that some properties may be manifested to a greater or lesser degree. Kaiser, 1990 [83] and Slyh, et al., 1999 [84] are two of a raft of papers on calculating certain acoustic properties.

A.3 Classifiers

The third consideration in determining emotion from voice has to do with the kinds of classifiers, or statistical tools, used to build models of speakers in which the acoustic parameters of vocal affect are statistically related to the perceived emotions. The underlying assumption is that there is a database of speech that is tagged with the names of the emotion or emotions purportedly evident from the various segments that comprise the speech. Associated with the tags are any acoustic parameters of interest. The classifier is used to build a speaker model from known speech samples, after which the model can be used to determine the emotions portrayed in future speech samples.

A variety of classifiers has been used by researchers, of which seven are: hidden Markov models, kernel regression, k-nearest neighbors, linear discrimination, maximum likelihood Bayes classifier, neural nets, and vector quantization. The list is hardly complete but it gives a sense of the eclectic tastes of the various

Table 3. List of Acoustic Properties

pitch mean, pitch median, pitch standard deviation, pitch extrema, median duration of falls or rises in pitch, speech rate, mean tune duration(segments separated by more than 180 milliseconds of silence),*long term average spectrum by frequency, spectral tilt* (a measure of the raising or lowering of the voice), *distribution of energy within various spectral ranges such as below 250Hz, jitter* (variation in pitch period), *shimmer* (variation in amplitude), *per-phoneme first formant mean, per-phoneme second formant mean.*

researchers. Pared down to essentials, given an emotion E and an acoustic parameter A, the model is intended to yield two probabilities: the probability of observing A when E is evident, and the probability of observing A when E is not evident. Mathematically, the former is written as P(A|E) and the latter as P(A|∼E). The ratio of these probabilities – P(A|E)/P(A|∼E) – gives the likelihood, or odds, that when A is detected in the speech, the emotion E is being experienced.

A simple arithmetic example makes this clear. Suppose that in 100 exemplars of speech wherein the speaker is said to experience sadness, the pitch falls 20% or more in 80 of the exemplars. Moreover, suppose that in 1000 exemplars of speech wherein the speaker is said not to experience sadness, the pitch falls 20% or more in 100 of those exemplars. Then the probability of the pitch falling when the speaker is sad is 80/100, i.e., P(A|E)=0.8, while the probability of the pitch falling when the speaker is not said is 100/1000, i.e., P(A|∼E)=0.1. The likelihood that the speaker is sad when the pitch falls 20% is therefore 0.8/0.1 = 8/1. Thus when the speaker's pitch drops 20% the odds are eight to one that the speaker is feeling – and expressing – sadness.

If other acoustic parameters are associated with the sadness emotion, their likelihoods may be computed similarly. If the parameters are known or assumed to be independent, then multiplying the likelihoods gives an overall likelihood ratio for the emotion given the acoustic parameters. (See Rose, 2002, [82] for an excellent, lucid explication of likelihood ratios.)

A.4 Databases

A fourth thrust of emotion detection research is the development of databases of speech tagged with emotions. The model building and likelihoods discussed above depend on having such a database.

Certainly one way to assemble such a database is to hire actors to exhibit the emotions of interest, and record their speech as they do so. This may be done directly by instructing the actor to speak a given sentence with a feeling of sadness, or it may be done by giving the actor a role to play and lines to read in which sadness is called for. This is a common source of data but one open to much question, perhaps best summed up in Douglas-Cowie, et al. (2000) [37] where the authors write, "At the very least, acted emotion cannot be a sufficient basis for conclusions about the expression of emotion".

Other recourses remain. One is to find "real" people and immerse them in emotion invoking situations while recording their speech. One would presumably not go so far as, say, to threaten to throw people from the top of the Roman Coliseum in order to collect fearful speech, though this suggests that Nero may have had the means to be an effective researcher. Rather, participants might be asked to recall and describe a particularly emotional event in their lives, or perhaps asked to read emotion-invoking passages aloud, either composed for the purpose, or drawn from the literature. Another approach, as taken by Douglas-Cowie, et al. (2000) [37], involves collecting data from media shows, either radio or television, featuring non-actors in verbal interactions that evoke emotions. (In general, negative emotions often come out in chat shows whereas positive emotions derive from religious programs.) For summaries of numerous human-based databases see Cowie, et al. (2005) [7].

A final possibility is to develop a database of emotionally charged synthetic speech (Iida, et al., 2003 [46]). Such a database would be useful for studies in human perception of emotions in the presence of particular acoustic features that could be tightly controlled. But a synthetic database should not be used to identify the acoustic correlates of a particular emotion as perceived by a human as that would clearly be circular, except in the framework of evaluating the efficacy of the database.

Application of Speaker Classification in Human Machine Dialog Systems

Felix Burkhardt[1], Richard Huber[2], and Anton Batliner[3]

[1] T-Systems International, Goslarer Ufer 35, 10589 Berlin, Germany
felix.burkhardt@t-systems.com
http://www.t-systems.com
[2] Sympalog Voice Solutions GmbH, Karl-Zucker-Str. 10, 91052 Erlangen, Germany
huber@sympalog.de
http://www.sympalog.de
[3] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Martensstr. 3,
91058 Erlangen, Germany
batliner@informatik.uni-erlangen.de
http://www.informatik.uni-erlangen.de

Abstract. This chapter deals with the application of automatic speaker classification in human machine dialog systems based on telephone operation. In a first step we introduce a taxonomy based on three features that such systems might have. We explain the features, namely *online*, *mirroring*, *critical* and their respective counterparts get explicated and are than used to characterize a part of the exemplary applications that illustrate the benefit of that approach. Furthermore prototypical application scenarios are described that shall illustrate the vast possibilities to utilize automated speaker classification in dialog applications.

Keywords: speaker classification, applications, voice portals, taxonomy.

1 Contents and Motivation

In human-human communication speaker classification takes place at all times and is very valuable for the communication process, as people constantly adapt their manner of speaking based on the assessment they have of their counterpart's age, gender, mood or mother tongue. Incorporating such strategies into (semi-)automated voice services can be very helpful to extend their benefit. On the other hand, the possibility of automated speaker classification based on telephone-calls makes new applications possible based on this feature in itself.

This chapter shall envisage some application that utilize speaker classification in a telecommunication scenario. We will propose a taxonomy of such applications based on features like online/offline or mirroring/non-mirroring. Further we will introduce a set of application scenarios and discuss in parts their place in the taxonomy. Further ideas for applications shall illustrate the many possibilities to utilize speaker classification with automated dialog systems.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 174–179, 2007.

2 A Taxonomy for Speaker Classification Applications

In [1] we introduced a taxonomy to distinguish between applications that utilize emotional awareness. Because emotion recognition can be seen simply as a subset of speaker classification, this approach is extensible with respect to other speaker classifications based on age, gender or accent.

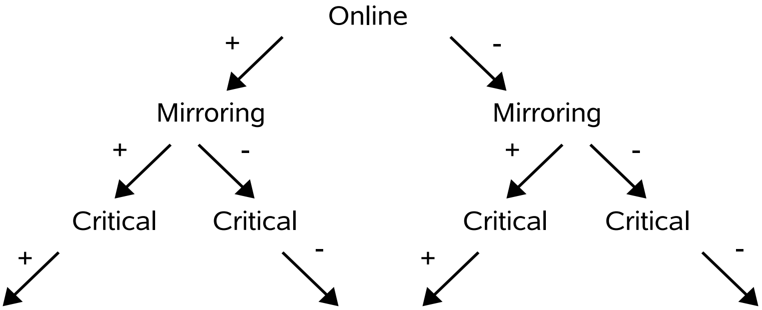


Fig. 1. A possible taxonomy of speaker classification applications

Such a taxonomy can be very useful to think up possible applications that utilize speaker classification by playing around with the inclusion or exclusion of certain features. In figure 1 a possible taxonomy is depicted as a tree. Now if we have an application that resides in one leaf of that tree, moving it to another place might give us ideas to think up a different application. Examples are given below. In the following paragraphs we introduce three binary features that could be used for such a taxonomy.

Online vs. Offline: With an online system the classification is made immediately while interacting with user, while with an offline system the classification takes place delayed, after the actual interaction, based on logged audio data. Take as an example an ad sponsored telephone service that presents target group optimized advertising based on user classification. In an online version the system would decide with each call which target specific advertisement to display whereas an offline version of the same application would simply analyze the main user groups, perhaps analyzed for certain time-intervals or branches of the dialog, and present user-targeted advertising based on that decision.

Mirroring vs. Not Mirroring: This distinction is based on whether the user gets a direct feedback based on the classification or whether he/she is not directly aware of it. Take as an example a system that analyzes the perception of accent. In a mirroring version such a system could help a language student to improve the pronunciation[1] while a non-mirroring variant might be used to offer the caller a dialog in his/her mother tongue.

[1] Although we admit that it is not straightforward to imagine this as a service offered via telephone.

Critical / Not Critical: If we talk about a *critical* application we mean that the feature "speaker classification" is indispensable for the success of the application's purpose. Of course all applications whose main feature consists of the speaker classification, like the example of the language-student that gets supported by accent detection, fall into that category.

A further differentiation of applications can be based on the speaker features that are classified. An application that classifies callers with respect to their age might make sense also with other features like accent, gender or emotion. E.g. the obvious idea to offer callers an optimized dialog design according to their age group can be mapped to emotion recognition in the anger-detecting voice portal scenario.

3 Application Scenarios

In the following we will introduce a set of applications that utilize speaker classification. Some of them were repeatedly envisaged in scientific literature, others have been mentioned already in the public media and few are even deployed as real-world applications. Many applications based on emotion recognition are further discussed in [2] or [1]. As such they can be taken as prototypical applications that stand for a family of related ideas. By applying the taxonomy mentioned in Section 2 the set can be enlarged by thinking up applications that differ with respect to a certain feature or utilize the classification based on a different speaker feature.

Dispatching callers to trained agents. This describes in a generic way the idea to classify the customers of a call center and forward them to an agent whose profile matches the caller's class. One obvious example would be the language of the caller in a multilingual call-center, e.g. a credit-card hotline. Another famous application foresees the measurement of anger in the caller's voice (e.g. for a complaint hotline), so that very angry customers can get handled by specially trained agents [3]. But generally spoken, the idea to let a specially talented agent take care of certain speaker classes can be generalized to classes like men, women, elderly, kids, strangers, or people coming from a specific part of country.

Adapted Dialog Design. If we take automated voice portals into account, dialog design becomes an important issue. State of the art technology in language understanding and artificial intelligence does not yet allow for totally free and open dialogs in human computer interaction (HCI). Dialog design comprises the way how the dialog flow is designed, i.e. which grammars are activated, which prompts will be played, which choices can be made by the user and which feedback strategies are implemented. In the case of static speaker classification like age or gender one possible application in this context would be to implement several designs and activate the one that fits best to the current user profile. This might consist of very subtle changes, for example elderly customers might prefer a slower speech rate in the system's prompts. A misclassification would then not lead to a perceptible difficulty for callers, resulting in a non-critical application with respect to the above mentioned taxonomy. On the other hand a dynamic speaker classification like e.g. emotion or language could be used to adapt the dialog dynamically to a change in the user's state. One of the most famous examples for such an application consists of the emotion-aware voice portal that detects user's anger and tries to soothe him/her by comforting dialog strategies as described in [3]. With this example a misclassification might lead to serious problems as callers that were not angry will probably get angry if accused unjustly. In that sense the application can be seen as critical.

Target-group specific advertising. Analogous to the development of Internet services it is foreseeable that a rising number of telephone services will by financed by advertising. Knowledge of the user group can help a great deal with respect to product choice and way of marketing. According to the taxonomy this could be used as an online application if the users get classified in the beginning of their conversation and tailored advertisement is presented to them in the course of further interaction, e.g. while waiting for a connection. As an offline application on the other hand the main user groups of a specific voice-portal or branch of a voice portal could be identified during a data collecting phase and then later advertisement targeted for the major user group be chosen. This is not a critical application as the main target of the interaction (a user gaining information) would not be endangered by the choice of an inappropriate advertisement. It is also not mirroring as the user is not directly aware of being classified.

Adapted Persona Design. With "persona design" we describe an automated voice-portal's character as a virtual persona. The character is represented by the wording of the prompts, the sound of the systems voice[2] and the expressions the grammar consist of. The design of such a "persona" can be enhanced to a great deal by background music or other sound-effects. Encountering an inconsistent persona design can be very confusing in an interaction, just like speaking to a person with multiple personalities. Usually the persona design is influenced by the topic the voice-portal application is all about, e.g. a banking application will prefer a serious persona, perhaps an elderly gentleman whereas a ring tone download application might use a trendy younger girl character. But speaker classification makes it possible to design one and the same application with different personas and activate them based on a prior application. Analogous to the target-group specific advertising application this would be conceivable in an online version and an offline version, where the major user groups are identified in a preliminary data collection phase.

Market analysis of target groups. Even if the system's reactions are not influenced by the speaker classification, a knowledge of which group called when or was interested in which product can be invaluable for marketing strategists. Because it's not important to gain such a knowledge during the course of interaction, this could be realized as an offline application, thereby allowing the

[2] Irrespective of the fact whether it's playback of prerecorded prompts or speech synthesis.

use of more sophisticated classification algorithms that don't need to follow real time requirements.

Call Center Quality Management. Call center managers have a strong interest in monitoring and optimizing the quality of the services provided by their agents. Speaker classification can be employed for this purpose in a variety of ways. For example, a classifier for the emotional state of the agents and/or the caller can be used to calculate numerical values that act as an indicator for the average quality of service provided over a certain period of time. Based on a large number of calls, even a classifier with a rather poor recognition rate on the utterance level is suitable to reliably detect relevant changes, such as an increasing number of angry callers or an increasing average stress level of the call center agents. Speaker classification can also be employed to identify individual calls in a corpus of recorded call center conversations. For example, calls with angry users can be selected for the purpose of training agents to cope with this situation.

Telephone surveillance. Clearly, speaker classification methods offer the potential for increasing the effectiveness of telephone surveillance measures by government authorities, e.g. by preselecting calls conducted by certain subgroups of the population.

Influence on staff planning. There are studies showing that the success of a call center is higher (this could be measured in customer satisfaction or sometimes even in revenue) when the characteristics of the caller and the agent matches, i.e. they are in the same age range, equal social level, etc. So if it is possible to create a profile of the caller groups over time (in the morning young male high professionals call and in the afternoon elderly housewives from middle class families) it makes sense to try to match the structure of the call center agent groups in that time to the caller structure.

Cross- and Up selling. Imagine an automated ordering hotline where people call in and place their orders. What usually should be done in such an application is that the system gives the callers some proposals what should be ordered in addition (this sometimes is also done by human agents). If the system in that case could not only use the already ordered product as the base of the decision what to propose as up selling item but also the gender of the caller and the age and other speaker characteristics, the chance that the proposed additional item is ordered can be maximized. Just as an example: if someone orders a digital camera in an voice application, the up selling item for younger callers perhaps should be some software product like Photoshop for manipulating taken photographs whereas for elderly people an additional suitcase is proposed.

Quiz Games/Prize Competitions. In automatic prize competitions over the telephone the set of questions can be matched against the callers characteristic. Usually, the idea behind those quiz games is that callers have a quite good chance to get the right answers so that there is the chance for the company to establish a relation to the caller. If the systems is aware of teens calling in perhaps questions on currently successful pop bands is the right choice so that the callers have a quite good chance to know the right answers whereas for the older generation questions on classic music are more preferable.

Gaming, Fun. A related field of applications is given by applications in the gaming or entertainment sector. For example the love detector by Nemesysco Ltd. attempts to classify speech samples based on how much “love” they convey. Other entertainment application ideas comprise online-telephone role games or horoscope applications that variegate their prognoses based on some automatically detected speaker characteristic.

4 Conclusion

Generally spoken, the use of automated speaker classification is of high potential when it comes to the design of more natural interfaces in human machine communication. We introduced a taxonomy that might be helpful to think up possibilities for the integration of speaker classification in various application fields. Furthermore we see from the large list of prototypical scenarios that can benefit from speaker classification strategies that this technology is indeed an important basis for enhancement in all kinds of application fields.

Acknowledgments

This work was partly funded by the EU in the framework of the NoE HUMAINE under grant IST-2002-507422, and by the German Federal Ministry of Education and Research (*BMBF*) in the framework of the SmartWeb project (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

References

1. Batliner, A., Burkhardt, F., van Ballegooy, M., Nöth, E.: A Taxonomy of Applications that Utilize Emotional Awareness. In: Erjavec, T., Gros, J. (eds.) Language Technologies, IS-LTC 2006, Ljubljana, Slovenia, Infornacijska Druzba (Information Society), pp. 246–250 (2006)
2. Picard, R.: Affective Computing: Challenges. Journal of Human-Computer Studies 59, 55–64 (2003)
3. Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R.: An emotion-aware voice portal. In: Proc. Electronic Speech Signal Processing ESSP, pp. 123–131 (2005)

Speaker Classification in Forensic Phonetics and Acoustics

Michael Jessen

Department of Speaker Identification and Audio Analysis (KT54),
Bundeskriminalamt, 65173 Wiesbaden, Germany
Michael.Jessen@bka.bund.de

Abstract. Speaker classification in forensic phonetics and acoustics is relevant for several practical tasks within this discipline, including voice analysis, voice comparison, and voice lineup. Six domains of speaker characteristics commonly used in forensic speech analysis are addressed: dialect, foreign accent, sociolect, age, gender, and medical conditions. Focussing on gender plus the less-commonly used characteristic of body size, it is argued that while auditory analysis is indispensable in forensic speaker classification, acoustic analysis can provide important additional information.

Keywords: dialect, foreign accent, sociolect, age, gender/sex, speech pathology, speaker size/height, formants, fundamental frequency.

1 Introduction

Within the field that is commonly referred to as “Forensic Speech and Audio Analysis” (e.g. by the European Network of Forensic Science Institutes, ENFSI) or “Forensic Phonetics and Acoustics” (from the name of the International Association for Forensic Phonetics and Acoustics) speaker classification is a regular and important task. The classification characteristics that are most commonly used in forensic phonetics and acoustics are gender, age, dialect, foreign accent, sociolect, and speech pathology. Speaker classification in forensic phonetics is primarily performed within two different practical tasks. At the German *Bundeskriminalamt* (German Federal Police Office, BKA) these two tasks are named *Stimmenanalyse* and *Stimmenvergleich*. The most direct translations of these terms are “voice analysis” and “voice comparison”, respectively, and these are the terms that will be used during this paper.

The difference between voice analysis, voice comparison, as well as further tasks under the more general heading of forensic speaker identification, is defined in practical terms and is based on the availability or non-availability of recorded speech material and the availability or non-availability of a suspect in a criminal case. From these two factors four possible combinations arise.

In the first combination of situations there is recorded material from an unknown speaker who is connected to a crime (e.g. the voice of a kidnapper making ransom demands over the telephone), but no suspect is present and hence no recording of a

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 180–204, 2007.

suspect can be made available which could be compared with the recording of the unknown speaker. Such a situation often arises in the early stages of a police investigation. In such a case the police asks an expert in forensic speech analysis for any characteristics of the unknown speaker that can be inferred from his recorded voice and that can be of help in finding a suspect. The task carried out by the forensic expert in response to this request is called *voice analysis*. Sometimes also the term "voice profiling" is used for the same activity. Voice analysis is almost exclusively a speaker classification task.[1] Depending on the circumstances of the case voice analysis often has to be carried out very quickly and sometimes with limited resources (when it has to be performed outside the speech laboratory), but voice analyses can be – and often have been in the past – of crucial help for the police in finding a suspect. Alternatively or in addition to carrying out a voice analysis it is possible to present the recorded speech sample to a wide public, using radio, television or related means.

In the second type of situation a suspect has been found and there is recorded material available not only from the unknown speaker but also from the suspect. (Presence of a suspect does not necessarily imply availability of recorded material from that suspect; sometimes the suspect refuses to provide a speech sample and there might be no other means of obtaining speech material from that person.) In such a case the two speech samples can be compared with respect to a wide variety of speech properties and a statement will be made as to whether or not the two speech samples were produced by the same person (depending on the expert or expert group, such a statement is usually expressed in probabilistic terms). This activity is called *voice comparison*. Speaker classification is part of the work involved in voice comparisons. But unlike voice analyses, which are almost entirely a speaker classification task, voice comparisons also involve (other) aspects of speaker identification such as the auditory assessment of voice quality or the measurement of average fundamental frequency, with no inference on speaker class.

In the combinations of situations addressed so far, recorded material was available. If no such material is available, an alternative is the presence of at least one witness who has heard the voice of the unknown speaker in association with the crime (often the witness is also the victim of the crime). Again one has to distinguish between the presence and the absence of a suspect. And if a suspect is present one has to distinguish between the case where the witness knew the speaker before and recognized him during the crime (which might be the very reason a suspect exists) and the situation where the witness did not know the speaker and hence could not recognize him during the crime. In the former situation the forensic specialist has to assess the reliability of the witness' recognition of the speaker during the crime. In the latter case a so called *voice lineup* (also called "voice parade") is designed and carried out. This is a complicated process involving many steps and precautions ([1] for guidelines that are agreed upon by most practitioners in forensic phonetics and acoustics). Speaker classification is one aspect in the preparation of voice lineups. The witness is asked by the expert which gender, age, dialect etc. the unknown

[1] "Almost", because some speaker characteristics potentially reported in voice analyses – such as an unusually high pitch level – are not speaker class characteristics in a strict sense but can still be useful in an investigation if police and laypersons understand what is meant by the particular speaker characteristic and can, for example, compare it with their memory of person they know or have encountered.

speaker had. Subsequently, the expert has to select foil speakers in the lineup that match the classification characteristics of the offender described by the witness and/or the classification characteristics that the suspect has according to an analysis by the expert ([2] for more in-depth discussion of this point). Finally, all the speakers in the lineup have essentially the same speaker classification and the witness has to concentrate on differences between speakers that go beyond speaker classification. For each speaker in the lineup the witness is asked whether the voice of that speaker corresponds to her/his memory of the speaker at the time of the crime. In this manner a pure speaker identification task is performed by the witness, because speaker classification had already been performed with the foil selection process. Speaker classification performed by witnesses (in collaboration with experts) is a different matter than speaker classification performed by experts that are trained in phonetics or linguistics. It is an interesting topic in its own right, though, and of interest beyond forensics: how do phonetically untrained laypersons classify speakers? But in the interest of space the paper will focus on speaker classification by experts and on situations where recorded material from at least the unknown speaker is available.

The fourth possibility in the taxonomy of situations shall not be left unmentioned although it is only rarely relevant in current practice. This would be the case where a witness has heard the speaker in association with the crime and does not know him, where no recordings exist, and where no suspects exist that would make a voice lineup possible. In such a case the expert can try to make a voice analysis in collaboration with the witness, arriving at a speaker classification plus some very salient speaker characteristics such as increased pitch level, if these occurred. This task could be aided by the use of speech databases containing multiple speaker classification combinations, so that the speaker classification task becomes a more illustrative and less abstract matter for the witness and so that terminological misunderstandings between witness and expert can be clarified. A more challenging way of responding to this situation would be to use a speech synthesizer in order to create some form of acoustic phantom picture. Although sketched as early as in [3], it is not apparent from published sources since then that anybody has approached such a forensic task yet. Given the good quality of current speech synthesis systems – including ways of changing speakers and voice qualities either with formant synthesis or concatenative synthesis plus signal manipulation – such a method of creating acoustic phantom pictures is not completely unrealistic any more.

This paper proceeds as follows. In the next section the classification characteristics most commonly used in forensic phonetics and acoustics are presented one at a time and elaborated upon as far as some of the most important aspects of forensic analysis are concerned. In the subsequent section one current issue in forensic speaker classification – the use of acoustic methods – is addressed, and finally a conclusion is presented.

In several locations during the paper it is mentioned how speaker classification is performed at the BKA. When in these cases the term “we” is used it refers to the members of the forensic phonetics and acoustics group of the BKA (the official name is *Fachbereich für Sprechererkennung und Tonträgeranalyse*, best translated as Department of Speaker Identification and Audio Analysis). Despite this usage of the first person plural the responsibility for the statements and suggestions made here rests with the author of this paper alone. Laying particular emphasis on the speaker classification procedures as used at the BKA is justified only insofar as we know from regular participation in the Annual Conference of the International Association for Forensic Phonetics and Acoustics and related national and international events that similar procedures are used by other institutes and individuals involved in forensic phonetics and acoustics.

For other overviews of speaker classification along with many other aspects of forensic phonetics and acoustics which cannot be addressed in this paper the reader is referred to textbooks and tutorials such as [3–13]. The reader is also recommended to consult the *International Journal of Speech, Language and the Law* (formerly: *Forensic Linguistics*) for many articles on forensic phonetics and acoustics.

2 Speaker Classification Characteristics in Current Forensic Practice[2]

2.1 Gender

Gender determination is usually a simple task. By far the most prominent phonetic correlate of gender is the average pitch level of the speaker, which among adult speakers (especially before the onset of old age) is on average much higher for female than male speakers ([5, 14] for German speaker data based on 100 men and 50 women; cf. also [15] for multi-stylistic data on 100 German-speaking men). This pitch difference can be explained by male-female differences in the length of the vocal folds [16]. Pitch level could be measured in terms of average fundamental frequency (f0), but usually the male-female pitch level differences are salient enough that measurement is not necessary. In practice, however, we encounter cases where pitch level cannot be determined or where pitch level is not informative for the gender determination task. Situations of this kind occur when whisper or unusual phonation types such as falsetto or creak are used as a voice disguise strategy or for other reasons.[3] Another difficulty of this sort arises when the pitch level of the speaker is

[2] The focus of this review is on speaker classification features with complete or strong stability over time. For example, acquired dialectal patterns remain permanent in adulthood or change only very slowly. What is not addressed here are classifications that can change rapidly. Not discussed, for example, is the detection of alcohol consumption or the identification and classification of stress or emotion. One way these transitional classifications are relevant in forensic phonetics and acoustic is when in voice comparisons there are indications of a mismatch in the occurrence or intensity of transitional classifications between the questioned and the reference material. In such a case it is necessary to know about the impact that these transitional classifications can have on the speaker identification parameters that are investigated in a voice comparison.

[3] Some disguise methods lead only to partial neutralisation of gender differences in pitch level (thanks to Stefan Gfroerer for emphasizing this point). According to [17], when women produce falsetto voice they tend to have higher average f0 than when men produce falsetto voice. However, women more rarely used falsetto voice when they were asked to increase pitch as a voice disguise than men, so that if male falsetto voice is compared to female high-pitched (by disguise) but still modal voice the pitch difference between men and women is usually very small. If male falsetto voice is compared to normal female voice, the males tend to have higher f0 (see Figs. 5, 6 in conjunction with Table 2 in [17]).

unusually high for a man and unusually low for a women, i.e. is located the ambiguous area between the Gaussian distributions of male and female f0 levels in the population. Furthermore, all different kinds of "gender bender" scenarios are possible (e.g. influence of added hormones or pathological hormone levels, some cases of homosexuality and transsexualism) that can make the task difficult (which are situations that might not only affect pitch).

In cases where pitch level fails, other indices of gender have to be relied upon and the conclusion on gender classification has to be stated in more probabilistic terms. One correlate of gender that will be addressed separately in section 3.2 are the formant frequencies. These are on average higher in female than male speech due to differences in vocal tract length. Although differences in pitch/fundamental frequency and formant frequencies between men and women have a strong foundation in anatomical differences, the anatomical effects can be enhanced or reduced by behavioural factors. For example, it has been found that male-female formant differences are not the same in every language; e.g. they are much larger among speakers of Russian than among speakers of Danish. Cultural factors could be responsible for these differences [18].

Aside from these more prosodic and paralinguistic phonetic correlates, cues to gender have also been found for more segmental and grammatically motivated phonetic properties. As one generalisation in this domain, segmental-phonetic evidence has been reported that female talkers tend to speak "more clearly" than male talkers. For example, based on an analysis of the TIMIT database, [19] found that females produce fewer instances of flapping or final stop deletion than males and that they speak more slowly than males. As a related tendency it has been found in many sociolinguistic studies that women use more standard-like and more "correct" pronunciations ([20] for overview). However, this tendency in female speech towards the linguistic standard of a language is not exceptionless, and it has been argued that this tendency is not a direct marker of gender but one that is mediated by differences in social variables such as class, prestige, or status, which can but need not cooccur with gender differences [20, 21].

It is possible that biological effects and sociological effects interact, whereby the tendency for clarity is more biologically motivated (see Note 6 for a perceptual explanation of clear speech) whereas "correctness" is sociologically motivated. Both clarity and correctness seem to be consistent and could be mutually enhancing.

Finally, it needs to be pointed out that cues to gender can also be found in domains of linguistics such as morphology, syntax, lexicon, and pragmatics. Some languages have strongly culturalised "genderlects" for male vs. female speech (for example, when certain lexical items are only tolerated among one gender group), whereas in many (most) others the gender differences are less categorical and more subtle.

2.2 Age

Age classification is a much more difficult task, not only because it involves more than a binary decision but also because there is not the kind of single dominant phonetic cue available that exists in gender determination. Due to the criminal statistics, age classification is most relevant within the range of about 20 to 50 years – most importantly among men. Consequently, the phonetic age correlates of adolescents or the elderly are only of limited importance in forensics. In our age classifications we distinguish between "chronological age" and "biological age". Chronological age is the calendar age as determined by a person's date of birth. Biological age refers to the aging level of the relevant organs and physiological mechanisms that are relevant in speech production. Biological aging in this sense can be accelerated, for example, due to abuse or overuse of the vocal folds through factors such as smoking, alcohol consumption, psychological stress/tension, or frequent loud/shouted speech production without vocal training ([22] for a comprehensive overview of vocal aging). As a rule of thumb, biological and chronological aging is patterned like a pair of scissors (thanks to Angelika Braun for this rule and metaphor): at an earlier chronological age (around 20 to 25) chronological and biological age are very similar, that is, a person of that age group usually sounds as old as s/he really is. With progressing age chronological and biological age can diverge. There are those speakers who take good care of their voices, and biological aging proceeds consistent with chronological aging, but there are also those who are abusive to their voices and therefore show more rapid biological than chronological aging. In our *Gutachten* (expert witness reports) we report the biological age, and if we find any indications why biological age could differ from chronological age in the case at hand we mention this in our report.

Age classification in forensic phonetics and acoustics is most commonly based on the overall auditory impression of the speaker, without performing any further phonetic analysis. Research has shown that listeners are to a certain degree able to estimate the age of the speaker ([22, 23] for overviews). [23] has investigated whether experts in forensic phonetics and acoustics are more accurate in estimating speaker age than listeners with no training in phonetics and no forensic experience. Among experts, the difference between perceived age and chronological age was 5.9 years on average, whereas among non-experts the average difference was 6.5 years. According to these results the experts performed better in age estimation, but not by much. When subdividing the data according to whether the speakers were smokers or non-smokers, experts showed the same performance as non-experts for the voices of smokers (both with 4.7 years deviation) but were better for non-smokers (8.4 years deviation for non-experts, 7.1 for experts). These data show that auditory age estimation is informative, but that it is not very accurate. Being aware of this situation, we report our age estimations in terms of ranges between about 10 or 20 years, that is we say, for example, that a particular speaker is between about 25 and 40 years old. In addition to this form of holistic age estimation it would be desirable to use phonetic and acoustic analysis to supplement the auditory impressions. In some cases age correlates might also be found in domains such as lexicon and stylistics. Clearly, the need for more research on the phonetics and acoustics of vocal aging is indicated.

2.3 Dialect

Dialect classification in forensic phonetics and acoustics involves the task of estimating – as much as possible given the available evidence – the region in which the relevant speaker has spent most of his life before the onset of adulthood. This goal is based on the assumption that after that time (we call it *Sprachprägephase* 'language acquisition period') speakers will not change their dialectal patterns by much, even if they move to a different region (although there are exceptions to this generalisation).

In German dialectology two major layers of dialect can be distinguished – deep and shallow. Deep dialect is spoken when individuals from local and mostly rural communities – especially those of the older generation – talk to each other, particularly on local issues. This form of deep dialect is the subject matter of by far most of the published dialectological studies ([24] for overview). However, deep dialect is only of very limited importance in forensic phonetics and acoustics, especially in institutes like the BKA, where due to high case load we have to concentrate on high crime, and cases with deep dialect often do not satisfy this criterion. More relevant forensically are shallower layers of dialect. One of these shallow layers is referred to as *regionale Umgangssprache* 'regional colloquial speech'. This roughly corresponds to the form of regional dialect that occurs when individuals speak to others from outside the local community but do not attempt to approach Standard German (even if they could), perhaps because the situation is considered too casual for that, or when they communicate among members of larger urban dialect areas without detailed dialectal subdifferentiation. The other type of shallow dialect is called *dialektal gefärbte Standardsprache* 'dialectally coloured standard German', which occurs when someone with a dialectal background attempts to speak standard German but is unable to achieve it fully. Much less literature is available on these shallow layers of regional differentiation. In order to compensate for this lack of documentation and in order to have the advantages of a database, DRUGS (*Dialektdatenbank Regionaler Umgangssprachen* 'database of regional colloquial speech') was created during the 1990s at the BKA [25]. More recently, another dialect database project has been initiated by the BKA in collaboration with the institute *Deutscher Sprachatlas* at University of Marburg, Germany. This system, which is called DIGS (*Dialektdatenbank gefärbter Standardsprache* 'database of (dialectally) coloured standard German'), contains data from conversations in emergency calls, where speech of the local emergency agents was recorded (in addition, read speech of a standard text will be elicited). This new database will reflect a more current state of regional German (DRUGS was mostly based on recordings from the 1950s and early 1960s) and has a tighter network of locations.

Phonetics and phonology are the most important disciplines in differentiating regional colloquial or dialectally coloured speech patterns. This differs from the analysis of deep dialects where linguistic domains such as morphology and lexicon are about equally important. In our laboratory and probably most other comparable institutes phonetic analysis for the purpose of dialect classification is primarily based on auditory analysis. One common method is to take the standard pronunciation, as documented in the pronouncing dictionaries, as a baseline and transcribe and analyse any deviations from that baseline.[4] Acoustic-phonetic analysis can be a useful

[4] Establishing such a baseline upon which dialectal deviation can be documented is complicated by the fact that the pronouncing dictionaries (e.g. [26] for German) usually provide transcriptions only for words as they are spoken in isolation. Connected speech phenomena on a sentence and discourse level are usually only explained briefly (and they could never be treated as explicitly as word-level phonetics given the infinite number of possible syntagmatic combinations). Yet it is clear that connected speech can contain numerous violations of word-level phonetic standards ([27] for many such connected speech phenomena in German) and it is equally clear that these violations are not necessarily dialectally motivated nor in any other ways "unusual". Therefore it depends on the experience and phonetic education of the expert to be able to establish a baseline upon which phonetic deviations are noteworthy for speaker classification and speaker identification purposes.

addition to auditory methods. One topic where acoustic analysis has proven useful in practice is monophthongisation in German dialects. Here formant structure has revealed monophthong status in situations where auditorily the situation was less clear.

2.4 Foreign Accent

As is well known, the term “accent” has multiple readings and can lead to confusion (even when disregarding the meaning of word- or sentence-level prominence). Accent could be used to address characteristics of (shallow) dialect on a phonetic-phonological level. Alternatively, the term can be restricted to foreign accent, which is the option that will be taken here. When talking about foreign accent we refer to L1-influenced deviations from L2 on all linguistic levels, not just phonetics and phonology. Foreign-accented speech has become much more important in our casework over time because in the cases we receive the number of criminal offenders with native languages other than German has increased considerably.

It is usually easy to determine that phonetic-phonological or other deviations from the target language (in our case Standard German) do not have a dialectal origin but must be due to foreign accent. It is much more difficult to infer the actual native language of the speaker that has caused the observable interferences with German. Often only broader characterisations of the native language can be given, for example by narrowing down the range of probable native languages to those of Slavic origin. One reason for these limitations in the accurate estimation of a native language can be the short duration of the available speech material. In case of short material, many distinctive properties of a certain foreign accent might not be determinable because lexical items or syntactic patterns with the relevant “diagnostic” structural configuration might not be present. It should be noted that not only negative transfer but also positive transfer can be informative. If for example – as has occurred in a case – the speaker has no problems producing the front rounded vowels of German, which are relatively rare cross-linguistically, it is probable that the native language of the speaker also contains front rounded vowels (such as Turkish).

The task of accent location is more complicated if the speaker has a multilingual background where more than the target language German and one native language are involved. Another complication that is becoming increasingly more severe in Germany is the use of “ethnolect” among fully native speakers of German. This is a form of foreign accent consisting of elements from mainly Turkish, which is featured in the media as “Kanak Sprak” and is popular among the younger generation of native and nonnative German speakers, especially from lower social classes [28]. As has become relevant in some casework recently, due to the increasing popularity of ethnolect it is not always possible to be certain about the presence of genuine foreign accent. Furthermore, ethnolect might blur the differences between real foreign accents among younger speakers with different L1 background.

Two other types of information beside accent location that can be useful in a police investigation are the competence level that a non-native speaker has in the target language and how long s/he is likely to have lived in the country where the target language is spoken. The competence level can be captured as the amount of deviation from the standard target language on different linguistic levels (small deviations indicating high competence). Duration of presence in the target country can be estimated from a high degree of fluency, e.g. a fast speaking tempo or the absence of dysfluencies and speech errors (as opposed to language errors). Fluent speech among non-native speakers does not imply that the level of competence in the target language must be high. As is well known from research on bilingualism, speakers can "fossilize" their language competence at a certain level that they consider sufficient for their everyday communication needs and in that case fluency may be high but competence low.

2.5 Sociolect

Of the different classifications relevant to sociolinguists some have turned out to be of particular interest in forensic work.[5] One of these sociolinguistic variables is the education level. Many clues to the level of education come from linguistic domains such as lexicon, syntax, and stylistics. For example, a high proportion of loanwords, a complex syntax, including subordination constructions, or a generally eloquent speech style are more likely to be found with higher than lower education levels. Pronunciation might also offer clues, for example if someone from an area where dialect is pervasive throughout the population (like in the Swabian or Bavarian area of Germany) speaks with little or no dialectal influence, this might be an index of higher education (however conversely, presence of dialect in these areas needs not imply low education because these dialects have high prestige).

Another sociolinguistic variable is the profession of the speaker. Some professions use specific terminology (cf. also the notion of *Fachsprache* 'language of specific purposes'[30]). Such terminology might slip into conversations that are unrelated to professional discourse. [7] mentions a case of blackmail where the anonymous caller used the German word *Langsamfahrstrecke* 'low-speed railway section'. This turned out to be specialized railway terminology, which let the suspicion arise that the caller was professionally involved in the railway system. Another example from our recent work was the use by a non-native speaker of the word *Qualitätskontrolle* 'quality control' in a case where a company was blackmailed.

[5] The field of sociolinguistics and the notion of sociolect can be defined in narrower and broader terms (for recent overview with special emphasis on phonetics and an acknowledgement of the importance of sociophonetics to forensic phonetics see [29]). If defined in broader terms categories such as age and gender can be included in the range of sociolinguistic variation. Sociolect can also be defined to include regional variation. In our casework a narrow definition of sociolect is assumed and categories such as gender, age, and dialect are treated separately. It is also possible to divide the categories age and gender into a biological component and a sociolinguistic component. In that case the latter category can be terminologically divided into "gender" (sociolinguistic) and "sex" (biological). In practical terms we do not make this distinction although we are aware that clues to age and gender in speech can be found both in biologically motivated and sociolinguistically motivated patterns.

The two variables education level and profession can also interact. This is the case when the level of eloquence is high enough that it can be inferred that the person is a "professional speaker" (e.g. business consultant, teacher, lawyer) or if the level of vocabulary and syntax is high enough that one can infer an influence from written language that points to a "professional writer". Conversely, particularly low levels on these dimensions might indicate that the speaker is not used to professional discourse or to writing (such as a construction worker or a farmer).

2.6 Medical Conditions

Indications of speech pathology – used here as a cover term for disorders of language, speech, and voice – can be extremely helpful for voice comparisons and voice analyses because they can narrow down the number of speakers considerably. Some speech pathological characteristics are a priori more informative forensically than others.

First, speech pathological characteristics that are more or less stable over time are more useful than those that are very short lived. One example of the latter type is (acute) laryngitis, which can result in voice source changes such as rough or breathy voice and a reduction of average pitch, but also changes in resonance [31]. Through medical treatment, voice rest, or abstinence from smoking the laryngitis can decay over a relative short time period. If for example – as has occurred in a recent case – no symptoms of laryngitis are present in the unknown recording but such symptoms turn out to be present when the voice sample of the suspect was taken, the voice comparison can become difficult and any voice characteristics related to laryngitis have to be dismissed because they are transitional and not characteristic of the speaker. To the extent that transitional voice disorders like laryngitis are difficult to distinguish from more permanent voice disorders based on auditory and acoustic evidence alone, voice disorders are generally of limited use in speaker classification.

Second, for speech pathological characteristics to be useful forensically they must not have a strong negative impact on general cognitive or communicative abilities. Some disorders of language, speech, and voice might affect other cognitive and physical capacities to a point that it is very unlikely that a person inhibited in this manner could commit the types of crimes that are relevant in our casework. Many brain disorders fall under this category. The speech pathological disorders might also by themselves be so severe that the sort of communications that are relevant forensically cannot be carried out. To take a drastic example, a person who suffers from anything more than the mildest forms of aphasia will not be able to conduct a negotiation about drug trafficking or the placement of ransom money. Related to this point, speech pathological conditions that occur only or predominantly at young or at old age will usually not be relevant forensically for obvious legal or criminological reasons.

Forensically useful speech pathological conditions that remain when these two criteria are applied are those that are very stable over time (for example due to resistance against therapeutic and other treatment) and do not strongly inhibit communication nor are associated with severe cognitive or physical deficits. One condition with this profile is stuttering – as long as the degree of stuttering is still relatively mild. According to studies quoted by [32] in a review paper, stuttering occurs in about 1% of the world's population. If this number included stronger forms of stuttering which would strongly inhibit communication (and hence violate one of the two mentioned criteria), the percentage of mild stutterers, that are relevant forensically, would be even smaller than 1%. Due to this rare occurrence of stuttering, finding stuttering in a speech sample can be a very informative characteristic towards finding a suspect. It can also be of high evidential value in voice comparisons – both towards identity, if stuttering is found in the sample of the unknown speaker and the sample of the suspect, and towards non-identity, if stuttering is found only in one of the samples. However, in order to determine the presence of stuttering the forensic specialist must be aware of other medical and non-medical conditions that can be confused with stuttering. If in doubt, a physician or speech-language pathologist should be consulted. The ability to detect stuttering also depends on the duration and quality of the material. In a case processed by the author the technical quality of the material was so low that intelligibility was reduced in the relevant passages. Due to this loss in intelligibility, passages that seemed to contain frequent syllable repetitions also allowed alternative interpretations where the spoken material was unmarked. Due to these difficulties no statement about whether or not stutter was present could be made. In addition to sufficient quality, the quantity of the material also has to be sufficient. With very short recordings it usually cannot be determined whether non-fluencies are within the range of normal behaviour, where such non-fluencies are possible as well, or whether they approach the quality and quantity of speech pathological behaviour.

Another type of speech pathological characteristic that has become relevant in forensic casework is the presence of articulation problems that are caused by disorders or delays in the acquisition of the sound system. Different sounds have different probabilities of being affected, and one of the most frequent cases are mispronunciations of the sound [s], often referred to as "sigmatism" [31]. In cases where this phenomenon might be involved it has to be ensured that the frequency range of the recording is sufficient in order to be able to distinguish a normal from a deviant [s]. One way of estimating the impact of frequency range and other technical conditions would be to select normal and deviant s-sounds from a medical reference recording and filter them in a way that simulates the frequency range and other conditions of the recording in the specific case. It can then be tested auditorily or acoustically if the pathological s-sound can still be identified and distinguished from the normal one.

Deviations from expected non-pathological speech patterns might also be caused by conditions that require medical attention but that are not classified as disorders of language, speech, or voice (hence the cover term "medical conditions"). One such case are breathing sounds that indicate pulmonary problems or are due to obstructions in the passage of air that occur at the level of the larynx or above. Deviant breathing sounds can be very distinctive auditorily and acoustically ([5] for spectrographic illustration). However, it needs to be considered that these breathing problems might be transitional, similarly to what was said about laryngitis above.

3 Current Issues in Forensic Speaker Classification: Auditory vs. Acoustic-Phonetic Analysis

3.1 Introduction

There is a tradition in forensic phonetics that speaker classification is based on auditory rather than acoustic methods – especially when carried out for the purpose of voice analysis. Among the reasons for this preference there are three practical aspects that have been of importance in our lab. First, voice analyses often have to be carried out very rapidly because they are required as important information in an ongoing investigation where lives can be at stake. Usually at least a preliminary speaker profile is expected within one day. In line with this requirement, auditory analysis allows a fast access to the speech material, whereas acoustic analysis might be much more time consuming. Secondly, there have been situations where the speaker identification team was required to perform a voice analysis outside the lab and at the particular location where an urgent investigation was directed. In such a case it would have been unrealistic to carry along the entire equipment necessary for acoustic analysis. This point was more urgent about ten or more years ago, where good speech analysis systems were not as portable and where electronic communication and data transfer was not as widespread and easy as it is now, but the argument still carries some weight. Thirdly, the quality of the material on which voice analyses have to be carried out can be very poor at times. Sometimes these quality reductions are so profound that the material cannot be used for a voice comparison. One example of these quality reductions is overlap between voices, where the task has been to profile the speaker in the background, not the one in the foreground. In cases like that it is possible that speaker classification information can still be extracted auditorily where acoustic measurements would be unreliable.

In spite of this well-founded tradition of preferring auditory over acoustic methods in forensic speaker classification it is important to keep an open mind considering arguments that emphasize the theoretical and practical value of acoustic methods as well.

3.2 New Information on Established Speaker Classification Characteristics: Gender

For some speaker classification characteristics acoustic analysis might add accuracy and objectivity but does not necessarily advance the task. Consider gender classification. As was mentioned in 2.1 the most powerful parameter to identify the gender is the average pitch level of the speaker. In most cases the pitch level difference is sufficiently large that auditory examination will enable a male and a female speaker to be accurately distinguished. Furthermore, there is a very transparent relation between the perception of pitch (and the description of this perception by the expert) and the measurement of fundamental frequency, to the effect that measurement of average f0 can add detail but is still concerned with the same phenomenon of pitch level differences between the sexes. Consider now the situation where pitch is not accessible or informative, perhaps because as a voice disguise the speech sample was whispered or produced in falsetto voice or creak, or perhaps there is reason to believe that the speaker under analysis has an unusually high or low voice compared to his gender group and that confusion with the opposite sex based on pitch level is very likely. In such a situation acoustic analysis can make an important contribution by providing formant frequency measurements. As is well known, women on average have higher formant frequencies than men [33, 34], due to the fact that the vocal tracts of women are on average shorter than those of men [35].[6] With formant data from a large number of male and female adult speakers the expert could make a scientifically motivated statement as to whether the formant values measured in one of these difficult cases are more likely to belong to a man or to a woman. (Gender determination in non-adults is still another story but only of limited importance forensically.)

From the auditory perspective it could be objected that gender perception is possible to a certain degree even without reliable pitch information and that in this perception process formant structure makes some contribution [39–41], so that acoustic measurement is not necessary. However, despite this ability of gender perception in the absence of reliable pitch information and the contribution of formant structure to this perception process there is still the following difference to the domain of average pitch. As mentioned above, measured f0 and perceived pitch stand in a very direct and transparent relationship, so that for the purpose of gender classification one source of information is largely redundant. For vocal tract characteristics of men vs. women, on the other hand, it is unclear what the perceptual correlate of formant frequency is and how it should be described. In contrast to pitch, where a voice with a high f0 can be described as "high" or "high-pitched" it is not clear how a voice with high formant positions should be described by the expert. One possibility would be to call such a voice "light" (or "bright") and, conversely, to call a voice with low formant frequencies "dark" (see also [16] with reference to singing).

[6] It has been found that female vowel spaces are not just uniformly upshifted in frequency relative to male ones, but that female vowel spaces also tend to be larger. This is mainly due to the fact that the difference between male and female formants is larger for /i/ and /a/ than for /u/. [36] present a review of non-uniform scaling of female and male formant frequencies and discuss previous anatomical explanations related to the fact that in adult males the pharynx takes up a greater proportion of the entire length of the vocal tract than in females. They conclude that this anatomical explanation cannot capture all the observable male-female differences in vowel spacing and that behavioural aspects must be at work at well. They formulate and provide support for the "sufficient contrast" hypothesis. According to this hypothesis female speakers need to produce more peripheral vowel targets (and in this sense need to speak more clearly than men) in order to compensate for the reduction of intelligibility that results from the fact that the high average f0 values in their speech implies an increase of harmonic spacing which leads to reduced accuracy in capturing the peaks of the vocal tract transfer function (i.e. the formants). [37, 38] makes the point that due to the anatomically given smaller cross-sectional distances that have to be covered in the articulation of female compared to male speech, it is easier for females to produce those more peripheral vowel targets compared to males. In this sense anatomical and behavioural factors interact in interesting ways.

More technically, one could also use the auditory feature terminology of [42] and call a light voice “acute” or “sharp” and a dark voice “grave” or “flat”.[7] Interestingly, they use “tonality features” as a cover term for the features (or feature values) grave, acute, sharp and plain, which underlines that some form of perceptual supralaryngeal tonal property is at work. However, terminology for supralaryngeal tonality is not as widespread in phonetic research and teaching as high vs. low laryngeal pitch. For example, the IPA provides symbols for intonational and tonal pitch features, but has no symbol or diacritic for light vs. dark voices due to vocal tract length.[8] This does not mean that the difference between light vs. dark voices due to vocal tract length differences is not perceivable. That it is perceivable can clearly be shown, for example by some speech and music processing programs, where the user can set different formant values appropriate for vocal tracts of different length and where a clear perceptual difference is obtained that many listeners could agree upon to call light vs. dark. The point is that auditory impressions that arise from differences in vocal tract length (which are usually strong between men and women and more subtle within the sexes; cf. following section) and that might be labelled on a light-dark dimension are not well enough captured in phonetics so that it is possible that two phoneticians would make very different assessments along this dimension because they are not trained in its use. Formant measurements, on the other hand, promise much greater consistency across specialists and are also more precise than auditory judgments. These are reasons why acoustic analysis in terms of formant

[7] Flat/sharp is probably better suited for present purposes than grave/acute, since, as far as vowels are concerned, the correlates of grave/acute are more concentrated on the second formant (high with acute vowels, low with grave vowels) whereas flat/sharp in principle affects all formants (low formant frequencies with flat vowels, high formant frequencies with sharp vowels). In vowel systems, grave/acute corresponds to the back/front distinction, respectively, whereas flat/sharp most commonly corresponds to the rounded/unrounded distinction. The fact that lip rounding in conjunction with lip protrusion leads to an effective increase in vocal tract length suggests that auditory flatness is the most appropriate way of capturing vocal tract length differences that result from other sources as well, viz. those from differences in gender, speaker height, and speaker identity (see also [43] on the feature [flat]). On the other hand, terminologically “light/dark” is most closely associated with “grave/acute” (see [44] who trace the use of this terminology back to the work of the psychologist Wolfgang Köhler in the early 20th century). [44] also point out that grave/acute is relevant in sound symbolism, where front vowels tend to occur in words denoting small size and back vowels in words denoting large size (likewise, see [45]), and that it is relevant in synesthesia, where an association has been found between front vowels and light colours as well as between back vowels and dark colours.

[8] It might be argued that different formant values are captured by the IPA in the form of different vowel symbols and their diacritics. However, these vowel symbols can be used by phoneticians only because phoneticians, just like any other language user, have performed an unconscious process of vocal tract or talker normalisation [18]. This includes normalisation against differences between male and female vocal tracts, which themselves are not captured by the IPA. Probably the closest thing to the perceptual characterisation of vocal tract length consistent with an IPA framework is the “longitudinal setting” [46], comprising larynx lowering/raising and labial protrusion/labiodentalisation (cf. the “VoQS chart”, addressed in [47]). However, these settings refer more to behavioural modifications of vocal tract length than to organically different vocal tracts [48].

measurements promises additional information at least in those cases of gender classification where pitch cues are absent or unreliable.[9]

One could go farther and recommend acoustic, rather than merely auditory analysis even for pitch if pitch is present and reliable. This would be particularly informative when background statistics are available containing average f0 values of large numbers of men and women [5, 14]. If, in addition, formant frequencies are measured as well there would be the opportunity of testing whether the gender classification that is obtained on the basis of f0 information is consistent with the one obtained on the basis of formant information. In terms of measurement methodology, formant frequencies are fairly independent of f0 characteristics, although in practice there are certain limitations due to the fact that the transfer function peaks of the vocal tract might lie at different locations with respect to the harmonics of the source spectrum; this is also a reason why formant measurements with very high-pitched voices (which could also arise because of loud or shouted speech) are difficult.

Whether such high degree of independence in the detection of source and filter characteristics also obtains on the perceptual level is not so certain. It is possible that listeners have difficulties separating "high-tonality" at the source level (i.e. high vs. low voices) from high-tonality at the filter level (i.e. light vs. dark voices) and that to a certain extent they integrate the two into a single general pitch percept. [51] has shown that the perception of differences between vowels, which where synthesized using variations in formant frequencies, was not independent from the perception of synthetically generated differences in f0. Analogous results for the integration of vocalic (and consonantal) differences on the filter level and f0 – as a source variable – have been reported by [52]. It is possible that such a perceptual integration is enhanced if the two relevant properties commonly exhibit covariation and the listener has knowledge about this association (either innately or by learning). This would be the case for gender perception, where low/high pitch usually goes hand in hand with low/high formant values. Evidence for perceptual integration between formant frequencies and f0 in gender perception has been found by [53] using a selective adaptation paradigm. Perceptual integration has also been found in other aspects of source-filter interactions. [54] found that perceptual integration occurs between advanced tongue root – cued by low F1 (filter) – and lax/breathy voice quality – cued by increased spectral tilt and open quotient (source).

The point of this discussion is that when untrained or even phonetically trained listeners make an assessment of the speakers' gender by attending to the speech material in a holistic fashion it is likely that they take into account both information

[9] When measuring formant frequencies in non-modal voice sources such as whisper and creak one has to keep in mind that those voice sources might change the formant frequencies to different values compared to modal voice. As for creaky voice, [49] has shown that this voice quality can change formant frequencies, though not by very much. Probably even less influential is whisper. Although because of tracheal coupling, which occurs in whisper due to glottal opening, there can be a change in the amplitudes and frequencies of formants (and the addition of "tracheal" formants) the only formant strongly affected is F1, and that formant is affected more in its amplitude than its frequency [50]. Falsetto voice can be more detrimental to formant measurements, mainly because of wide harmonic spacing that occurs with very high-pitched voices. It also needs to be considered that falsetto might change formant patterns, e.g. due to larynx raising.

about the source and about the vocal tract. Even when phonetically trained listeners proceed analytically by attending separately to the tonal contribution coming from the voice source and the tonal information coming from vocal tract length differences they might not be fully able to “fight” the perceptual integration forces between the two types of tonality. Under normal circumstances there is nothing wrong with such a perceptual integration because usually men have both lower laryngeal and lower supralaryngeal tonality than women. In that case the perceptual integration of laryngeal and supralaryngeal contributions can actually enhance the perceptual recognition of masculinity vs. femininity. However there can be situations where it is advisable to examine the two contributions separately. Voice disguise or one of the “gender bender” scenarios mentioned in 2.1 can result in the fact that laryngeal and supralaryngeal correlates of gender are no longer correlated. For example, the tonality of the voice source might be unusually high for a transsexual man who has undergone hormone treatment and surgical reduction of the vocal folds, but vocal tract length cannot be modified in the same fashion and its acoustic correlates reveal his original masculinity. Likewise, inconsistencies found by careful separate examination of supralaryngeal and laryngeal characteristics might reveal a form of voice disguise where the person tries to sound like somebody from the opposite gender by changing laryngeal tonality, but does not know about the importance of the supralaryngeal contribution to gender perception. That the opposite situation can be found as well has been shown in a recent case involving a bomb threat at an airport. As was revealed after the fact, the woman making the bomb threat produced lip rounding (and perhaps larynx lowering) in order to sound more masculine.

In general, to the extent that it is advisable to perform gender classification analytically by separate examination of laryngeal and supralaryngeal tonality, acoustic analysis is probably better suited for this task because the source information can be separated better from the filter information with acoustic methods than with auditory methods.[10]

Having concentrated on average f0 and formant structure, as well as their perceptual correlates, and on gender classification, the possible importance of using acoustic in addition to auditory information can also be examined with other parameters and classification domains (e.g. age). Let us briefly consider voice source

[10] Commenting on this paper, Olaf Köster mentioned the possibility that another difference between male and female speech lies in vocal loudness (vocal effort), whereby men tend to speak more loudly than women. Such a difference would be consistent with an ethological framework [55, 56] where louder speech would signal larger size and strength, for example in terms of larger lung volume and pulmonic force. As supporting evidence for this view it has been found by [57] that among various voice quality parameters the strongest gender difference was in terms of H1*-A3*: men had smaller values of this parameter and hence a less-steep spectral tilt than women. Although they attribute this difference to more glottal leakage in female than male speech (which is another, partially independent male-female difference, especially if turbulence is involved), the evidence could also be interpreted – at least on the perceptual level, where the ethological explanation is most relevant – as a difference in vocal loudness, whereby men speak more loudly than women. Unfortunately, such a spectral tilt difference is hard to measure under forensic conditions, as will be addressed separately towards the end of this section.

characteristics such as roughness (also called harsh voice) on the perceptual level and jitter on the acoustic level. As has been found in many studies, there is a certain correlation between jitter and roughness, but this correlation is by no means perfect or in a one-to-one fashion. Auditory roughness is also correlated with other acoustic parameters such as harmonics to noise ratio (negative correlation), and jitter also cues other auditory voice qualities such as breathy voice [58] (see also [59] within a forensic framework). In such a case, providing both auditory and acoustic data for the same general voice quality is not redundant but provides additional information. The idea would be that providing both roughness ratings and jitter measurements can provide more accurate speaker classification information than if only one of them is used. Furthermore, reference information on speaker classification characteristics in the literature might be expressed in acoustic rather than auditory terms (e.g., jitter values for different ages, reported by [60]). This is a more general argument for the inclusion of acoustic methods that also holds for average fundamental frequency. As far as voice source characteristics are concerned it has to be pointed out, however, that due to the common quality limitations of forensic material – especially in terms of bandpass filtering due to telephone transmission and the presence of technical or environmental noise – certain voice quality parameters cannot be measured at all (e.g. "H1", the amplitude of the first harmonic, especially in men) or their measurement and interpretation has to proceed with great caution (e.g. spectral tilt, breathiness turbulence) (cf. [48]). What is still largely unknown is the way and extent to which measurements of jitter (or shimmer) are affected by the types of adverse conditions found in forensic work. [61] presents relevant results but also points out the need for more research.

3.3 New Speaker Classification Characteristics: Speaker Height

Some speaker classification characteristics might be inaccessible or unreliable auditorily but accessible and more reliable by acoustic methods. Consequently, without acoustic analysis such a speaker classification characteristic could not be used at all. In this subsection it will be discussed whether such a situation obtains for the classification of speaker height.[11]

Research by Lass, including [62], presented an optimistic picture about the ability of listeners (including phonetically untrained ones) to estimate the body height of a speaker. However, [63]) showed that this result was mainly due to the fact that results for female and male speakers had been analysed together rather than separately, hence conflating height perception with aspects of gender perception. When he reanalysed the raw data of [62], calculating correlations between perceived and real body height separately for male and female speech, the correlations dropped below the level of statistical significance.

In [64] a new empirical study on the issue was reported where height measurements and read-speech recordings were made of 15 male and 15 female

[11] Some studies address not only speaker height but also speaker weight and usually find similar acoustic and perceptual effects with weight as with height. For reasons of space the influence of speaker weight will not be discussed here but the reader will find further information in the cited literature.

Norwegian-speaking students. Height (and weight) was estimated by 10 male and 10 female student listeners (most likely phonetically untrained) and acoustic measurement were made. With this design the researchers were able to determine correlations between each pair of the three types of evidence: actual height, perceived height, and acoustic measurements. Significant correlations between perceived and actual height were essentially limited to those situations where males listened to male speech. Female speech caused no significant effects – whether perceived by males or females – and female listeners showed only very limited ability to estimate males' body height. Despite this restricted pattern of correlations between perceived and actual height, correlations between perceived height and acoustic parameters were generally stronger and more uniform across male and female speakers and listeners. Every existing significant correlation showed that large perceived body height is associated with low average f0 and long vocal tract (acoustically derived using the F2 value of the neutral vowel schwa) and small height with high f0 and short vocal tracts. The fact that the correlations between perceived height and acoustics were greater and more uniform than the ones between perceived and actual height indicates that the listeners were sometimes mislead by the probably unconscious expectation that low/high f0 and long/short acoustic vocal tract is an index of large/small body height, respectively.[12] Such an expectation is a problem for the perceptual estimation of body height: it introduces some form of stereotype about the relations between body height and vocal characteristics that can guide the estimation of height into the wrong direction. A similar pattern of stereotype-guided perception is shown by [66], who had females listen to male speech and found high correlations between low f0 and perceived maleness attributes such as being muscular and having chest hair. She also found a significant correlation between low f0 and large perceived height although the correlation between f0 and actual height was not significant.

Having shown that the ability of human listeners to estimate body height is very limited and subject to bias the question arises whether acoustic phonetics offers a better access to this speaker classification characteristic. Previous studies – including [14] as well as [66], just mentioned – have shown that average f0 is not a reliable correlate of speaker height at all. This result has been confirmed again in [64], who reported that correlations between body height and average f0 were non-significant. On the other hand, vocal tract size, as indicated by formant patterns, offers a more promising correlate of speaker height and deserves closer attention.

[64] found generally no significant correlations between formant-based vocal tract length and body size but report one exception where a significant correlation (with

[12] [64] suggest that evolutionary factors might be responsible for this expectation bias. Referring to [55] they mention that in the communication systems of many animals low frequency sounds (like in this case: low f0 and low formant positions) signal large, potentially threatening, vocalizers (see also [56]). As far as the relation between vocal tract size, body size, and formant structure is concerned, the ethological foundations and implications of this pattern have been worked out in detail by [65]. [64] suggest that evolutionary factors could also explain their result that only when males listen to male voices the height estimation is reasonably accurate. This pattern would be consistent with a scenario where male hominids (and many other animals) had to have a realistic estimate of the body size and hence degree of threat of male rivals, and that in some situations acoustic cues were the only ones available.

longer vocal tracts indicating larger bodies) was obtained among female speakers as they were reading one of two text paragraphs. They note, however, that not too much emphasis should be given to this exception. [67] does not agree with the assessment of [64] that the significant correlation between height and vocal tract length among females is accidental. He claims that the total number of 15 speakers per gender was too small to detect significant correlations between height and vocal tract length, where such correlations could have been shown to exist with a larger speaker corpus. He also criticises that the range of the population height was too small and that vocal tract calculations were solely based on F2 while disregarding other formants.

[67] carried out a speaker height study of his own based on 48 female and 43 male speakers (mostly students) who produced the vowels of German in isolation. Instead of inferring vocal tract length from formants he worked with the formants F1 to F4 directly, as well as with various combinations of formants, and made formant measurement separately for each vowel. Except for F1, he found negative correlations between formant frequency and body height (which is the expected pattern since formants decrease with an increase in tube length). For the vowel [ø:] he reports that he could establish a regression line and that the correlation was statistically significant. (He mentioned that the correlation was strongest for this vowel, but does not address systematically which other vowels achieved significance with which formants.) Discussing why the strength of the correlation differs between vowels, he mentions that with [ø:] the formants are well-spaced. A similar beneficial formant spacing could occur with schwa, which was used by [64] but was not included in the stimuli of [67] (perhaps because schwa can hardly be spoken in isolation by phonetically untrained subjects without changing its character). If it were the case that only the vowel [ø:] offered significant correlations between formant frequency and body height the practical implications would be limited because the token frequency of this vowel in German running speech is quite low. But it would be interesting to find out whether significant correlations also apply to at least those other vowels that have a clear formant spacing, such as schwa or the “äh”-type vowels used in filled pauses. Another interesting issue is the existence of differences between male and female speakers. From the correlation coefficients [67] reports (for the vowel [ø:], as shown in his Table 3) it can be seen that for each individual formant and for all combinations of formants where a negative correlation between formant and height are found (hence excluding F1) the correlation is slightly higher in female than male speech.

In his conclusion [67] makes a very useful comment when he says that although the generally weak correlations between formant frequency and body height prevent any precise prediction of the latter from the former, it can be said at least that “low formant frequencies make it very unlikely that a short person is responsible for their production, and high formant frequencies make it equally unlikely that a tall person is involved” (p. 276). A statement like this would be quite useful in a police investigation, where it could help to narrow down the range of possible suspects.

The relation between body height and formant frequencies was addressed again by [68] in two experiments. In the first experiment 22 male and 55 female Spanish-speaking students produced sustained versions of each Spanish vowel. In the second experiment running speech from a read text was elicited from 29 male and 62 female Spanish-speaking students, of which 60 were not included in the first experiment. In the first experiment F1 to F4 were measured for each vowel, in the second experiment the same formants were measured based on Long Term Average Spectra, in which the information for all vowels (as well as other speech sounds) was combined. Body height (and weight) was measured for each individual and the analysis of the results was performed separately for male and female speakers. In the first experiment, significant negative correlations between body height and formant frequencies were found for several combinations of formant, vowel, and gender. Most significant correlations were found for F2, followed by F3, significant correlations were only found for /i/, /e/, and /a/, not for /o/ and /u/ and females showed more significant correlations than males. Specifically, for female speech significant negative correlations between body height and formant frequencies were found with F2 in /a, e, i/ and with F3 in /e, i/. For male speech the only significant correlation was with F2 measured in the vowel /e/. In the second experiment significant negative correlations between body height and formant frequencies were found only for females. With them it was the formants F3 and F4 that yielded the strongest effects. Based on these results the correlation between formant structure and body height is much stronger for women than for men. The same effect was also found in [64] and [67], only that in those studies this gender effect was more subtle.

Opposite results as far as the gender effect on the influence of body height on formant frequencies is concerned were obtained by [69]. They recorded lists of isolated vowels, words, and sentences spoken in Canadian English by 34 male and 34 female students. Body measurements including height were performed. Formants F1 to F4 as well as f0 were measured of each of the English vowels. From the formant measurements "formant dispersion" (FD) was calculated as the mean frequency difference between successive formants (FD is predicted to increase with a decrease in vocal tract length), but the frequencies of the four formants were also analysed by themselves. Separate analysis was performed on the neutral vowel schwa. For the female subjects no significant correlations between height and f0, FD, or the frequency of any individual formant were found. For the male subjects, on the other hand, significant negative correlations were found for FD, both when all vowels were analysed together and when tokens of schwa were analysed separately (together with a non-significant contribution from f0). Consistent with earlier studies, f0 did not significantly depend on height. In a second multiple regression test where the values of individual formants were used instead of the overall FD parameter, the only parameter that was significantly (negatively) correlated with height was F4 when all vowels were considered. When separate analysis of schwa was performed a significant effect was obtained in which all four formants and f0 were involved, but where the strongest effect was for F4. These results contradict those of [68], where it was (almost always) the female and not the male population that showed significant correlations between formant frequencies and body height. From a practical point of view it needs to be mentioned that the success of F4 as a certain predictor of body height is of only limited use in forensic speech analysis since in telephone-transmitted speech that formant is so close to the upper frequency boundary of the telephone pass band (which usually lies at around 3500 Hz) that this formant cannot be detected at all or not be measured with sufficient reliability.

This literature survey has shown that body height is to a certain extent correlated with the length of the vocal tract as it can be predicted from the formant frequencies or the frequency distance between the formants (cf. also [35] for correlations with direct, MRI-based, measurements of vocal tract length and [69] for comments on that study). The size of this correlation depends on various factors, one of them being the gender of the population for which the height-formant correlations are obtained. According to one study [68] the correlation is reasonably strong for women but quite weak for men. If this were the result of all studies the practical value for forensic work would be limited due to the much higher number of males than females in the population of criminal offenders. However, another study [69] showed the opposite result and in two further studies [64, 67] the advantage of females was only slight. This gender issue and other open questions, as well as the important practical implications in forensic work, show that the association of acoustically inferable vocal tract characteristics with body height (as well as other body measures) should be investigated further. One line of research could be to investigate correlations between body height and anatomical properties other than vocal tract length. For example, if it turned out that the cross-section of the vocal tract correlates with body height, acoustic-phonetic correlates and consequences of individual differences in vocal tract cross-section could be investigated as indices of body height (cf. [37, 38] mentioned in Note 6).

3.4 Conclusion

In this section some arguments have been advanced for why forensic speaker classification can benefit from the inclusion of acoustic methods to supplement the auditory methods that are currently in practice. Focussing on gender and body height classification, where both pitch and vocal tract information are important, the essential arguments are as follows.

First, there is no well-recognized terminology for the perceptual effects of organic variations in vocal tract length. Without phonetic standards and training in the perceptual classification of vocal tract length this cue to gender and height classification cannot be accessed reliably if it were entirely approached with auditory methods. Formant frequency measurements, on the other hand, are a more reliable way of capturing vocal tract length.

Secondly, it cannot be assumed that the perception of the source property pitch and the perception of the filter property that results from variations in vocal tract length are fully independent. From the literature there is evidence that at least some form of perceptual integration occurs between these and other source and filter characteristics. Such a perceptual integration would entail that listeners have some difficulties distinguishing tonality that is caused by vocal fold vibrations from tonality that is caused by vocal tract filtering. Although there are aspects where source-filter separation is also difficult to achieve acoustically, it is likely that acoustic measurements of f0 and formant structure allow for a better source-filter separation than auditory methods.

Thirdly, the perception of gender and body height can be biased by the expectation that low/high f0 and low/high formant positions are cues of male/female gender and large/small body size, respectively. For the normal cases of gender perception these expectations are realistic. However in the more marginal areas of gender perception and in the presence of voice disguise such an expectation can be misleading. And as far as height classification is concerned, the evidence has shown that despite the presence of a certain correlation between vocal tract length and body height there is a large amount of decoupling between those two anatomical properties. A decoupling between (low) f0 and (large) body height is even much stronger. [65] argues that the expectation that large individuals have low f0 and low formants (which he also demonstrated based on a perception experiment with synthetic stimuli) might go back to an old and persistent trait in the evolution of communication (see also [55, 56]). The same is likely to be the case with gender perception. Listeners, even those who are phonetically trained, might not be fully able to fight such a phylogenetically founded perceptual bias. Acoustic measurements, on the other hand, are free from such a bias.

4 General Conclusion

This contribution has provided an overview of the role of speaker classification in forensic phonetics and acoustics. In an introduction it has been shown that forensic speaker identification essentially divides into the subdisciplines voice analysis, voice comparison, and voice lineup, and that speaker classification is relevant to all three of these. In the second section six different speaker classification characteristics that are commonly used in forensic work were addressed. Some of these characteristics have a stronger foundation in biological and organic factors (age, gender, medical conditions) and some have a stronger linguistic foundation (dialect, foreign accent, sociolect). The remaining part of the paper was dedicated to a discussion of current issues in forensic speaker classification, focussing on the question of whether acoustic methods can supplement the auditory methods that are currently in use. It was argued that there are situations and areas in forensic speaker classification – including the largely uncharted one of body height estimation – where the inclusion of acoustic methods can be very informative.

Acknowledgments. Many thanks to Kirsty McDougall, Olaf Köster, Stefan Gfroerer, and Nadja Tschäpe for comments on this paper.

References

1. Broeders, A.P.A., van Amselvoort, A.G.: Lineup Construction for Forensic Earwitness Identification: a Practical Approach. In: Proceedings of the 14th International Congress of Phonetic Sciences, vol. 2, pp. 1373–1376 (1999)
2. Broeders, A.P.A.: Earwitness Identification: Common Ground, Disputed Territory and Uncharted Areas. Forensic Linguistics 3, 3–13 (1996)
3. Nolan, F.: The Phonetic Bases of Speaker Recognition. Cambridge University Press, Cambridge (1983)
4. Nolan, F.: Speaker Recognition and Forensic Phonetics. In: Hardcastle, W.J., Laver, J. (eds.) The Handbook of the Phonetic Sciences, pp. 744–767. Blackwell, Oxford (1997)
5. Künzel, H.J.: Sprechererkennung: Grundzüge forensischer Sprachverarbeitung. Kriminalistik Verlag, Heidelberg (1987)

6. Künzel, H.J.: Field Procedures in Forensic Speaker Recognition. In: Windsor Lewis, J. (ed.) Studies in General and English Phonetics. Essays in Honour of Professor J.D. O'Connor. Routledge, London, pp. 68–84 (1995)
7. Künzel, H.J.: Die forensische Sprachverarbeitung. Ein Überblick über den gegenwärtigen Stand. Kriminalistik 11, 676–684 (2003)
8. Künzel, H.J.: Tasks in Forensic Speech and Audio Analysis: A Tutorial. The Phonetician 90, 9–22 (2004)
9. Hollien, H.: The Acoustics of Crime. In: The New Science of Forensic Phonetics, Plenum Press, New York (1990)
10. Hollien, H.: Forensic Voice Identification. Academic Press, San Diego (2002)
11. Rose, P.: Forensic Speaker Identification. Taylor and Francis, London (2002)
12. Gfroerer, S.: Auditory-Instrumental Forensic Speaker Recognition. In: Proceedings of EUROSPEECH 2003, Geneva, pp. 705–708 (2003) (on CD)
13. Gfroerer, S.: Sprechererkennung und Tonträgerauswertung. In: Widmaier, G. (ed.) Münchener Anwaltshandbuch Strafverteidigung. Beck, München, pp. 2505–2526 (2006)
14. Künzel, H.J.: How Well Does Average Fundamental Frequency Correlate with Speaker Height and Weight? Phonetica 46, 117–125 (1989)
15. Jessen, M., Köster, O., Gfroerer, S.: Influence of Vocal Effort on Average and Variability of Fundamental Frequency. The International Journal of Speech, Language and the Law 12, 174–213 (2005)
16. Titze, I.: Principles of Voice Production. Englewood Cliffs, Englewood Cliffs (1994)
17. Künzel, H.J.: Effects of Voice Disguise on Speaking Fundamental Frequency. Forensic Linguistics 7, 149–179 (2000)
18. Johnson, K.: Speaker Normalization in Speech Perception. In: Pisoni, D.B., Remez, R.E. (eds.) The Handbook of Speech Perception, pp. 363–389. Blackwell, Oxford (2006)
19. Byrd, D.: Relation of Sex and Dialect to Reduction. Speech Communication 15, 39–54 (1994)
20. Smith, P.M.: Sex Markers in Speech. In: Scherer, K.J., Giles, H. (eds.) Social Markers in Speech, pp. 109–146. Cambridge University Press, Cambridge (1979)
21. Eckert, P.: The Whole Woman: Sex and Gender Differences in Variation. Language Variation and Change 1, 245–267 (1989)
22. Linville, S.E.: Vocal Aging. Singular, San Diego (2001)
23. Braun, A.: Age Estimation by Different Listener Groups. Forensic Linguistics 3, 65–73 (1996)
24. Russ, C.V.J. (eds.): The Dialects of Modern German: A Linguistic Survey. Routledge, London (1990)
25. Köster, O.: Die Datenbank regionaler Umgangssprachen (DRUGS). Ein neues Expertensystem für die forensische Sprechererkennung. Kriminalistik 55, 46–50 (2001)
26. Mangold, M.: Das Aussprachewörterbuch, 5th edn. Dudenverlag, Mannheim (2003)
27. Kohler, K.J.: Segmental Reduction in Connected Speech in German: Phonological Facts and Phonetic Explanations. In: Hardcastle, W.J., Marchal, A. (eds.) Speech Production and Speech Modelling. Foris, Dordrecht, pp. 69–92 (1990)
28. Dirim, I., Auer, P.: Türkisch sprechen nicht nur die Türken. Über die Unschärfebeziehung zwischen Sprache und Ethnie in Deutschland. De Gruyter, Berlin (2004)
29. Foulkes, P.: Sociophonetics. In: Brown, K. (ed.) Encyclopedia of Language and Linguistics, 2nd edn., pp. 495–498. Elsevier, Amsterdam (2006)
30. Hoffmann, L.: Fachsprache/Language of Specific Purposes. In: Ammon, U., Dittmar, N., Mattheier, K.J., Trudgill, P. (eds.) Sociolinguistics/Soziolinguistik, 2nd edn., vol. 1, pp. 232–238. De Gruyter, Berlin (2004)

31. Wendler, J., Seidner, W., Kittel, G., Eysholdt, U.: Lehrbuch der Phoniatrie und Pädaudiologie. Thieme, Stuttgart (1996)
32. Smith, A.: Stuttering: Physiological Correlates and Theoretical Perspectives. In: Blanken, G., Dittmann, J., Grimm, H., Marshall, J.C., Wallesch, C.-W (eds): Linguistic Disorders and Pathologies. De Gruyter, Berlin, pp. 864–870 (1993)
33. Peterson, G.E., Barney, H.L.: Control Methods Used in a Study of the Vowels. Journal of the Acoustical Society of America 24, 175–184 (1952)
34. Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic Characteristics of American English Vowels. Journal of the Acoustical Society of America 97, 3099–3111 (1995)
35. Fitch, W.T., Giedd, J.: Morphology and Development of the Human Vocal Tract: A Study Using Magnetic Resonance Imaging. Journal of the Acoustical Society of America 106, 1511–1522 (1999)
36. Diehl, R.L., Lindblom, B., Hoemeke, K.A., Fahey, R.P.: On Explaining Certain Male-Female Differences in the Phonetic Realization of Vowel Categories. Journal of Phonetics 24, 187–208 (1996)
37. Simpson, A.: Dynamic Consequences of Differences in Male and Female Vocal Tract Dimensions. Journal of the Acoustical Society of America 109, 2153–2164 (2001)
38. Simpson, A.: Gender-Specific Articulatory-Acoustic Relations in Vowel Sequences. Journal of Phonetics 30, 417–435 (2002)
39. Coleman, R.O.: Male and Female Voice Quality and its Relationship to Vowel Formant Frequencies. Journal of Speech and Hearing Research 14, 565–577 (1971)
40. Coleman, R.O.: A Comparison of the Contributions of Two Voice Quality Characteristics to the Perception of Maleness and Femaleness in the Voice. Journal of Speech and Hearing Research 19, 168–180 (1976)
41. Lehiste, I., Meltzer, D.: Vowel and Speaker Identification in Natural and Synthetic Speech. Language and Speech 16, 356–364 (1973)
42. Jakobson, R., Fant, C.G.M., Halle, M.: Preliminaries to Speech Analysis. In: The Distinctive Features and their Correlates, MIT Press, Cambridge, Mass. (1952)
43. Ohala, J.J.: Around Flat. In: Fromkin, V.A. (ed.) Phonetic Linguistics: Essays in Honor of Peter Ladefoged, pp. 223–241. Academic Press, Orlando (1985)
44. Jakobson, R., Waugh, L.R.: The Sound Shape of Language. 2nd edn. Mouton de Gruyter, Berlin (1987)
45. LaPolla, R.J.: An Experimental Investigation into Phonetic Symbolism as it Relates to Mandarin Chinese. In: Hinton, L., Nichols, J., Ohala, J.J. (eds.) Sound Symbolism, pp. 130–147. Cambridge University Press, Cambridge (1994)
46. Laver, J.: The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge (1980)
47. Ball, M.J., Esling, J., Dickson, C.: The Transcription of Voice Quality. In: Kent, R.D., Ball, M.J. (eds.) Voice Quality Measurement. Singular, San Diego, pp. 49–72 (2000)
48. Nolan, F.: Forensic Speaker Identification and the Phonetic Description of Voice Quality. In: Hardcastle, W.J., Mackenzie Beck, J. (eds.) A Figure of Speech. A Festschrift for John Laver, pp. 385–411. Lawrence Erlbaum Associates, Mahwah (2005)
49. Moosmüller, S.: The Influence of Creaky Voice on Formant Frequency Changes. Forensic Linguistics 8, 100–112 (2001)
50. Klatt, D.H., Klatt, L.C.: Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. Journal of the Acoustical Society of America 87, 820–857 (1990)

51. Miller, J.L.: Interactions in Processing Segmental and Suprasegmental Features of Speech. Perception and Psychophysics 24, 175–180 (1978)
52. Repp, B.H., Lin, H.-B.: Integration of Segmental and Tonal Information in Speech Perception: a Cross-Linguistic Study. Journal of Phonetics 18, 481–495 (1990)
53. Mullenix, J.W., Johnson, K.A., Topcu-Durgun, M., Farnsworth, L.M.: The Perceptual Representation of Voice Gender. Journal of the Acoustical Society of America 98, 3080–3095 (1995)
54. Kingston, J., Macmillan, N.A., Dickey, L.W., Thorburn, R., Bartels, C.: Integrality in the Perception of Tongue Root Position and Voice Quality in Vowels. Journal of the Acoustical Society of America 101, 1696–1705 (1997)
55. Ohala, J.J.: An Ethological Perspective on Common Cross-Language Utilization of F0 of Voice. Phonetica 41, 1–16 (1984)
56. Ohala, J.J.: The Frequency Code Underlies the Sound-Symbolic Use of Voice Pitch. In: Hinton, L., Nichols, J., Ohala, J.J. (eds.) Sound Symbolism, pp. 325–347. Cambridge University Press, Cambridge (1994)
57. Hanson, H.M., Chuang, E.S.: Glottal Characteristics of Male Speakers: Acoustic Correlates and Comparison with Female Data. Journal of the Acoustical Society of America 106, 1064–1077 (1999)
58. Kreiman, J., Gerratt, B.: Measuring Voice Quality. In: Kent, R.D., Ball, M.J. (eds.) Voice Quality Measurement. Singular, San Diego, pp. 73–101 (2000)
59. Köster, O., Köster, J.-P.: The Auditory-Perceptual Evaluation of Voice Quality in Forensic Speaker Recognition. The Phonetician 89, 9–37 (2004)
60. Baken, R.J., Orlikoff, R.F.: Clinical Measurement of Speech and Voice. Singular, San Diego (2000)
61. Wagner, I.: A New Jitter-Algorithm to Quantify Hoarseness: an Exploratory Study. Forensic Linguistics 2, 18–27 (1995)
62. Lass, N.J., Phillips, J.K., Bruchey, C.A.: The Effect of Filtered Speech on Speaker Height and Weight Identification. Journal of Phonetics 8, 91–100 (1980)
63. Van Dommelen, W.: Speaker Height and Weight Identification: a Re-Evaluation of Some Old Data. Journal of Phonetics 21, 337–341 (1993)
64. Van Dommelen, W., Moxness, B.H.: Acoustic Parameters in Speaker Height and Weight Identification: Sex-Specific Behaviour. Language and Speech 38, 267–287 (1995)
65. Fitch, W.T.: Vocal Tract Length Perception and the Evolution of Language. Ph.D. Dissertation, Brown University (1994)
66. Collins, S.A.: Men's Voices and Women's Choices. Animal Behaviour 60, 773–780 (2000)
67. Greisbach, R.: Estimation of Speaker Height From Formant Frequencies. Forensic Linguistics 6, 265–277 (1999)
68. González, J.: Formant Frequencies and Body Size of Speaker: a Weak Relationship in Adult Humans. Journal of Phonetics 32, 277–287 (2004)
69. Rendall, D., Kollias, S., Ney, C.: Pitch (F0) and Formant Profiles of Human Vowels and Vowel-Like Baboon Grunts: The Role of Vocalizer Body Size and Voice-Acoustic Allometry. Journal of the Acoustical Society of America 117, 944–955 (2005)

Forensic Automatic Speaker Classification in the "Coming Paradigm Shift"

Joaquin Gonzalez-Rodriguez and Daniel Ramos

ATVS (Speech and Signal Processing Group), Escuela Politecnica Superior
Universidad Autonoma de Madrid, E-28049 Madrid, Spain
{joaquin.gonzalez,daniel.ramos}@uam.es

Abstract. A new paradigm for forensic science has been encouraged in the last years, motivated by the recently reopened debate about the infallibility of some classical forensic disciplines and the controversy about the admissibility of evidence in courts. Standardization of procedures, proficiency testing, transparency in the scientific evaluation of the evidence and testability of the system and protocols are emphasized in order to guarantee the scientific objectivity of the procedures. In this chapter those ideas and their relationship to automatic forensic speaker classification will be analyzed in order to define where automatic speaker classification is and which direction should it take under this context. Following the DNA methodology, which is being regarded as the scientific "golden" standard for evidence evaluation, the Bayesian approach has been proposed as a scientific and logical methodology. Likelihood ratios (LR) are computed based on the similarity-typicality pair, which facilitates the transparency in the process. The speaker classification is performed by the fact finder, who defines the possible hypotheses involved in the classification process. Thus, the prior probability of the hypotheses and the LR computed by the forensic system are used to assign a class to each suspected speaker depending on the defined hypotheses. The definition of this hypotheses typically refer to the speaker identity, thus leading to a speaker recognition task, but they can be defined in a more general context of speaker classification. The concept of calibration as a way of reporting reliable and accurate opinions is also addressed. Application-independent evaluation techniques (C_{llr} and APE curves) are addressed as a proper way for presenting results of proficiency testing in courts, as these evaluation metrics clearly show the influence of calibration errors in the accuracy of the inferential decision process. In order to illustrate the effects of calibration, we conclude with new experimental examples used as blind proficiency test following the NIST SRE 2006 evaluation protocol.

Keywords: forensic, calibration, likelihood ratio, paradigm shift, Daubert rules, speaker classification, DNA.

1 Introduction

In the recent years, the interest in the classical debate about the presentation of the forensic evidence in a court of law has significantly increased [1,2,3]. Some

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 205–217, 2007.

reasons for that are found in the establishment of the American Daubert rules for the admissibility of the scientific evidence in trials [4]. These rules claim that scientific techniques presenting standard procedures and demonstrating their testability, accuracy and acceptance in the scientific community are likely to be accepted in a U.S. federal court of law. On the other hand, non-scientific statements, such as expert testimonies lacking of scientific foundations, are likely to be rejected. The implications of these rules are in accordance to many opinions of forensic experts worldwide [1,5,2,3]. The debate also considers that existing techniques which have been assumed by the court as error-free are starting to be questioned. This has been partly due to some critical errors in positive identification reports, highlighted by the mass media (like the Mayfield case in Madrid terrorist attacks in 11 March 2004 [6]).

In order to cope with this emerging requirements, the speaker recognition community has been investigating ways of converging to this new paradigm in forensic science in the last years [7,8]. Standardization and proficiency testing should be key points in this convergence process, as a way of presenting the accuracy of the systems in a clear and standard way. In order to converge in the evidence evaluation process, the Bayesian approach for evidence analysis [9] has been proposed as a common framework for forensic interpretation of the evidence, following the DNA standard. This approach has been successfully applied to forensic speaker classification, using automatic [10,11], phonetic-acoustic [7,12] or semi-automatic approaches [7]. In this Bayesian framework, speaker classification is performed by the fact finder, who also defines the hypotheses in the case. Then, classification is achieved by considering the prior probabilities of the hypotheses and the LR computed by the forensic scientist or system. In a forensic case the fact finder will be typically interested in the identity of the speaker, leading to a speaker recognition task. However, the hypotheses may be defined as more general classes of speakers, and therefore it will be speaker classification.

One of the main advantages of Bayesian methods is their testability. Opinions about the hypotheses are expressed in the form of posterior probabilities. Therefore, there is a need of measuring not only the discrimination capabilities of the system, but the reliability of such confidences. Highly discriminant (or *refined* [13]) systems may lead to wrong posterior probabilities if they do not elicit reliable (or *calibrated*) confidences [13,14,15].

In this chapter, we define a framework for the use of automatic speaker classification following the criteria needed by the coming paradigm shift in forensic science [1]. The chapter is organized as follows. Section 2 describes the motivation and main differential characteristics of the “coming paradigm shift” [1], including the use of the Bayesian framework for evidence analysis following the DNA standard. The concepts of *calibration* and *refinement* are introduced in Section 3 and a methodology for the assessment of miscalibration effects is presented. An experimental example is shown in Section 4, where the effect of calibration is highlighted by simulating a comparative proficiency testing of several robust approaches proposed in the literature for Bayesian forensic speaker classification for the two-class problem. Finally, conclusions are drawn in Section 5.

2 The Coming Paradigm Shift and Forensic Speaker Recognition

In 1993, the United States Supreme Court stated that [4] in order for scientific evidence to be accepted in a U.S. federal court of law, any technique must satisfy the following conditions: *i)* it has been or can be tested. *ii)* it has been subjected to peer review or publication, *iii)* there exist standards controlling its use, *iv)* it is generally accepted in the scientific community, and *v)* it has a known or potential (and acceptable) error rate. These so-called *Daubert rules*, added to the evidence of errors in some well-established forensic areas, have lead to reconsider the procedures used for forensic interpretation and reporting [1,2]. Transparent and standard methodologies and proficiency testing are being highlighted as essential for a proper use of scientific evidence. Moreover, it has been pointed out that no forensic discipline is really error-free, even considering some well established disciplines which were viewed as error-free in the past (e. g., fingerprints [16]). These demonstrations have come maily due to important mistakes in real trials [16,6], but there also is an important part of the scientific community who supports those ideas [3,2]. In this sense, positive identification as a result of forensic analysis constitutes an arbitrary decision adopted by the experts in a subjective way, usually justified by their experience in the field [2,3]. This obscurity and arbitrariness in positive identification statements leads not only to usurp the judge's role in the decision making process [17], but also to a hardly testable framework.

In [1], DNA analysis is proposed as a model in order to avoid these difficulties. The main characteristics of forensic DNA analysis, highlighted in [1,2] may be summarized in: *i)* it is scientifically based, avoiding expert opinions based on experience [2]; *ii)* it is clear and standard in their procedures, allowing scrutinizing and inspection by fact finders and forensic scientists [1]; and *iii)* it is probabilistic, avoiding hard *match* or *non-match* statements [1,3]. This forensic discipline, much newer than fingerprint analysis, has been characterized by the use of a Bayesian methodology, which has been addressed as a logical and scientific framework for evidence analysis [9,5]. Under this framework, DNA experts have computed a degree of support for the prosecutor or defense hypothesis in the form of a *likelihood ratio*. This value is obtained in a data-driven way by computing: *i)* a similarity factor which supports that the questioned sample was left by a given suspect, and *ii)* a typicality factor which supports that the questioned sample was left by anyone else in a relevant population. In order to adopt this methodology for forensic speaker classification, during the last years several works have demonstrated that any score-based speaker classification system can be adapted to work following the Bayesian methodology [18,10,11].

2.1 The Bayesian Methodology

The Bayesian framework for interpretation of the evidence represents a mathematical an logical tool for the evidence analysis process. This Bayesian framework presents many advantages in the forensic context. First, it allows the forensic scientists to estimate and report a meaningful value to the court [17]. Second, the role of the scientist is clearly defined, leaving to the court the task of using prior judgements or costs in the decision process [19]. Third, probabilities can be interpreted as degrees of belief, allowing the incorporation of subjective opinions as probabilities in the inference process in a clear and scientific way [20].

Classically, Bayesian interpretation of the forensic evidence using automatic systems has been performed by generative statistical models [21,9,18,11], whereas discriminative techniques have been also recently applied to this task [10]. In both cases, the objective is to compute the likelihood ratio (LR) as a degree of support of one hypothesis versus its opposite. This LR can be estimated from similarity scores computed by an automatic system [10,11]. We assume that the evidence E is the comparison of a questioned mark recovered from a scene of crime (e. g., a wire-tapping) with some material from a known source, which can be a suspect (e. g., a recording from the suspect in controlled situations). Typically, using automatic systems this E will be a similarity score between the mark and the suspect material. However, other kind of meta-information (such as signal to noise ratio, transmission channels, subjective quality of the speech signal, etc.) may be also used in order to compute this LR value [10]. Bayes' theorem states that:

$$\frac{P\left(H_p \mid E, I\right)}{P\left(H_d \mid E, I\right)} = LR \cdot \frac{P\left(H_p \mid I\right)}{P\left(H_d \mid I\right)} \tag{1}$$

$$LR = \frac{f\left(E \middle| H_p, I\right)}{f\left(E \middle| H_d, I\right)} \tag{2}$$

where H_p (a given suspect is the author of the questioned recording involved in the crime) and H_d (another individual is the author of the questioned recording involved in the crime) are typically the relevant hypothesis and I is the background information available in the case. The hypotheses are defined in the court from I, the prosecutor and defense propositions and often because of the adversarial nature of the criminal system. We will use the hypotheses defined above through all the chapter, as the typical definition at trial will be realted to the identity of the speaker, and thus we can talk about sepaker recognition. However, they can be defined in a wider context, leading to a more general speaker classification task. For instance, the judge may be interested in knowing if the speaker was a smoker or not.

As it can be seen in Equation 2, the LR is the ratio of two magnitudes. The likelihood $f\left(e \middle| H_p, I\right)$ in the numerator in Equation 2 is known as the within-source distribution, and models the variability of the speaker between sessions. The evaluation of this function in $e = E$ gives a measure of the similarity between the questioned material and the suspect. On the other hand, the likelihood $f\left(e \middle| H_d, I\right)$ in the denominator is known as the between-source distribution, and its evaluation in $e = E$ can be seen as a measure of the typicality or rarity of the suspect in a relevant population of individuals. Both values, similarity and typicality, are computed in a transparent way by the speaker recognition system or expert, and it is the duty of the forensic scientist, following the background information of the case (I), to select the population of individuals which will be proper for the case at hand. This approach for LR computation can be easily documented by the forensic scientist and understood by fact finders [9,2].

2.2 Proficiency Testing in Automatic Forensic Speaker Classification

As it has been mentioned above, proficiency testing is addressed as a key issue for the admissibility of forensic systems in courts [1]. According to Daubert, in order to improve the clarity in the presentation of the performance of the technique in use, we will need unified protocols for system evaluation. We identify two main factors as critical for the achievement of this goal in forensic speaker recognition and classification. First, there should be standard and accessible speech databases and protocols in order to perform the test in comparable conditions. In this sense, the work by NIST and NFI/TNO in their respective SREs has been fundamental in the last years [22], and such databases and protocols are a reference for any scientific evaluation of performance of speaker recognition systems. Second, the use of a common methodology for presenting results in court will measure and clarify the reliability of the system to be used for forensic analysis. In Section 3.1 we propose a method for the common evaluation of forensic speaker classification systems.

3 Calibration in Bayesian Forensic Speaker Classification

The concept of calibration was introduced in [13] in the context of weather forecasting. There, posterior probabilities (also known as confidences) were used as degrees of belief about a given hypothesis (tomorrow it will rain) against its opposite (tomorrow it will not rain). The accuracy of the forecaster was then assessed by means of *strictly proper scoring rules*, which may be viewed as cost functions which assign a penalty to a given confidence depending on: *i)* the probabilistic value of the forecast, and *ii)* the true hypothesis which actually occurred (see [14,15] for details). For example, if a probabilistic forecast gives a high probability of rain for tomorrow (value of the forecast) and tomorrow it does not rain (true hypothesis), a proper scoring rule will assign a high penalty to the forecast, and vice-versa. Strictly proper scoring rules have interesting properties. First, if we know the true value of the hypotheses for each trial and use that knowledge for building a *perfect system*, the *only* posterior probability value which optimize a strictly proper scoring rule with respect to that perfect system is the posterior probability which would be given by the perfect system itself [13]. Thus, any opinion expressed by the forecaster which deviates from the one elicited with knowledge of the true answer will lead to a higher penalty. Second, in [13] it is demonstrated that any proper scoring rule can be split into a *refinement* component, measuring the discrimination capabilities of the confidence values elicited, and a *calibration* component, which measures the deviation of such confidence values from those elicited by the perfect system.

The use of proper scoring rules in order to assess speaker classification and recognition systems delivering LR values has been recently proposed in the literature [14,10,15]. In a speaker classification context, each *forecast* is represented with the confidence on the hypothesis "the speaker is the author of the test utterance" or its opposite or opposites, which may be inferred from the LR computed by the speaker recognition system and the prior probabilities (not necessarily estimated by the system). This assessment framework is perfectly suited for the methodology proposed in Section 2.1 for forensic speaker recognition considering: *i)* the hypotheses used are H_p and H_d as defined in Section 2.1, *ii)* the prior judgements are province of the court, and *iii)* the LR is computed by the forensic speaker recognition system.

3.1 Assessing Calibration in Forensic Speaker Classification

The common assessment methods widely used among the speaker recognition community for assessing the performance of systems have been mainly proposed in NIST SREs. There, DET plots have been used to measure the discrimination performance of speaker detection technology. However, LR values are not only used as a discrimination score, but as a measure of the degree of support to a hypothesis against its opposite. Using the LR and the prior odds (province of the court [17]) we obtain a posterior probability for each hypothesis. Thus, the accuracy of the LR values does not only depend on their discrimination power for trials where H_p or H_d is true (measured by the refinement of the LR values), but in their actual values (calibrated LR values will lead to reliable confidences).

In order to assess the actual values of the LR, in Bayesian analysis of forensic evidences Tippett plots have been classically used for performance evaluation [11]. In this representation, the distribution of the LR values being H_p or H_d respectively true are plotted together. Important values shown by these curves (and not by DET plots) are the distributions of the computed LR values and the rates of misleading evidence. The rate of misleading evidence is defined as the proportion of LR values giving support to the wrong hypotheses ($LR > 1$ when H_d is true and $LR < 1$ when H_p is true). In Figure 1 an example of Tippett plots is shown.

Recent approaches for speaker classification evaluation have proposed the use of application-independent metrics such as C_{llr} [14], where *application*, as defined in [14,15], is the set of prior probabilities and decision costs involved in the inferential process [19]. C_{llr} is a single scalar value defined as:

$$C_{llr} = \frac{1}{N_{H_p}} \sum_{i for H_p = true} log_2\left(1 + \frac{1}{LR_i}\right) + \frac{1}{N_{H_d}} \sum_{j for H_d = true} log_2(1 + LR_j) \tag{3}$$

where N_{H_p} and N_{H_d} are respectively the number of LR values in the evaluation set for H_p or H_d true. As it can be seen in Equation 3, hypothesis-dependent

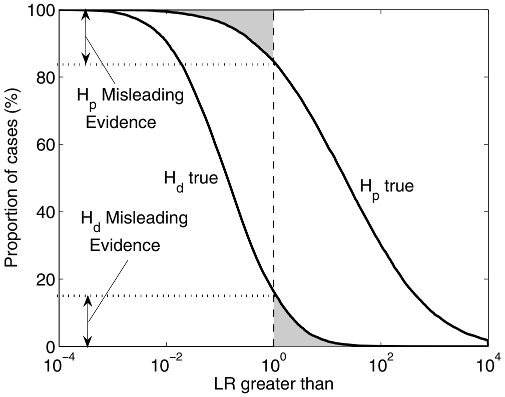
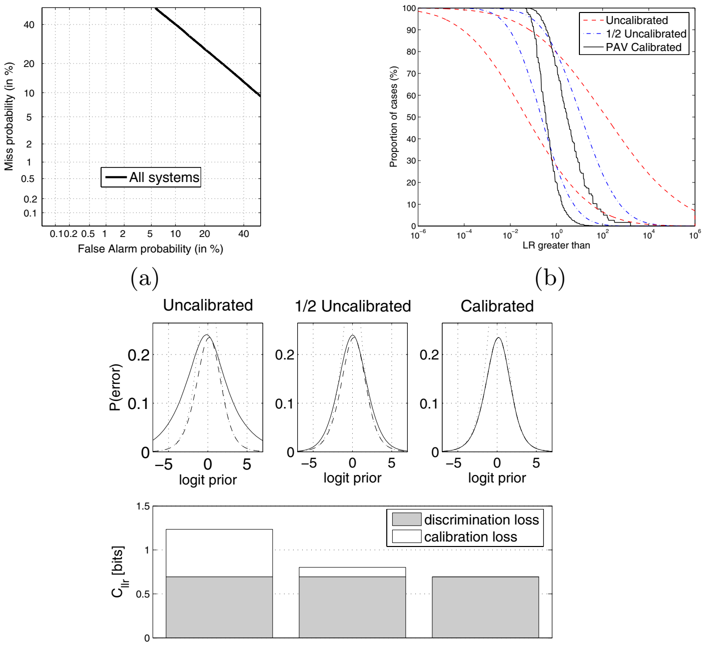


Fig. 1. Example of Tippett plots showing the actual LR distributions (with its meaningful values) and the rates of misleading evidence when H_p and H_d are respectively true

logarithmic cost functions are applied to the LR values being evaluated, and thus they are assessed depending on their numerical value: highly misleading LR values will have a strong penalty (high C_{llr}) and viceversa.

C_{llr} presents several interesting properties. First, the LR values are evaluated in an application-independent way, which in forensics means *case-independent*, where different costs and priors may be involved in the decision process of each different case [19]. Second, as a single scalar value, C_{llr} is very useful in order to easily compare and rank systems. Third, it can be demonstrated that C_{llr} is a strictly proper scoring rule [14], and it can be split into discrimination loss (C_{llr}^{min}) and calibration loss ($C_{llr} - C_{llr}^{min}$). The C_{llr}^{min} value is obtained by optimal calibration via a monotonic transformation of the LR values, knowing the actual hypothesis occurred for each LR value. The Pool Adjacent Violators (PAV) algorithm is used for obtaining such monotinic transformation. Details may be found in [14,15].

Related to this C_{llr} value, the APE-curve (Applied Probability of Error) [14] has been also proposed as a way of measuring the probability of error of the LR values computed by the forensic system in a wide range of applications (different costs and priors). This probability of error is represented for the actual LR values computed by the speaker recognition system and also for optimally calibrated LR values obtained as cited above for C_{llr}^{min}. Therefore, this representation clearly illustrates the effects of a lack of calibration: highly discriminant LR values may lead to a high probability of erroneous decisions if they are not properly calibrated. Because of their interesting properties, APE curves and C_{llr} has been used as an evaluation metric in NIST 2006 SRE [23]. In this article, we have used the evaluation tools for C_{llr} and APE curve computation included in the toolkit FoCal [24].



(a) (b)

(c)

Fig. 2. DET curves (a), Tippett plots (b) and APE curves (c) for three simulated systems (System 1, System 2 and System 3). *LR* values have been randomly generated in order to plot these curves.

3.2 Calibration Example

The effects of calibration in forensic speaker classification are illustrated in this section with an example using synthetic data. Here, an uncalibrated set of LR values have been synthetically generated for each of the H_p and H_d hypotheses. This system is then transformed by two different monotonic mappings: i) a linear scaling by a 0.5 factor and ii) the PAV algorithm trained with the same syntethic data (a-posteriori training). Figure 2 shows the performance of these *synthetic systems*, which we will call *Uncalibrated* system, 1/2 *Uncalibrated* system and *Calibrated* system. Results are presented in terms of DET curves, Tippett plots, C_{llr} values and APE curves. DET curve in Figure 2(a) shows that the discrimination performance of all systems is the same. This is due because a monotonic transformation does not change the ordering of the scores. However, Tippett plots in Figure 2(b) show that confidences inferred from LR values computed by the *Uncalibrated* system will lead to important errors because of the high proportion of misleading LR values. However, the rates of misleading evidence are dramatically reduced for system 1/2 *Uncalibrated* because of the linear scaling. Thus, LR values will be more moderate after the scaling, but less misleading.

The same effects in LR values are observed for system *Calibrated*, but in this case the mapping is non-linear.

These results are clearly observed in Figure 2(c), which presents the same results in the form of C_{llr} values and APE curves. Overall performance is given by C_{llr}, split into discrimination loss (C_{llr}^{min}) and calibration loss ($C_{llr} - C_{llr}^{min}$). It is observed that all systems present the same discrimination performance (discrimination loss). However, C_{llr} values for the *Uncalibrated* system are much higher than for the rest, mainly because the effects of misleading evidence observed in Figure 2(b). On the other hand, the calibration performance of the *Calibrated* LR values is the best for all systems.

In order to complete the analysis, APE curves in Figure 2(c) show the probability of error for all possible values of prior probabilities and decision costs (horizontal axis). The dashed line shows the performance of optimally calibrated LR values obtained by monotonic transformation from LR values given by the system [14,15]. The solid line shows the actual probability of error of the LR values computed. It is observed that the probability of error dramatically increases when the system is not properly calibrated. Due to this lack of calibration, posteriors inferred using the *Uncalibrated* LR values will have a much higher probability of error, even when it has the same discrimination performance as the rest of systems.

4 Experiments

In order to confirm the effects presented in Section 3 using automatic speaker classification systems, we present an experimental example using different LR computation techniques.

The scores needed for LR computation have been obtained using the ATVS GMM-SVM-NAP system, which is based on the classification of GMM mean-supervectors using support vector machines. Details may be found in [25]. The comparative results presented here consider two techniques for the evaluation of the forensic evidence found in the literature, namely: *i)* suspect-independent LR computation [26] and *ii)* suspect-adapted Maximum A Posteriori (MAP) LR computation. We briefly describe each interpretation technique below.

In suspect-independent within-source estimation a framework is proposed assuming that an accurate model of the within-source distribution for a given suspect can be obtained using target scores from different individuals in the same conditions, namely $X_G = \{x_{G1}, \ldots, x_{GN}\}$. On the other hand, suspect-adapted MAP estimation of within-source distributions adapts the global distribution $f_G(e) = N(\mu_G, \sigma_G)$ to the *suspect* distribution $f_S(e) = N(\mu_S, \sigma_S)$, estimated from a set of M suspect target scores $X_S = \{x_{S1}, \ldots, x_{SM}\}$ obtained from the suspect speech involved in the trial. Therefore, an *adapted* within-source pdf $f(e|H_p, I) \equiv f_A(e) = N(\mu_A, \sigma_A)$ is obtained. See [27] for details.

Experiments have been performed using the evaluation protocol proposed in NIST 2006 SRE for the 1 conversation side training and 1 conversation side testing task (1c-1c, see [23] for details). Suspect target scores set X_S consists of

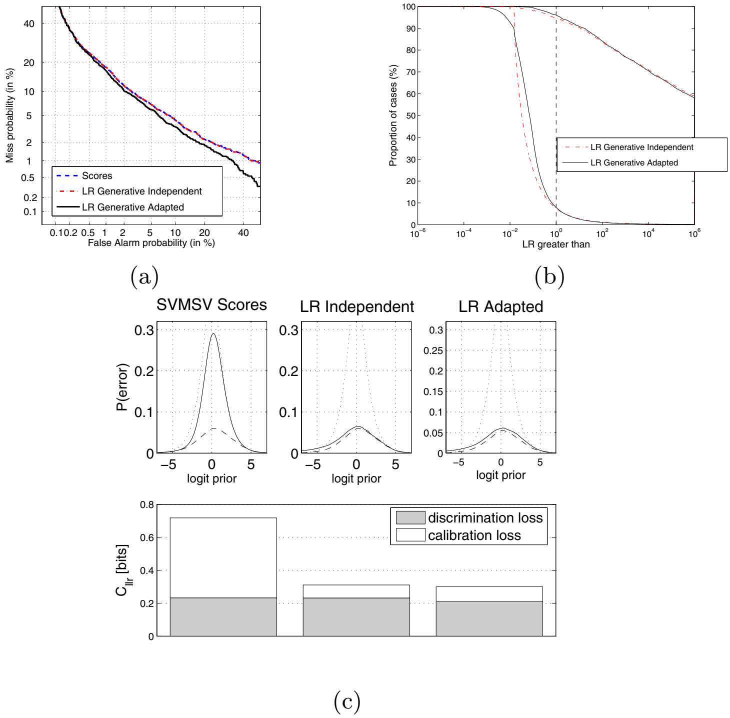


Fig. 3. DET curves (a), Tippett plots (b) and APE curves (c) comparing suspect-independent, WDP and suspect-adapted within-source computation with scarce suspect data (M=2) in the selected subset from 8c-1c in NIST 2005 SRE

all the target scores for each speaker from the whole score set in the evaluation, except the score used as evidence in each LR computation. More than 50.000 trials have been performed in this condition. Background data, including global target score set X_G, have been extracted form NIST 2004 SRE database and protocol [22].

4.1 Results

In Figure 3 we compare the performance of the different evaluated techniques. Results are presented in DET curves, Tippett plots, C_{llr} values and APE curves. In Figure 3(a) we can see the DET plots showing the discrimination performance of the different evaluated LR computation techniques using the GMM-SVM-NAP speaker recognition system. It can be observed that the discrimination performance is better for the suspect-adapted case. Thus, suspect-adapted LR computation exploits suspect-specificities in order to lead to a more efficient extraction of information about the speaker identity.

Figure 3(b) shows the distribution of LR values in the form of Tippett plots for each LR computation technique. It is shown that the magnitude of the strength of misleading evidence (LR values supporting the wrong hypothesis) is quite limited for both cases. This is a good result, because the LR values obtained using these two different LR computation techniques are homogeneous. However, the observation of the Tippett plots does not allow us to easily conclude which technique is better. Moreover, the calibration of the techniques is not explicitly measured.

Figure 3(c) shows the proposed methodology for presenting forensic system testing results. It is observed that the discrimination performance (dashed curve) is better for the suspect-adapted case, as it was shown in 3(a). The calibration performance is very similar for both LR computation techniques, significantly outperforming the scores. On the other hand, C_{llr} values, represented under the APE curves, give an overall performance metric which allows the ranking of the different techniques. In this case, suspect-adapted technique outperforms suspect-indpendent LR computation.

5 Conclusions

In this chapter we have discussed the current state and future directions that forensic speaker classification should take in order to cope with the rising needs being debated in the forensic science community. Questioning the infallibility of any forensic technique and demanding scientifically-sound methods for the admissibility of forensic evidence in the court are the main reasons for these new requirements. Some main guidelines for the use of forensic speaker recognition in courts may be drawn from this debate, such as the need of transparency, accuracy and testability for any technique to be admissible. This work has presented a methodology which copes with these interrelated requirements and therefore fulfills these needs. The transparency of the reasoning process under uncertainty is guaranteed by the use of the scientific and logical Bayesian framework for evidence analysis, as it happens in forensic DNA profiling. Under such Bayesian framework, the roles of the fact finder and the forensic system are perfectly defined. It is the role of the forensic system to output a LR value about the hypotheses involved in the case. On the other hand, the fact finder defines the hypotheses, gives value to the prior odds and performs the final speaker classification step. The hypotheses are typically related to the identity of the speaker (speaker recognition) but they may be defined in a wider context, leading to a more general speaker classification task. The discussion about the effects of a lack of calibration in automatic forensic speaker classification systems has been supported by heuristic examples and experimental results. The conclusions from such discussion can be extended to any other forensic speaker classification approaches (semi-automatic, phonetic-acoustic, etc.) based on the Bayesian framework and reporting LR values. Moreover, several methods for the evaluation of forensic systems have been addressed, from classical techniques based on DET curves and Tippett plots to more recent application-independent approaches based on C_{llr} and APE curves. These two last metrics have been emphasized as a proper way of presenting results, as they show and highlight the calibration performance as a measure of reliability of the LR values computed by the forensic system. All these evaluation techniques, added to a clear and standard protocol such as those developed by NIST in their yearly SREs, give a method to perform proficiency tests in a controlled and transparent way. Therefore, the proposed methodology looks forward to fulfilling the needs of testability and standardization stated by the Daubert rules and demanded from forensic experts worldwide.

Acknowledgements

This work has been supported by the Spanish Ministry for Science and Technology under project TEC2006-13170-C02-01. The authors also wish to thank Niko Brümmer, Colin Aitken and Christophe Champod for fruitful discussions and useful comments.

References

1. Saks, M.J., Koehler, J.J.: The coming paradigm shift in forensic identification science. Science 309, 892–895 (2005)
2. Cole, S.A.: A history of fingerprinting and criminal identification (2005), Available at http://www.nasonline.org/site/PageServer?pagename=sackler_forensic_presentations
3. Champod, C., Evett, I.W.: A probabilistic approach to fingerprint evidence. Journal of Forensic Identification 51, 101–122 (2001)
4. Court, U.S.: Daubert v. Merrel Dow Pharmaceuticals [509 U.S. 579] (2003)
5. Evett, I.W.: Towards a uniform framework for reporting opinions in forensic science casework. Science and Justice 38, 198–202 (1998)
6. Heath, D., Bemton, H.: Portland lawyer released in probe of Spanish bombings. Seattle Times (May 21, 2004), Available at http://www.law.asu.edu/?id=8857
7. Rose, P.: Technical forensic speaker recognition: Evaluation, types and testing of evidence. Computer Speech and Language 20, 159–191 (2006)
8. Ramos-Castro, D., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: Likelihood ratio calibration in transparent and testable forensic speaker recognition. In: Proc. of Odyssey (2006)
9. Aitken, C.G.G., Taroni, F.: Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley & Sons, Chichester (2004)
10. Campbell, W.M., Reynolds, D.A., Campbell, J.P., Brady, K.J.: Estimating and evaluating confidence for forensic speaker recognition. In: Proc. of ICASSP, pp. 717–720 (2005)
11. Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J.: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech and Language 20, 331–355 (2006)
12. Jessen, M.: Speaker classification in forensic phonetics and acoustics. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNCS), vol. 4343, Springer, Heidelberg (2007) (this issue)
13. de Groot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. The Statistician 32, 12–22 (1982)

14. Brummer, N., du Preez, J.: Application independent evaluation of speaker detection. Computer Speech and Language 20, 230–275 (2006)
15. van Leeuwen, D., Brümmer, N.: An introduction to application-independent evaluation of speaker recognition systems (this issue). In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007)
16. Cole, S.A.: More than zero: Accounting for error in latent fingerprint identification. Journal of Criminal Law & Criminology 95, 985–1078 (2005)
17. Champod, C., Meuwly, D.: The inference of identity in forensic speaker recognition. Speech Communication 31, 193–203 (2000)
18. Meuwly, D.: Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique. Ph.D. thesis, IPSC-Universite de Lausanne (2001)
19. Taroni, F., Bozza, S., Aitken, C.G.G.: Decision analysis in forensic science. Journal of Forensic Sciences 50, 894–905 (2005)
20. Taroni, F., Aitken, C.G.G., Garbolino, P.: De Finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: A commentary for forensic scientists. Science and Justice 41, 145–150 (2001)
21. Curran, J.: Forensic Applications of Bayesian Inference to Glass Evidence. University of Waikato, New Zealand (1997)
22. van Leeuwen, D., Martin, A., Przybocki, M., Bouten, J.: The NIST 2004 and TNO/NFI speaker recognition evaluations. Computer Speech and Language 20, 128–158 (2006)
23. NIST: NIST speech group website: http://www.nist.gov/speech
24. Brummer, N.: (Focal toolkit) Available at http://www.dsp.sun.ac.za/~nbrummer/focal/
25. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. Signal Processing Letters 13(5), 308–311 (2006)
26. Botti, F., Alexander, A., Drygajlo, A.: An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data. In: Proc. of Odyssey, pp. 63–68 (2004)
27. Ramos-Castro, D., Gonzalez-Rodriguez, J., Montero-Asenjo, A., Ortega-Garcia, J.: Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation. In: Proc. of Odyssey (2006)

The Many Roles of Speaker Classification in Speaker Verification and Identification

Judith Markowitz

J. Markowitz, Consultants
5801 N. Sheridan Rd, #19A
Chicago, IL 60660
judith@jmarkowitz.com

Abstract. Speaker classification is a fundamental component of speaker identification and verification (SIV) technologies. This paper provides and overview of the many guises that classification takes within SIV systems.

Keywords: biometric, speaker identification, speaker authentication, speaker verification, authentication, identification, verification, speaker classification, SIV, anti-speaker, disguised voice, speaker segmentation, speaker clustering, speaker variability.

1 Introduction

One of the most widely-deployed application domains of speaker classification is within systems that perform automated speaker identification and verification (SIV). The purpose of a speaker-verification (SV) system is to determine whether the speaker is making a true or a false claim of identity. The object of speaker identification (SI) is to attach a speaker identity to a sample of speech from a previously unknown speaker. The use of both technologies is growing for security, forensics, and intelligence (Markowitz, 2000 [1], 2006 [2]).

The aim of both SV and SI is to link a speech sample to a specific individual, which is not classification. Yet, SI and SV systems (and other biometric verification and identification systems) perform a number of classification tasks in order to accomplish their goals.

2 Variability

The reason classification is used is that the data in SIV/biometric samples are variable. In fact, spoken utterances are like unique creations produced by similarities and differences arising from both external sources and the speaker. Variability is such an inherent part of SIV and other biometrics that if a sample is found to be a perfect or near-perfect match with the enrollment data from the claimed identity the system sounds a “replay” or “spoofing” attack alarm

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 218–225, 2007.

(Markowitz, 2005 [3]). In replay/spoofing attacks an imposter attempts to fool the biometric security system by re-using a sample taken from the claimed identity. Replay/spoofing in SV generally employs a tape recording (called a "tape attack")[1].

Resolution of variability involves classification of the speakers acoustic patterns as well as classification operations related to the communication environment (noise, device/handset type, and channel). Intra-speaker variability can be produced by speaking at different speeds, by stress, illness, fatigue, whispering; or simply by positioning the articulators (lips, teeth, or tongue) differently.

SIV systems capture and encode some intra-speaker variability during enrollment by asking for several utterances or by having the enrollee talk for up to thirty seconds while the system captures and analyzes the speech. The enrollment data are clustered into a "codebook" that describes the enrollee's voice. This information is stored as the enrollee's voice model (sometimes called "voiceprint"). It is, essentially, a delineation of the class of vocal behaviors of the enrollee.

When a new utterance is submitted to an SIV system by someone claiming to be the enrollee, the system compares the codebook for that utterance with the codebook(s) for one or more stored voice models. This process is often called the "classification" step of SIV. SV, for example, evaluates whether and how well the new sample fits into the class of acoustic patterns defined by the voice model of the person the speaker claims to be.

The most widely-used approaches for accomplishing this classification task are nearest neighbor, vector quantization, neural networks, and binary trees. Each of these techniques calculates the similarities and differences between the new sample and other voice models for each of the features utilized by the system. This process is consolidated into an overall similarity score. SV uses the score determine whether the speaker's claim of identity will be accepted or rejected; SI uses the score to rank speaker candidates for the speech sample under analysis.

Philips Speech Recognition Systems employs a variant of this technique in its speech-recognition (SR) dictation product for physicians. SR dictation systems create a separate user model for each speaker and continually update that model as the person speaks. Philips noticed that physicians often hand dictation off to assistants who use the physicians user model to do their work. If the acoustic patterns of the assistants were incorporated into the model it would degrade accuracy. The classification metric determines whether or not the current speaker is the enrolled physician. If not, it will not update the user model.

SV systems also employ a set of techniques for enhancing the accuracy of the classification called anti-speaker modeling.

[1] A human mimic could also be used to spoof an SV system but this is rare and much trickier. SIV systems employ features that reflect the size and shape of the vocal apparatus (throat, mouth, and nose) In order to mount a viable attack, the mimic must have physiology that is similar to the claimed identity or the system will detect differences.

3 Anti-speaker Modeling

Virtually all commercial and research SV systems employ some form of anti-speaker modeling. Anti-speaker modeling is designed to enhance the accuracy of an SV system by comparing the claimant's speech with voice models from speakers other than the model for the claimed identity. These additional evaluations allow the SV system to perform better in "adverse" environments, such as those with a great deal of background or channel noise, or when there is a mismatch between the handset or channel used for enrollment and that used by the claimant.

One kind of anti-speaker modeling, discriminant training, entails categorization of a newly-enrolled voice model based on comparison with all the other voices in the system. This approach is an inherent part of how neural networks and, to some extent, binary trees operate.

Another widely-used type of anti-speaker modeling is the "world model" (also called "background model"). It is a class model that is derived from the speech of a diverse population of speakers. Well-designed world models contain a balance of voices that would be representative of the voices of potential imposters.

In the world-model approach, the claimant's speech is compared with the voice model of the claimed identity and with the world model. The score is computed as a ratio of the divergence of the claimant's speech from the model of the claimed identity over the divergence of the claimant's speech from the world model (Equation 1).

$$\text{score} = \frac{\text{claimed identity}}{\text{world model}} \tag{1}$$

A high score indicates that the claimant's speech is more akin to the voice of the claimed identity than it is to the world model and that there is a high probability that the claimant is who she/he claims to be. A low score suggests that the claimant is likely an impostor.

From the perspective of speaker classification, the most interesting variant of anti-speaker modeling is cohort normalization (Higgins et al, 1991 [4]). Cohort normalization is performed when an individual enrolls in an SV system. After creating the codebook for the enrollee, the system examines its database for voice models that are similar to the newly-created model. The cohort class differs for each enrollee.

When a claimant supplies speech data to an SV system with cohort normalization the system retrieves the voice model for the claimed identity and the voice model for each of its cohorts. The claimant's speech is compared to all of those models with the expectation that, if the claimant is making a valid claim, the score for the claimed identity model will be higher than the scores for anyone in the cohort class.

The IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence), a research institute in Switzerland, employed a combination of a world model of English speakers, Arabic-speaking cohort models, and numerous examples of Osama bin Laden's speech to determine whether the 2002 recording attributed to bin Laden was faked. Figure 1 shows that what this procedure does is determine whether or not a given sample can be categorized as being within the bin Laden class.

IDIAP [5] concluded that, “While this study does not permit us to draw any definite (statistically significant) conclusions, it nonetheless shows that there is serious room for doubt” about whether the voice on the tape could be categorized as that of Osama bin Laden.

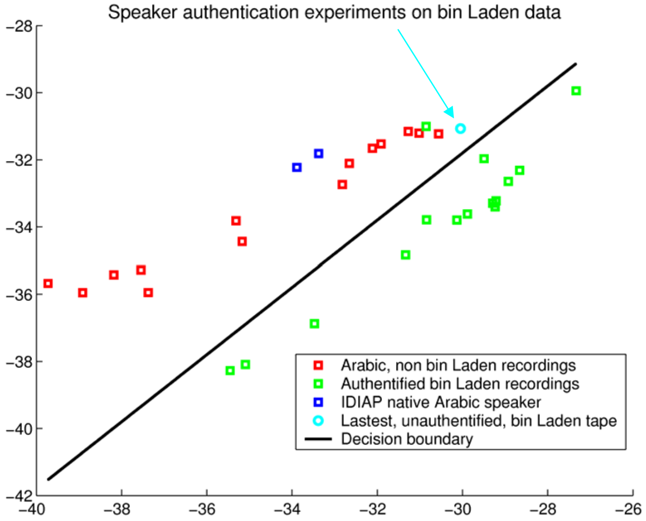


Fig. 1. 2002 IDIAP analysis of bin Laden tape [5]

4 Disguised Voices

The identification, analysis, and reversal of voice disguise are promising areas of investigation for speaker classification that are applicable to forensics and intelligence. The most systematic study of voice disguise was done by Robert Rodman (Rodman, 1998 [6]) who positioned his research on this subject within speaker classification. Rodman partitioned disguised voices into the four categories shown in Table 1 and has been since creating a database of samples for use in the development and testing of systems for identifying, categorizing, and reversing the effects of voice disguise.

The ability to detect and reverse intentional electronic disguise will be essential for the viability of SIV in the future because sophisticated voice disguise could easily merge with the work on voice forgery (usually called “voice conversion” or “voice morphing”). Voice conversion is simply the intentional electronic alteration of vocal features and patterns into the voice of a specific individual. Perrot, et al [7] assessed the threat of voice conversion to SIV systems using data

Table 1. Kinds of voice disguise [6]

Broad taxonomy of voice disguise:	DELIBERATE	NONDELIBERATE
ELECTRONIC	Electronic scrambling, etc.	Channel distortions, etc.
NONELECTRONIC	Speaking in falsetto, etc.	Hoarseness, intoxication, etc.

from the NIST speaker recognition evaluation of 2004 and found that it could pose a serious threat to existing commercial SIV technology.

5 Stress and Lie Detection

The ability to detect stress is valuable for a broad spectrum of situations in both the pubic and private sectors. It would be critical to know, for example, whether the stress levels of key employees working in nuclear weapons facilities or as international peacekeepers are too high for them to perform their jobs. A similar metric could also apply to police officers, corporate executives, and child-care workers. Being able to determine whether a suspect, informant, or witness is telling the truth would be invaluable for law enforcement and intelligence. It is equally important for business transactions and personal relationships.

Speech is an almost universal human ability. It is, therefore, fortunate that research has shown that stress affects speech in well-defined ways (Hansen and Clements, 1987 [8]; Jameson, et al, 2005 [9]; Scherer, et al. 2002 [10]). This means that stressed and unstressed speech constitute different classes of spoken behavior and that the manifestation(s) of stress in speech could be applied to the uses enumerated above.

The dominant technique for identifying stressed speech is based on "microtremor" research done in the mid-twentieth century (Lippold, 1971 [11]). Microtremors are involuntary muscular contractions that generate low-frequency oscillations (8-12 Hz) that appear to reflect the tension within muscles and seem to be part of the communication between the muscles and the nervous system. Virtually all commercial voice stress analysis and lie-detection systems utilize this approach and subsequent testing by the Air Force Research Laboratory found that these systems can distinguish stressed from unstressed speech (Haddad, et al, 2002 [12]).

Recent research reveals that stress manifests itself in a variety of ways in a person's speech (Müller et al, 2001 [13]) and that different kinds and levels of stress affect speech in different ways (Hansen, et al 2000, [14]) which indicates that stressed speech consists of a set of classes. The NATO Research Study Group (Hansen, et al, 2000 [14]) postulated four basic stressed-speech categories based on its research with military personnel. Their categories are tied to the source of the stress: physical (e.g., vibration, pressure, acceleration, equipment/physical load), physiological (e.g., alcohol, medicines, narcotics, fatigue, illness), perceptual (e.g., noise, poor communication channel, a listener who is having problems understanding), and psychological (e.g., emotion, lying, workload, anxiety) and produce unique constellations of effects on speech. Within and between their categories, unique constellations of effects on speech are produced. Lombard speech, for example, is a well-documented response to noise (perceptual stress) that has the following characteristics: increased vocal effort, greater duration of words due to increased vowel length, shifts in formant locations for vowels, increased formant amplitudes, and deletion of some word-final consonants (Markowitz, 1996, [15]).

The ability to go beyond microtremors is of particular interest to developers of speech recognition and SIV products because the acoustic manifestations of stress are known to cause the performance of these systems to deteriorate (Hansen, et al, 2000 [14]; Müller et al, 2001 [13]). Work by the NATO Research Study Group on Speech (Hansen, et al, 2000 [14]), the European Union Esprit VeriVox project (Karlsson, et al, 2000 [16]), and others on developing methods for transforming knowledge about stressed speech into tools for enhancing speech recognition and SIV products is still in its infancy.

6 Speaker Segmentation and Clustering

Speaker segmentation and clustering apply to the analysis of multispeaker environments. Those environments range from two-wire telecommunications channels that encode both (or all) speakers on the same channel to transcription and/or indexing of meetings and news broadcasts. In most cases, the number and identities of speakers is generally not known beforehand.

The goal of speaker segmentation is to identify all the boundaries between the speech of different speakers in the audio signal. In order to segment, the system must first determine whether the current speaker has changed. The most primitive method of detecting that a speaker has changed is to look for silence. This is useful as an alert to the system that the speaker may change but, used by itself, it is unreliable because speakers often pause in their speech (no speaker change) or talk over each other. The most common techniques for detecting that the speaker has changed are log likelihood ratio, Bayesian information criterion, and similar distance metrics (Ajmera, et al, 2004 [17]). They measure similarity/dissimilarity between the features extracted from consecutive slices (called "windows") of the signal. These approaches may be supplemented by higher-level change detectors, such as gender, language, dialect, and even topic. Boundaries are set at points where the distance measure is sufficiently large.

The next stage, speaker clustering, aims to identify, group all of the segments uttered by the same speaker, and assign a unique label to them (e.g., male No. 10, female No. 5) which are really speaker classes. Clustering employs variants of some of the same distance measures employed for establishing boundaries between speakers (Gish, et al, 1991 [18]; Reynolds, et al, 1998 [19]).

These techniques have been incorporated into automated indexing of broadcast news (Maybury, 2000 [20]), films, speeches, meetings, telephone conversations, and other multi-speaker audio sources. These systems still represent emerging technology but their utility has already been demonstrated in indexing of broadcast news transmissions and intelligence gathering.

Some of these systems offer semi-automated assistance to forensics and intelligence operations. Typically, such systems identify one or more classes of speakers that match a set of criteria. One commercially-available example is the Loquendo Voice Investigation System which can be used to monitor cellular call traffic looking for speakers classifications of special interest to its law-enforcement or intelligence agency clients.

7 Conclusion

This paper has demonstrated that speaker classification is a core component of SIV applications in the real world. The "classification" step within an SIV system represents the application of speaker classification in the core SIV engine. Anti-speaker technologies extend classification to enhancements to SIV systems based on comparison of spoke data with classes of speakers. Voice disguise is an area of research for forensics and intelligence that has already been partitioned into several major classes of disguise that are currently the object of research. Systems that detect stressed speech due to emotion, cognitive load, illness, and even lying are already being used commercially. At the same time, more refined analysis of the effects of different kinds of stressors is an active area of research that is designed to make SIV more robust to intra-speaker variability caused by stress. Classification is also a critical element of systems charged with transcribing, indexing, and otherwise analyzing multi-speaker communications.

References

1. Markowitz, J.: Voice Biometrics: Speaker Recognition in the Real World. Communications of the ACM 49(9), 66–73 (2000)
2. Markowitz, J.: Speaker Biometrics: The State of the Industry. In: Proceedings of the SpeechTEK West, San Francisco (2006)
3. Markowitz, J.: Anti-Spoofing for Voice. In: Proceedings of the Biometric Consortium, Washington (2005)
4. Higgins, A., Bahler, L., Porter, J.: Speaker Verification Using Randomized Phrase Prompting. Digital Signal Processing 1(2), 89–106 (1991)
5. IDIAP: Analysis of the latest bin laden tape (2002) http://www.idiap.ch/press_news.php/
6. Rodman, R.: Speaker Recognition of Disguised Voices: A Program for Research. In: Demirekler, M., Saranli, A., Altincay, H., Paoloni, A. (eds.) Proceedings of the Consortium on Speech Technology in Conjunction with the Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications, Ankara, Turkey, COST250 Publishing Arm, pp. 9–22 (1998)
7. Perrot, P., Aversano, G.: R., B., Charbit, M., Chollet, G.: Voice Forgery using ALISP: Indexation in a Client Memory. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), vol. 1, pp. 17–20 (2005)

8. Hansen, J.H.L., Clements, M.A.: Evaluation of Speech under Stress and Emotional Conditions. Journal of the Acoustic Society of America 82(1), 17–18 (1987)
9. Jameson, A., Großmann-Hutter, B., Müller, C., Wittig, F., Kiefer, J., Rummer, R.: Recognition of Psychologically Relevant Aspects of Context on the Basis of Features of Speech. In: Proceedings of the Second International Workshop on Modeling and Retrieval of Context in Conjunction with IJCAI'05, Edinburgh (2005)
10. Scherer, K.R., Grandjean, D., Johnstone, T., Klasmeyer, G., Bänziger, T.: Acoustic Correlates of Task Load and Stress. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP '02), Denver (2002)
11. Lippold, O.: Physiological Tremor. Scientific American 224(3), 65–73 (1971)
12. Haddad, D., Walter, S., Ratley, R., Smith, M.: Investigation and Evaluation of Voice Stress Analysis Technology (Final Report). Technical report, United States Department of Justice (Document No.: 193832) (2002)
13. Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., Wittig, F.: Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study. In: Bauer, M., Gmytrasiewicz, P., Vassileva, J. (eds.) UM 2001, User Modeling: Proceedings of the Eighth International Conference, pp. 24–33. Springer, Heidelberg (2001)
14. Hansen, J.H., Swail, C., South, A.J., Moore, R.K., Steeneken, H., Cupples, J.E.A., Vloeberghs, T.C.R., Trancoso, I., Verlinde, P.: The Impact of Speech Under 'Stress' on Military Speech Technology. In: NATO PROJECT 4 REPORT AC/232/IST/TG-01 Research Study Group on Speech. NATO IST/TG-01 (2000)
15. Markowitz, J.: Using Speech Recognition. Prentice Hall, Upper Saddle River (1996)
16. Karlsson, I., Banziger, T., Dankovicov, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K.: Speaker Verification with Elicited Speaking-Styles in the Verivox Project. Speech Communication 31(2), 121–129 (2000)
17. Ajmera, J., Mccowan, Iain Bourlard, H.: Robust Speaker Change Detection. IEEE Signal Processing Letters **11**(8) (2004) 649
18. Gish, H., Siu, M.H., Rohlicek, R.: Segregation of Speakers for Speech Recognition and Speaker Identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '91), Toronto, Canada, pp. 873–876 (1991)
19. Reynolds, D., Singer, E., Carlson, B., O'Leary, G., McLaughlin, J., Zissman, M.: Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP '98), Sydney (1998)
20. Maybury, M.: News on Demand. Communications of the ACM 43(2), 32–79 (2000)

Frame Based Features

Stefan Schacht, Jacques Koreman, Christoph Lauer, Andrew Morris, Dalei Wu, and Dietrich Klakow

Spoken Language Systems/Phonetics
Universität des Saarlandes
Saarbrücken, Germany
stefan.schacht@lsv.uni-saarland.de

Abstract. In this chapter we will discuss feature extraction methods for speaker classification. We introduce linear predictive coding, mel frequency cepstral coefficients and wavelets and perform experimental studies on AURORA and TIMIT data. For the speaker identification task, we can show that wavelets are beneficial.

Keywords: Speaker Identification, Feature Extraction, LPC, MFCC, Wavelets.

1 Introduction

In pattern classification problems feature extraction is very often one of the most important steps for a successful system. Surprisingly, in speaker classification, very little work has been done on task specific features. Instead, the standard feature extraction techniques, developed for speech recognition are also used for speaker classification: linear predictive coding and mel frequency cepstral coefficients. In this article, those two standard techniques will be described. In addition we will give an introduction to a wavelet feature extraction. In the final section we will describe experiments comparing wavelets to MFCCs. Unlike in speech recognition, wavelets seem to be able to outperform MFCCs on the task of speaker identification under special conditions.

2 Linear Prediction Coding

Experiments with linear prediction analysis of speech started during the 1960's mainly in search of an efficient method for coding of speech [1], [2]. Later LPC was also used for other speech related tasks as speech recognition and speaker recognition. The derivation of mel cepstrum based features reduced the relevance of LPC features, especially for speech recognition. Linear prediction is still in use in many speech coding systems, because the coefficients can be calculated efficiently and it is possible to code speech with a few kBit/s without compromising understandability. A popular example for the use of LPC is the GSM standard for mobile communication[3].

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 226–240, 2007.

2.1 A Simple Model of Speech Tract

Linear Prediction Coding(LPC) is based on a simple model of speech production. The vocal tract is modeled as a set of connected tubes with equal length and piecewise constant diameter, see Figure 1. It is assumed, that the glottis produces buzzing sounds(voiced speech) or noise(unvoiced speech). Under certain assumptions (no energy loss inside the vocal tract, no nonlinear effects ...) it can be shown that the transfer function of this model is an all-pole Filter with the z-transform

$$A(z) = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}} \tag{1}$$

where P is the number of tube segments. The coefficients $a_1 \dots a_P$ are directly related to the resonance frequencies of the vocal tract, called formants, and bear information about the shape of the vocal tract. The coefficients of the transfer function can be directly calculated from the signal through minimizing the linear prediction error:

$$e(n) = s_n - \sum_{i=1}^{P} a_i s_{n-i} \tag{2}$$

A detailed presentation of the tube model and its connection to linear prediction can be found in [4].

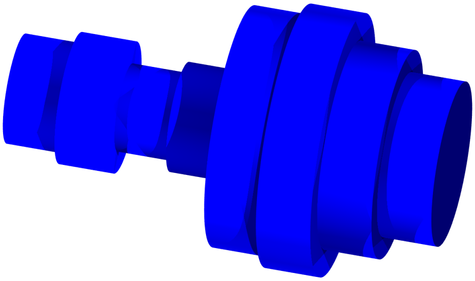


Fig. 1. A simple tube model of speech tract

2.2 Yule-Walker-Equations

There are different criteria for minimizing the linear prediction error (2). If we choose the squared expectation value we use the so called *autocorrelation method*. Other methods like the *covariance method* lead to slightly different equations. In either case it is assumed that the configuration of vocal tract and the signal don't change during a speech frame. So the following function has to be minimized:

$$\begin{aligned} J(\mathbf{a}) &= \mathbf{E}(e(n)^2) \\ &= \mathbf{E}\left(\left(s_n - \sum_{i=1}^{P} a_i s_{n-i}\right)^2\right) \end{aligned} \tag{3}$$

Using the linearity of the expectation value we get

$$J(\mathbf{a}) = \mathbf{E}(s_n^2) - 2\sum_{i=1}^{P} a_i \mathbf{E}(s_n s_{n-1}) + \sum_{i=1}^{P}\sum_{j=1}^{P} a_i a_j \mathbf{E}(s_{n-i} s_{n-i}) \tag{4}$$

The signal is assumed to be stationary, so the expectation values in this equation are independent of n and depend only on the absolute value of $i - j$. These are in fact the values of the autocorrelation function of the signal. We denote them with $R_{ss}(|i - j|)$. In practice these values have to be estimated from the signal frame. This can be done in a fast way using the *Wiener-Kinchin* theorem and the Fast Fourier Transform(FFT).

To find the minimum error solution, we calculate the gradient of the $J(\mathbf{a})$ and set it to Zero:

$$\nabla_{\mathbf{a}} J(\mathbf{a}) = 0 \tag{5}$$

A short calculation leads to the following set of linear equations:

$$\sum_{i=1}^{P} R_{ss}(|i - j|)\, a_i = R_{ss}(j) \quad j = 1 \ldots P \tag{6}$$

These are the *Yule-Walker* equations, which are well known in the theory of signal processing.

The usual way solve a set of linear equation is the Gauss algorithm, or in case of a symmetric matrix the Cholesky algorithm. These methods need $O(P^3)$ operations and don't really exploit the special structure of the *Yule-Walker* equations. A faster way is the so called levinson durbin recursion, which was first proposed by Levinson in 1947 [5] and later improved by Durbin. It needs approximately $P^2 + 2P$ operations.

We use the following notations:

$$\mathbf{a}^{(n)} = \begin{pmatrix} a_1^{(n)} \\ \vdots \\ a_n^{(n)} \end{pmatrix}, \quad \tilde{\mathbf{a}}^{(n)} = \begin{pmatrix} a_n^{(n)} \\ \vdots \\ a_1^{(n)} \end{pmatrix}$$

$$\mathbf{r}^{(n)} = \begin{pmatrix} R_{ss}(1) \\ \vdots \\ R_{ss}(n) \end{pmatrix}^{\mathrm{T}}, \quad \tilde{\mathbf{r}}^{(n)} = \begin{pmatrix} R_{ss}(n) \\ \vdots \\ R_{ss}(1) \end{pmatrix}^{\mathrm{T}}$$

initialization

$$a_1^{(1)} = \frac{R_{ss}(1)}{R_{ss}(0)}$$

step 1

$$a_n^{(n)} = \frac{R_{ss}(n) - \tilde{\mathbf{r}}^{(n-1)} \cdot \mathbf{a}^{(n-1)}}{R_{ss}(0) - \tilde{\mathbf{r}}^{(n-1)} \cdot \tilde{\mathbf{a}}^{(n-1)}}$$

step 2

$$\begin{pmatrix} a_1^{(n)} \\ \vdots \\ a_{n-1}^{(n)} \end{pmatrix} = \mathbf{a}^{(n-1)} - a_n^{(n-1)} \tilde{\mathbf{a}}^{(n)}$$

step 3 repeat **step 1** and **step 2** until $n = P$

The vector $\mathbf{a}^{(P)}$ is the solution of (6) we searched for.

3 Mel-Frequency Cepstrum Coefficients

Since their derivation approximately 30 years ago [6] mel frequency cepstrum coefficients(MFCC) became the standard feature set for various speech applications. Although originally developed for speech recognition many state-of-the-art systems for speaker classification use MFCC's as features,see [7][8][9].

3.1 Derivation

After discretization and quantization the signal gets filtered by simple high pass. This step tries to compensate the effect of the lips, which act as a low pass. This high pass has the form

$$s'_n = s_n - \alpha s_{n-1} \tag{7}$$

with values for α around 0.95.

The signal is split up into overlapping frames and the spectrum of each frame gets calculated with the *windowed discrete Fourier transformation*:

$$F_k = \sum_{n=0}^{N-1} w_n s_n e^{-\frac{2\pi i n k}{N}} \quad k = 0 \ldots N-1 \tag{8}$$

where N is the number of samples in each frame. The w_n are the coefficients of a window function. The multiplication with the window function is necessary to minimize distortions of the spectrum resulting from the finite length of the frame. A typical choice is the *Hamming window*:

$$w_n = 0.53836 - 0.46164 \cos\left(\frac{2\pi n}{N-1}\right) \quad n = 0..N-1 \tag{9}$$

We take the absolute value of each F_k because we are not interested in phase information. $|F_k|$ represents the energy of the signal at the frequency

$$\frac{k}{T}$$

where T is the duration of the frame. Since the signal is real valued the equation $|F_k| = |F_{N-k}|$ holds. The first coefficient F_0 is just the sum of the signal values of the frames, which bears no useful information for our purposes. So we get $\frac{N}{2}$ unique features from the discrete Fourier Transform.

A naive calculation of (8) needs N^2 complex multiplications. Using the *Fast Fourier Transform* it is possible to reduce the number of multiplications to $O(n \log n)$. Further reductions are possible if we use fact that the signal has only real values.

The ability of the human ear to discriminate signal frequencies decreases with higher frequencies. To model this a set of overlapping band pass filter is applied to the $|F_k|$. The center frequencies of these band pass filters are equally spaced on the *mel scale*

$$mel_f = 1127.01048 \cdot \ln\left(1 + \frac{f}{700}\right) \tag{10}$$

which was first proposed in [10]. A common choice in speech recognition are 24 triangle shaped filters which start and end at the center frequencies of its neighbors, see figure 2.

In the next step the logarithm is taken from the output of the filter bank. This serves two purposes. First these values are can be easier modeled by Gaussian distributions or Gaussian Mixture Models(GMM). This allows the application of well established classification methods. The second aspect is even more important:

We model speech generation as an LTI system with impulse response $h(n)$ and input signal $e(n)$

$$s(n) = h(n) * e(n) \quad . \tag{11}$$

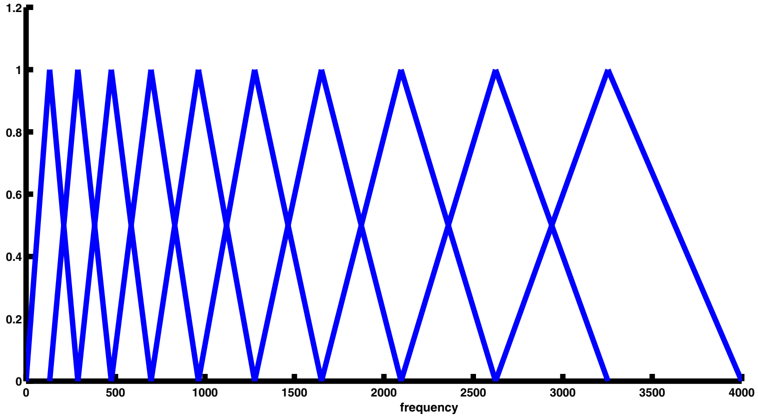


Fig. 2. Mel scale filterbank

A short calculation shows that the signal processing steps so far lead to a separation of the input signal and the impulse response:

$$\begin{aligned}\log\left|\mathcal{F}\left\{h(n) * e(n)\right\}\right| &= \log\left|\mathcal{F}\left\{h(n)\right\} \cdot \mathcal{F}\left\{e(n)\right\}\right| \\ &= \log\left(\left|\mathcal{F}\left\{h(n)\right\}\right| \cdot \left|\mathcal{F}\left\{e(n)\right\}\right|\right) \\ &= \log\left|\mathcal{F}\left\{h(n)\right\}\right| + \log\left|\mathcal{F}\left\{e(n)\right\}\right|\end{aligned} \tag{12}$$

This can be used to remove the effects of the transmission channel, for instance a telephone line. By subtracting the mean feature vector over a certain period of time from each feature vector the distortions from a linear transmission channel can be removed. This procedure is known as *Cepstral Mean Subtraction* (CMS). It is very popular in speech recognition, because it leads to major improvements in recognition results under severe conditions. Unfortunately CMS also removes speaker dependent information about the configuration of the vocal tract. So it actually decreases the performance in speaker recognition tasks.

As a last step a *Discrete Cosine Transform*(DCT) is applied to the logarithmized output values m_j of the filter bank:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}\left(j - 0.5\right)\right) \tag{13}$$

It serves as an easier to calculate approximation of the *Principal Component Analysis*, see [11] . The output values c_i are nearly uncorrelated. So if Gaussian Mixture Models are used for classification, see [12] [13], diagonal covariance matrices are sufficient. That means fewer parameters have to be estimated. Usually only the first 12 output values of the DCT are taken as feature values.

3.2 Dynamic Features

The features derived so far only contain informations about a single speech frame. To capture informations about the changes of the signal in time the vector can be supplemented with approximations of the first and second derivation of the vector components. A simple method to calculate these approximations are difference quotients.

$$\begin{aligned}\varDelta c_k^{(m)} &= c_k^{(m+\tau)} - c_k^{(m-\tau)} \\ \varDelta^2 c_k^{(m)} &= \varDelta c_k^{(m+\tau)} - \varDelta c_k^{(m-\tau)}\end{aligned}$$

But more common is the use of linear regression, first suggested in [14]. The slope of the regression line yields a more robust estimate of the first derivation and can be easily calculated

$$\delta c_k^{(m)} = \frac{\sum_{i=-\tau}^{\tau} i c_k^{(m+i)}}{\sum_{i=-\tau}^{\tau} i^2} \tag{14}$$

Then the same formula can be used to calculate estimates for the second derivation from the $\delta c_k^{(m)}$. A common choice for the window size is $\tau = 2$, see figure 3.

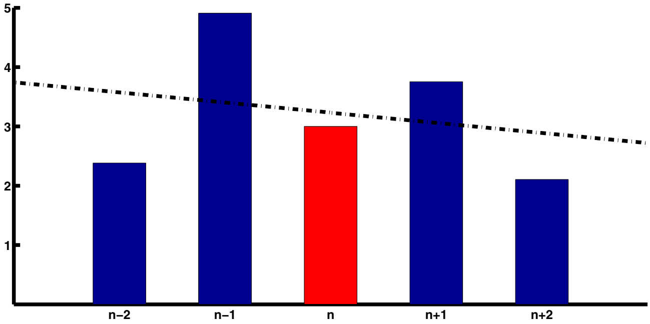


Fig. 3. Estimation of the first derivation using linear regression

4 Wavelet Based Features

A disadvantage of the Fourier transform, used in the mel cepstrum coefficients, is the missing ability to resolve the temporal characteristics of the signal. For this reason, the assumption of a stationary signal within the analyzed frame has to be made.

A method to capture temporal and frequency properties of the signal is the wavelet transform. It is currently very popular for task of image compression, see [15] [16], but it was also tried in audio applications like speech enhancement, speech recognition or speaker recognition. Despite its theoretical advantages, the wavelet transform has not yet been used in the majority of speech processing systems.

The wavelet transform is based, like the Fourier transform, on an approximation of the signal with a set of orthogonal functions. Starting with a so called "mother wavelet" $\psi(t)$ the set of orthogonal functions

$$\psi_{m,k}(t) = 2^{m/2}\psi(2^m t - k) \quad k, m \in \mathbb{Z} \tag{15}$$

is generated trough translation and dilatation. Nearly all wavelets that are used in practice posses a compact support, that means only a limited part of the wavelet is different from zero. So each approximation coefficient, calculated through

$$a_{m,k} = \int_{-\infty}^{\infty} \psi_{m,k}(t) s(t)\, dt \tag{16}$$

captures only informations from a limited part of the signal $s(t)$. A detailed description of the wavelet transform and its derivation can be found in [17].

The discrete version of the wavelet transform can be implemented as a set of finite impulse response filters. Starting with the sampled signal $(s_0, s_1, \ldots, s_{N-1}$ with sampling frequency f_s a high pass and a low pass are applied simultaneously to the signal. The output of the high pass represents the details of the signal in the frequency band $\frac{f_s}{4} \ldots \frac{f_s}{2}$. The output of the low pass gets down sampled by a factor of two and the same set of filters will be applied to the down sampled output. Through repeating these steps a multi resolution decomposition of the signal is generated.

A generalization of the discrete wavelet transform is the wavelet packet transform, see [18]. Here the set of filters is also selectively applied to the high pass branch of the signal decomposition. Sarikaya et al. used this transform to generate an a approximation of the mel scale filter bank, see [19]. They report an improvement for speaker recognition compared to the standard MFCC coefficients. Other groups simply calculated the energies of the sub-bands generated by the wavelet decomposition and used these numbers as features for speech recognition, see [20] [21]. But with both methods the localization informations which the wavelet coefficients provide get lost.

5 Comparison of MFCCs and Wavelets for Speaker Recognition and Speech Recognition

We compare the performance of MFCC features and wavelet packet features, as suggested by Sarikaya , for speaker and speech recognition under noisy conditions.

5.1 Data Sets

We used the AURORA data base for the speech recognition experiments. The AURORA database was created to evaluate the robustness of speech recognition in real-life, noisy conditions. The database is described in [22] and is based on the TIDigits database, which contains connected English digit sequences recorded in a quiet room using a high-quality microphone [23]. The original sound recordings were filtered with impulse responses typical for telecommunication equipment and different noises(suburban train, babble, car, exhibition hall , restaurant, street, airport and train station noise) were added at different signal to noise ratios. The training set of the data base consists of clean speech and noisy speech with a signal to noise ratio from 20dB to 0dB. The test set has three different parts: set A for additive matched noise, set B with additive mismatched noise and set C with additive noise and convolutional noise.

For the speaker recognition tests we used the TIMIT data base, which is described in [24]. As in [19] the original 16 kHz data was down sampled to 8 kHz, partly to simulate telephone speech and partly because speaker identification with the original data is too easy (near 100% correct). This also permits a better comparison with the speech recognition results on the 8 kHz AURORA database. Since TIMIT was recorded for speech, not speaker recognition, the standard division into training and test sets is not well suited for work on speaker identification. We therefore created our own division into training, development and evaluation data. All sentences of type SA_{1-2}, SI_{1-2} and SX_{1-2} were used for training, SX_3 and SI_3 for development and SX_4 and SX_5 for evaluation. For testing under noisy conditions we processed the speech data in same way as in the AURORA data base and added different ambient noises. These noises were taken, as in the development of the AURORA data base, from the NOISEX data base [25].

5.2 Feature Derivation

For both experiments we used the Hidden Markov Modeling Toolkit HTK [26] to compute the mel cepstrum based features. For speech recognition we used 12 MFCCs plus energy, as well as the corresponding first and second time difference parameters fitted over a 5 frame window, as in the reference set-up described in [22]. Although the noises in all our tests were additive only, as cepstral mean subtraction (CMS) is easy to apply with MFCCs, CMS was performed for noise cancellation.

The MFCCs for speaker recognition were derived with 24ms window and a 10ms shift, a pre-emphasis factor of 0.97, a Hamming window and 20 Mel scaled feature bands. As in [19] and contrary to the speech recognition features all 20 MFCC coefficients were used except c_0. Because neither silence removal nor dynamic features enhanced performance, these were not used. Cepstral mean subtraction was also tested, as for the Aurora data.

Wavelet based features for both experiments were extracted as in [19]. After using the same setup for window size, skip rate and pre-emphasis as in the derivation of the MFCCs, a wavelet packet tree were applied to the speech frame. The tree used 32nd order daubechies wavelet filter coefficients and represents a roughly Mel-scaled distribution of the subbands across frequency. The log energy in the 24 subbands was decorrelated by DCT, the same as for MFCCs, resulting in subband based cepstral parameter (SBCs). As an alternative to the DCT analysis we also used a second wavelet transform (level 3 transform using Daubechies 4 tap filters) to orthogonalize the data. Since the results for these features(WPPs) were very similar to the SBC features, they are not reported here.

Because there is no convolution theorem for the wavelet transform, cepstral mean subtraction isn't possible for SBCs. So we used another method for noise removal, called Super soft thresholding. The method is described in [27]. The assumption behind this procedure is that noise is more evenly distributed over all coefficients than speech, and by performing thresholding on the wavelet coefficients only the noise is suppressed. The continuous transfer function of Super soft thresholding is given by

$$y = \begin{cases} x - sign(x)(1-\alpha)t & \text{if } |x| \geq t \\ \alpha x & \text{if } |x| < t \end{cases}$$

and was chosen because it is assumed to lead to less distortion in the speech signal then simple thresholding. For our experiments we choose $\alpha = 0.5$ and $t = 2.5$.

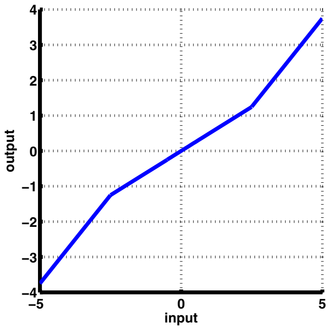


Fig. 4. Super soft thresholding transfer function for $\alpha = 0.5$ and $t = 2.5$

5.3 The Speech Recognition System

The speech recognition system used is the standard HTK hidden Markov model (HMM) reference recognizer for AURORA [22], but uses 20 Gaussians per state as in the Microsoft complex baseline system. This leads to an improvement in word accuracy compared to the three Gaussians used in the standard AURORA speech recognizer. All HMMs were trained on multi-condition Aurora data.

5.4 The Speaker Recognition System

For speaker identification a Gaussian mixture model (GMM) consisting of 32 Gaussians is trained for each speaker. As in [19] [28] , GMMs were trained by k-means clustering, followed by EM iteration. This was performed by the Torch machine learning toolkit [29] . The training and testing were always performed on matched noise conditions. We used a variance threshold factor of 0.01 and minimum Gaussian weight of 0.05 (performance falling sharply if either were halved or doubled).This simple model gives state-of-the-art speaker recognition performance. With TIMIT (though not with other databases, such as the CSLU speaker recognition database) no gain was found in training speaker models by adaptation from a world model, so no world model was used.

5.5 Results

Speech Recognition. For speech recognition on Aurora the SBC features generally lead to lower word accuracies than the MFCCs, with the exception of test B when no noise cancellation is used. In all cases, noise cancellation (either CMS or thresholding) improves word accuracy. The effect is strongest for CMS on the MFCC features, particularly for test C (for which CMS is clearly better suited). The full results are given in Table 1.

Table 2 focuses on a subset of the data. It shows the average word accuracy across all noises, but only for very low SNRs (5, 0 and -5 dB). The overall tendencies are the same as in Table 1, but the effect of CMS is much greater, while the effect of thresholding in the wavelet domain does not change so strongly

Table 1. Speech recognition word accuracy for different feature types with multi-condition trained HMMs (best results in bold type)

TEST	MFCC	MFCC+CMS	SBC	SBC+T
A	92.29	**93.22**	91.50	92.07
B	90.03	**92.63**	90.20	90.62
C	87.44	**91.87**	86.67	87.82

(and even has a negative effect for test B). As for the overall data in Table 2, the effect of CMS is greatest for test C.

Table 2 also shows that SBCs with thresholding consistently leads to a higher word accuracy at low SNRs. The tendency for SBCs with thresholding to outperform MFCCs (without CMS) can also be observed in Table 1, except for test A.

When we compare car noise and babble (AURORA noise conditions A2 and A3) to clean speech, we again find that MFCCs with CMS give the best performance (except for clean speech, where there is no change in word accuracy). A large effect of noise cancellation, both for CMS and thresholding, can be observed for car noise, but not babble, at SNR=0 , see table 3.

Note that the advantage of SBC features with thresholding over MFCCs (without CMS), which was observed for the overall results in Tables 1 and 2, disappears.

Speaker Recognition. For the clean TIMIT data, speaker identification results for SBC wavelet features are consistently higher than for the MFCC features, both with and without noise cancellation. Noise cancellation, both wavelet thresholding and especially cepstral mean subtraction) has a negative effect on the speaker identification accuracy, as should be expected for clean speech. For babble, wavelets perform better than MFCC features for two of the three noise conditions. With noise cancellation, the wavelets perform better than MFCC features, but in all cases the speaker identification accuracy without noise cancellation is higher. The car noise results are better for MFCC features than for wavelets in all cases. With CMS, though, the results for MFCCs are worse than for wavelets, except at SNR=0 dB. The results are summarized in Table 4.

5.6 Discussion

In the speech recognition experiments for the AURORA database, the MFCC features with CMS consistently lead to the best performance. MFCCs with CMS

Table 2. Speech recognition word accuracy for different feature types with multi-condition trained HMMs, but only SNRs from 5 to -5 dB (best results in bold type)

TEST	MFCC	MFCC+CMS	SBC	SBC+T
A	62.83	**66.97**	60.96	63.67
B	59.07	**65.49**	59.74	59.29
C	51.43	**64.55**	49.47	52.88

Table 3. Speech recognition word accuracy for different feature types with clean and noisy data (best results in bold type)

NOISE	SNR	MFCC	MFCC+CMS	SBC	SBC+T
clean	-	**99.53**	**99.53**	99.25	99.18
babble	20	98.91	**98.97**	98.85	98.67
babble	10	96.77	**97.25**	96.83	96.28
babble	0	68.68	**69.62**	67.87	68.62
car	20	99.19	**99.28**	98.87	98.63
car	10	97.38	**97.70**	96.39	96.45
car	0	62.54	**70.27**	54.88	60.66

Table 4. Speaker identification accuracy for different feature types with clean and noisy data (best results in bold type)

NOISE	SNR	MFCC	MFCC+CMS	SBC	SBC+T
clean	-	91.11	78.81	**94.05**	92.62
babble	20	76.19	62.54	**76.43**	75.79
babble	10	**50.63**	39.21	46.83	45.00
babble	0	6.75	4.37	**9.44**	8.65
car	20	**83.73**	70.08	81.75	81.19
car	10	**71.11**	60.40	66.51	65.79
car	0	**46.43**	41.35	38.76	36.90

are clearly more viable than SBC features, with or without thresholding, for the AURORA speech recognition tasks. CMS is useful to deal with the varying noises in the training data, and enhances the signal match with the test conditions.

In a comparison between MFCCs and wavelet based features in noisy telephone speech (WPP features, which lead to similar results as SBCs in our experiments, as noted in section 5.2) for the Slovenian and English SpeechDat2 data, WPPs without thresholding were found to even outperform MFCCs (without CMS), especially for the noisier English data [30]. Instead our experiments with the AURORA data showed only a slight advantage for wavelet based features when we used thresholding.

But MFCC features with CMS do not always outperform wavelet-based features, as was the case for the Aurora speech recognition experiments. In the speaker identification experiments on TIMIT, the two noise cancellation techniques always cause a deterioration in performance, see table 4. This is particularly true for CMS. Since CMS was developed to deal with differences in the convolutive noise in the data, it is not very surprising that it does not enhance speaker identification in the additive noise conditions, where the noise was moreover identical in the training and test conditions (unlike for the AURORA experiments, as noted above). CMS removes speaker characteristics together with any residual convolutive noise. But if CMS only works for convolutive noise, this does not explain why CMS works in speech recognition tests A and B, see tables 1 and 2, since the same filter is used for the training and test conditions.

Obviously, CMS also homogenizes the data for the different noise conditions used in the AURORA training and testing.

The negative effect of noise cancellation by thresholding in the wavelet domain is much smaller for the TIMIT speaker recognition experiments, probably because only small scale spectral details are discarded even if these details represent speaker properties instead of noise.

In the TIMIT speaker identification task we find that SBCs outperform MFCCs for clean speech, see table 4 as well as for 2 out of 3 babble conditions.

There are several possible reasons for the better performance of SBCs compared to MFCCs in these conditions. Of course, the task was different for the AURORA and the TIMIT data (speech versus speaker recognition). But as was pointed out above, another possible reason is that the match between the training and test data is better for TIMIT than for AURORA data, since for the latter the training data always includes several noise conditions, while the testing is only for a subset of them. Further speech and speaker recognition experiments on the two databases may help us to better understand the reasons for the different results.

Lastly, the wavelet processing used in this article (and many others) does not exploit the advantages of the optimal time-frequency resolution that wavelets offer. This is because a standard HMM/GMM decoder is not able to deal with data streams which have different data rates. However, decoders which allow asynchronous stream combination which have recently been developed and applied to multi-modal feature combination [31,32] could be used for this purpose. The analysis at multiple scales in time and frequency which wavelets can give would also provide complementary data streams whose combination by such methods may add to noise robustness.

References

1. Itakura, F., Saito, S.: Analysis synthesis telephony based upon the maximum likelihood method. In: Kohasi, Y. (ed.) Reports of 6th Int. Cong. Acoust. (1968)
2. Itakura, F., Saito, S.: Analysis synthesis telephony based on the partial autocorrelation coefficient, Acoust. Soc. of Japan Meeting (1969)
3. Mouly, M., Pautet, M.B.: The GSM System for Mobile Communications. Telecom Publishing (1992)
4. Markel, J.D., Gray, A.H.: Linear prediction of speech. Springer, Heidelberg (1976)
5. Levinson, N.: The wiener rms error criterion in filter design and prediction. J. Math. Phys. 25, 261–278 (1947)
6. Davis, S., Mermelstein, P.: Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences. IEEE Trans. on Acoustics, Speech, and Signal Processing 28, 357–366 (1980)
7. Heerden, C.J.v., Barnard, E.: Durations of context-dependent phonemes: A new feature in speaker verification. In: M"uller, C. (ed.) Speaker Classification. LNCS(LNAI), vol. 4441, Springer, Heidelberg (2007) (this issue)
8. Bellegarda, J.R.: Language–independent speaker classification over a far–field microphone. In: M"uller, C. (ed.) Speaker Classification. LNCS(LNAI), vol. 4441, Springer, Heidelberg (2007) (this issue)

9. Garcia, G., Jung, S.-K., Eriksson, T.: Bayes-optimal estimation of gmm parameters for speaker recognition. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007) (this issue)
10. Stevens, S.S., Volkmann, J., Newmann, E.B.: A scale for the measurement of a psychological magnitude pitch. Journal of the Acoustical Society of America 8, 185–190 (1937)
11. Jain, A.: A sinusoidal family of unitary transforms. In: PAMI (1979)
12. Schulz, T.: Speaker characteristics. In: Müller, C. (ed.) Speaker Classification. LNCS (LNAI), vol. 4343, Springer, Heidelberg (2007) (this issue)
13. Sturim, D.E., Campbell, W.M., Reynolds, D.A.: Classification methods for speaker recognition. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007) (this issue)
14. Furui, S.: Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustic, Speech, and Signal Processing 34, 52–59 (1986)
15. Bradley, J.N., Brislawn, C.M., Hopper, T.: Fbi wavelet/scalar quantization standard for gray-scale fingerprint image compression. In: Proc. SPIE. vol. 1961, pp. 293–304 (1993)
16. Christopoulos, C.A., Ebrahimi, T., Skodras, A.: Jpeg 2000: the new still picture compression standard. In: Proceedings of the ACM workshops on Multimedia, pp. 45–49 (2000)
17. Daubechies, I.: Ten Lectures on Wavelets (C B M S - N S F Regional Conference Series in Applied Mathematics). Soc. for Industrial & Applied Math. (1992)
18. Vetterli, M., Kovacevic, J.: Wavelets and Subband Coding. Prentice-Hall, Englewood Cliffs (1995)
19. Sarikaya, R., Pellom, B., Hansen, J.: Wavelet packet transform features with application to speaker identification. In: NORSIG'98, pp. 81–84 (1998)
20. Erzin, E., Cetin, A.E., Yardimici, Y.: Subband analysis for robust speech recognition in the presence of car noise. In: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE Computer Society Press, Los Alamitos (1995)
21. Kim, K., Youn, D.H., Lee, C.: Evaluation of wavelet filters for speech recognition. In: Proc. of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE Computer Society Press, Los Alamitos (2000)
22. Hirsch, H.G., Pearce, D.: The AURORA experimental framework for the performance evaluation of speech recognition under noisy conditions. In: Proceedings of the ISCA ITRW ASR (2000)
23. Leonard, R.: A database for speaker independent digit recognition (1984)
24. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N.: Darpa timit acoustic-phonetic continuous speech corpus cd-rom (1993)
25. Varga, A., Steeneken, H.: Assessment for automatic speech recognition: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12(3), 247–251 (1993)
26. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 2.2. Entropic (1999)
27. Nordstrm, F., Holst, J., Lindoff, B.: Time and frequency dependent noise reduction in speech signals. In: Proc. Int. Conf. on Signal Proc. Appl. and Techn. (1999)
28. Reynolds, D.A.: Experimental evaluation of features for robust speaker identification. IEEE Transactions on SAP 2, 639–643 (1994)
29. Collobert, R., Bengio, S., Marithoz, J.: Torch: a modular machine learning software library. Technical report (2002)

30. Modic, R., Lindberg, B., Petek, B.: Comparative wavelet and mfcc speech recognition experiments on the slovenian and english speechdat2. In: Proc. Isca-ITRW NOLISP (2003)
31. Bengio, S.: Multimodal speech processing using asynchronous hidden markov models. In: Proc. Information Fusion (2004)
32. Bengio, S.: Multimodal authentication using asynchronous HMMs. IDIAP-RR 03-02 (2003)

Higher-Level Features in Speaker Recognition

Elizabeth Shriberg

SRI International, Menlo Park, CA
International Computer Science Institute, Berkeley, CA
ees@speech.sri.com

Abstract. Higher-level features based on linguistic or long-range information have attracted significant attention in automatic speaker recognition. This article briefly summarizes approaches to using higher-level features for text-independent speaker verification over the last decade. To clarify how each approach uses higher-level information, features are described in terms of their type, temporal span, and reliance on automatic speech recognition for both feature *extraction* and feature *conditioning*. A subsequent analysis of higher-level features in a state-of-the-art system illustrates that (1) a higher-level cepstral system outperforms standard systems, (2) a prosodic system shows excellent performance individually and in combination, (3) other higher-level systems provide further gains, and (4) higher-level systems provide increasing relative gains as training data increases. Implications for the general field of speaker classification are discussed.

Keywords: Speaker recognition, speaker verification, higher-level features, high-level features, long-range features, prosodic features, stylistic features, automatic speech recognition, prosody, phonetic speaker recognition, speaker idiosyncrasies.

1 Introduction

The broad field of speaker classification makes use of a wide range of properties of spoken language—from lower-level features reflecting voice parameters to higher-level features that capture phonetic, prosodic, and lexical information. In subfields such as emotion, language, and dialect classification, among others, higher-level features play an important role in both human-based and automatic classification. In forensic phonetics, for example, it is common practice for human experts to use not only voice characteristics but also speaker characteristics based on pronunciation, prosody, and lexical information to assess the match between a suspect's speech and speech in a recording of interest.

In contrast, in automatic speaker recognition, the dominant approach in both commercial and research systems has been the use of Gaussian mixture models (GMMs) to model distributions of spectral information from short time frames of speech [1,2,3]. This approach, which reflects information about a speaker's vocal physiology, is highly successful, is simple, and has the further advantage of applicability to text-independent recognition since it does not rely on phonetic content. Nevertheless, it fails to capture

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 241–259, 2007.

a wealth of longer-range and linguistic information that also resides in the signal. As we will see, such higher-level information can significantly improve performance when combined with lower-level cepstral information. Higher-level information also offers the possibility of increased robustness to channel variation, since features such as lexical usage or temporal patterns do not change with changes in acoustic conditions. And finally, higher-level features can provide useful metadata about a speaker, such as what topic is being discussed, how a speaker is interacting with another talker, whether the speaker is emotional or disfluent, and so on.

The last decade has seen increased interest in exploring such higher-level features in automatic speaker recognition. One facilitating factor in this regard has been the greater availability of high-performance automatic speech recognition (ASR) systems. A second is the development of data resources and task definitions that encourage the study of higher-level features, which typically operate at longer ranges and thus require longer train and test samples. An influential task has been the "Extended Data" task in speaker recognition evaluations (SREs) conducted by the National Institute of Standards and Technology (NIST) [4]. Since its introduction in 2001, the task has provided speaker training and test data consisting of whole conversation sides, with multiple sides available in training.

The goals of this chapter are (1) to provide a brief overview of work on higher-level features, and (2) to demonstrate how higher-level features can contribute to performance in a state-of-the-art system. Since the term "higher-level" (as well as a host of similar terms) has had different meanings in the literature, a working definition is certainly in order. We will take a broad view and include as higher-level any features that involve either *linguistic information* or *information at longer time spans than used in frame-based systems*. As discussed in more detail in the section to follow, linguistic information will refer to information that requires an automatic speech recognition system. Linguistic information is further specified according to its use for either feature extraction or feature conditioning. Longer-time-span information refers to features that are either extracted over regions longer than a frame or to approaches that model frame sequence dynamics. As will be shown, approaches using linguistic information are typically longer-range, but this is not always the case. And conversely, approaches using longer-range information often, but not always, use linguistic information. An example of approaches that fall outside the scope of this review (i.e., that are not considered higher-level) are those based on distributions of frame-level pitch or energy [5,6,7,8], which, although often called "prosodic," involve neither linguistic nor long-range information.

Section 2 provides a summary of approaches to using higher-level features for text-independent speaker verification over the last decade. To clarify how each approach uses higher-level information, features are described in terms of a classification framework that specifies their type, time span, and reliance on automatic speech recognition for feature extraction and feature conditioning. Section 3 presents an analysis of higher-level features in a state-of-the-art system that includes multiple low-level and higher-level subsystems. Section 4 concludes with general implications for speaker recognition and the wider field of speaker classification.

2 Overview and Classification of Approaches

Although not exhaustive, the overview in this section aims to illustrate the wide array of methods and techniques used in higher-level feature modeling. A summary is presented in Table 1. In the first column of the table, and in the text to follow, features are grouped according to their feature *type*—progressing from lower-level cepstral features and features that essentially tokenize the acoustic space, to prosodic and finally to word-based features. The *description* of the feature is meant to convey its nature and contrast it to other features in the table; it may not match the original term used by the researcher(s).

To further specify just how each approach uses high-level information, and to contrast it with other approaches, three additional dimensions are introduced:

1. The temporal span of the feature
2. The level of ASR used for *feature extraction*
3. The level of ASR used for *region conditioning*

A feature's *time span* refers to the temporal region over which the feature is defined. We make a contrast here between frame-level and longer regions; this, of course, is a continuum. Note that a longer time span can be the result either of using a longer feature extraction region (e.g., a region based on lexical information) or of modeling sequential information based on frame-level features (e.g., pitch or energy dynamics over a sequence of many frames).

ASR used for feature extraction will refer to the highest level of ASR information needed to define and extract the feature. Features that require the output of an automatic speech recognition system necessarily involve some amount of linguistic information, but ASR systems can utilize varying degrees of linguistic constraints. At one end of the continuum are “open loop” phone recognizers, which decode using acoustic phone models but no phonotactic, lexical, or syntactic constraints. These systems essentially provide a means of tokenizing the acoustic space according to recognizer phone models. They often produce unusual (even unpronounceable) phone sequences that differ from those associated with the possible dictionary pronunciations for the words a speaker actually says. It is precisely because of these mismatches that such systems are useful in speaker recognition: the mismatches can reveal acoustic tendencies correlated with particular speakers. A step further in the direction of linguistic constraints involves imposing phonotactic constraints obtained from an N-phone language model. This approach restricts output to phone sequences that are observed in the language. At the extreme, the recognizer uses pronunciation dictionaries and word-level N-gram language models to hypothesize phones and words that make sense as part of complete sentence hypotheses. Higher-level features based on such output aim to capture information associated with specific words or word sequences, including not only their frequency of occurrence but also their acoustic realization, pronunciation, and prosodic rendering.

Finally, *ASR used for region conditioning* will be used to refer to the highest level of ASR required for filtering the output stream of features. If chosen appropriately, conditioning can improve speaker recognition in two ways: by reducing variability or by shifting means. Conditioning can reduce the variance of feature distributions by collecting data over more constrained (and thus more homogeneous) regions. And it can focus on regions that exhibit greater inherent between-speaker variation, i.e., that move the means of one speaker's feature distribution farther away from those of other speakers. Both effects result in improved speaker discrimination.

While in principle any features can restrict comparison regions to subspans of speech, region conditioning fragments the data. Thus, there is typically some linguistic motivation that the cost of data fragmentation will be more than offset by the reduction in variability and/or shift in means brought about by the conditioning. A simple example is that of text-constrained cepstral features. The features themselves are neither long-range nor ASR-dependent. The only sense in which they are higher-level is in their region conditioning, which compares cepstral features of specific words or subword units to reduce within-speaker spectral variability associated with phonetic content. A second example is the maximum likelihood linear regression (MLLR) approach (see below), which factors out phonetic content both at the level of features (by using transforms derived from ASR phone information) and at the level of conditioning (by comparing transforms for specific phones individually).

Region conditioning is not restricted to variability reduction for phonetic content. For example, as described further below, the constrained prosody sequence approach conditions on sets of words that behave similarly prosodically. Although in principle region conditioning includes any means of reducing within-speaker variability on the feature of interest, in practice it has typically involved information from ASR. An exception, however, is the conditioned syllable-based prosody sequence model (see below), which in addition to conditioning on words makes significant use of pause contexts (obtainable without ASR). One can imagine other contexts (e.g., regions of high energy) that could also prove useful in constraining regions of interest.

A description of the approaches in Table 1 follows. Because studies differ in terms of data sets, amount of training data, ASR systems, combination of the approach with other systems, and other factors, it is not possible to compare performance directly. Nevertheless, we can look at performance in two ways. First, we can look at the relative error reduction that higher-level features contribute when combined with a baseline cepstral GMM system. Such information is provided at the end of the discussion of each feature type. Second, a within-site study allowing for direct comparisons of the performance of high-level systems is provided in Section 3. The analysis includes only a subset of feature types but uses state-of-the-art systems and recent NIST evaluation data.

2.1 Cepstral and Cepstral-Derived Features

Several approaches use the output of a word or phone recognizer to condition the extraction of cepstral features, thereby reducing variability associated with phonetic content. A review of some of these approaches is provided in [9]. Note that constraining the features to specific words essentially confers on text-independent speaker models some of the advantages of text-dependent speaker verification. The approach in [10] conditions a cepstral GMM on the identities of frequent words, based on recognizer word alignments. A variant conditions on syllables rather than words [11]. Another approach is to use multi-state HMMs as used in ASR as speaker models, thereby conditioning at the phone level but also capitalizing on a more detailed model of the sequential aspects of

Table 1. A Multidimensional Classification of Higher-Level Features in Speaker Recognition. *Ref.* = selected reference(s), *GMM* = Gaussian mixture model, *SVM* = support vector machine, *HMM* = hidden Markov model, *DTW* = dynamic time warping, *MLLR* = maximum likelihood linear regression, *unc.* = unconstrained, *LR* = likelihood ratio, *rec.* = recognition, *artic.* = articulatory, *freq.* = frequencies, *POS* = part of speech.

Feature	Feature	Time	ASR Used for		Model	Ref.
Type	Description	Span	Feature Extraction	Region Conditioning		
Cepstral	phone-conditioned cepstral models	frame	none	phones, classes	GMM, SVM	[9]
	text-conditioned GMMs	frame	none	words, syllables	GMM	[10, 11]
	phone HMMs	frame	phone, word	phone	HMM	[12, 13]
	whole-word models	longer	none	frequent word N-grams	word HMM	[14]
	DTW word models	longer	none	frequent word N-grams	template	[15, 16]
Cepstral-derived	MLLR transforms	frame	word, unc. phone	phone	SVM	[17]
Acoustic tokenization (“phonetic”)	phone N-gram freq.	longer	unc. phone	none	LR, SVM	[18–22]
	word-conditioned phone N-gram freq.	longer	unc. phone	frequent word N-grams	SVM	[23]
	conditioned pronunciation model	longer	unc. phone + word	phones from word rec.	LR	[24]
	conditioned pronunciation model	longer	unc. phone + artic.	phones from unc. phone rec.	LR	[25]
Prosodic	prosody dynamics	longer	none	none / phone	LR	[26, 27] / [27]
	DTW word-pitch models	longer	none	word	template	[27]
	interpause / conversation-level statistics	longer	word	none	GMM / LR	[28] / [29]
	word-constrained phone duration	longer	word	word	GMM	[30]
	phone-constrained state duration	longer	word	phone	GMM	[30]
	syllable-based prosody sequence	longer	word	none / words, POS	SVM	[31, 32] / [33]
Lexical	word N-grams	longer	word	none	LR, SVM	[34, 35]
Lexico-prosodic	duration-conditioned word N-grams	longer	word	none	SVM	[36]

speech, in contrast to the bag-of-frames model used in GMM-based approaches. HMM speaker modeling can be based on phone recognition [12] or word recognition [13]. A more recent variant [14] uses whole-word HMMs, thereby enabling even more detailed modeling; the HMMs represent not only words but frequent bigrams and trigrams as well. Whole words and phrases are also modeled by [15], but in a nonparametric fashion. The cepstra for a given phrase are aligned by using dynamic time warping to fit a standard length, after which the stacked cepstral feature vectors can be compared directly and a match score computed. In [16], dynamic time warping is used both to find frequent words and to score them against the speaker model of the word.

The MLLR approach [17] uses speaker-specific model adaptation transforms from a speech recognizer (either phone or word level) as features, modeled by a support vector machine (SVM). Instead of cepstral features, it uses the *difference* between speaker-adapted Gaussian means and corresponding speaker-independent means as features. This difference is expressed as the coefficients of an affine transform that rotates and shifts the speaker-independent model to obtain a speaker-dependent model, computed with maximum likelihood linear regression. Furthermore, the Gaussian models used in this approach are not unstructured GMMs but the detailed context-dependent phone models used in a speech recognizer, making the resulting features text independent. This has the advantage that features are text independent while being shared among all instances of a given phone, thus avoiding the data fragmentation implied by the conditioning on words. Transforms specific to different phone classes are combined for greater representational detail.

Cepstral models are usually the most accurate speaker recognition models when used on their own. State-of-the-art cepstral systems give about 4% to 5% equal error rate (EER) on the most recent NIST SRE test set when trained on 1 conversation side per speaker, and roughly 2% to 3% EER with 8 sides of training data, after intersession variability compensation. Small gains (about 10% to 15%) can be achieved by combining more than one state-of-the-art cepstral system. Systems using phone- or word-conditioned cepstral models typically are not much better than standard (unconditioned) cepstral models when used on their own. But they can provide substantial gains when combined with the latter, with reported improvements of up to 50% for 8-side training [10,14]. It is not yet known how such systems combine when multiple cepstral systems are available.

2.2 Acoustic Tokenization ("Phonetic") Features

A large body of work, often referred to as "phonetic" recognition or modeling, employs unconstrained phone recognition essentially as a means by which to discretize the acoustic space and enable acoustic sequence modeling. (An alternative acoustic tokenization approach using GMM-generated events is described in [37].) Unconstrained-phone-based speaker models capture an assortment of speaker-dependent factors—including spectral characteristics, pronunciation idiosyncrasies, and lexical preferences—and can therefore be difficult to interpret. The basic approach obtains the top phone decoding hypothesis and then evaluates likelihood ratios of speaker-specific and generic (background) phone N-gram models [18]. Results can be improved by running several language-dependent or gender-dependent phone recognizers. The phone N-gram distributions are modeled by bigram or trigram language models; refinements of this approach include the use of decision tree models for better smoothing [21] and the modeling of phone N-grams representing simultaneous outputs from multiple phone recognizers [22].

An important advance was the use of SVMs instead of likelihood models to model phone N-gram frequencies [19]. Improvements can also be obtained by modeling not just the top hypothesized phone sequence from the recognizer, but rather the expected phone N-gram frequencies extracted from phone recognition lattices [20]. In [23], lattice-based phone N-gram frequency modeling is combined with word conditioning. This approach is thus analogous to that used for the word-conditioned cepstral models discussed earlier. The phone N-grams occurring in specific words and frequent phrases are tallied and assembled into a more detailed feature vector that is modeled by SVMs.

A unique combination of phone- and word-based modeling is described in [24,38]. The output of an unconstrained phone recognizer is time-aligned with the phone sequence from a word recognizer, and the conditional probabilities of the former given the latter are modeled. Thus, this model captures phone-specific pronunciation realizations, albeit averaged over all words. An interesting variant aligns hypothesized articulatory features with the unconstrained phone recognition sequence [25].

Approaches based on unconstrained phone recognition show about 2 to 3 times the EER of the best cepstral systems, but can provide substantial gains when combined with them. Reported results show EER reductions of about 25% for 1-side and 44% for 8-side training [20]. Recent experiments with word-constrained phone N-gram methods also give promising results [23] for 8-side training. How systems such as [20] combine when multiple cepstral systems are available is less clear, since preliminary work did not find large gains, but further research is warranted.

2.3 Prosodic Features

Prosodic approaches attempt to capture speaker-specific variation in intonation, timing, and loudness. Because such features are suprasegmental (are not properties of single speech segments but extend over syllables and longer regions), they can provide complementary information to systems based on frame-level or phonetic features. One of the most studied features is speech fundamental frequency (or as perceived, pitch), which reflects vocal fold vibration rate and is affected by various physical properties of the speaker's vocal folds, including their size, mass, and stiffness [39]. Distributions of frame-level pitch values have been used in a number of studies [5,6,7,8]. Although they convey useful information about a speaker's distribution of pitch values, such statistics do not capture dynamic information about pitch contours and are thus not viewed as high-level here.

Dynamic variation in pitch operates at longer temporal spans and is used to convey not only message content (e.g., syntactic units, semantic focus) but also paralinguistic information. Modeling of prosody dynamics (which captures longer-range information and is thus included as higher-level) was used in early work on text-dependent speaker recognition [40]. In [26], a method is described for contour modeling for text-independent recognition. The speaker's pitch movements are modeled by fitting a piecewise linear model to the pitch track to obtain a stylized pitch contour. Parameters of the stylized model are then used as statistical features for speaker verification. Variants are described in [27], which looks at rises and falls of the fitted pitch and energy values based on [26] and models the symbol sequence as a simple bigram. Additional integrated information includes rise and fall durations and phone context. Related work looks at piecewise linear fitting of pitch to help recover from performance degradations in speaker recognition for low-bit-rate coded speech [41]. Two interesting alternatives to using piecewise linear approximations of prosodic contours are proposed in [42], which uses wavelet analysis, and [43], which models sequences of quantized prosody symbols using latent semantic indexing.

An approach to prosodic modeling that is loosely analogous to the whole-word and DTW word modeling methods described for cepstral features is also explored in [27]. In this case, frequent words are matched for F0 contour, rather than for cepstral features. Thus, like its cepstral counterpart, this approach uses no linguistic information for feature extraction but conditions on word-level information from ASR.

A small number of studies have looked at linguistically conditioned duration, pitch, and energy statistics in longer spans of speech. In [28], prosody statistics are computed for units between pauses. The interpause unit is but one example of a larger world of features that could be defined at different temporal spans; the focus is on modeling approaches and modifying GMMs to cope with undefined or inherently missing features (such as pitch, which is missing during unvoiced regions). In [29,38], statistics are computed over an entire conversation side, and distances of each conversation-level feature vector from vectors for target versus impostor speakers are compared using log likelihood ratios. Earlier work on conversation-level statistics [44] includes lexical features such as disfluency rates. Finally, [29,38] explore sequential modeling of "turn"-level prosodic feature statistics. Because turns were automatically inferred from pause and speaker change information, they bear some similarity to the interpause extraction units used in [28], although features and models differ.

Two approaches that use ASR for *conditioning* (as opposed to merely for extraction) are described in [30]. One method, the phone-in-word-duration GMM, models the durations of phones within specific words. Unlike the previous prosodic approaches, it employs ASR for conditioning because it compares durations on a per-word basis. A second method, the state-in-phone-duration GMM, uses the durations (numbers of frames) of the three states in phone HMMs as features, and phones are used for conditioning. In each case, the durations for different positions form a feature vector and are modeled in the adapted-GMM framework used for standard cepstral GMM systems.

A recent method models syllable-based prosodic feature sequences [31,32]. In contrast to interpause-based and conversation-level prosody statistics, this approach uses smaller time units (resulting in more features) and models sequential information. Syllables are automatically inferred from ASR output, and a variety of F0, duration, and energy values are extracted from each syllable. In the unconstrained model, features are extracted for all syllable N-grams in a conversation side. To turn the variable-length sequences of feature vectors into a single conversation-level vector, a set of GMM models is created for each feature sequence (sequence of syllables and pauses). Given a sample, the posterior probabilities of each Gaussian in each GMM are computed and concatenated into the final conversation-level feature vector. These features are provided to an SVM to perform regression on the class labels.

A further refinement is a conditioned version of the syllable-based prosodic feature sequence SVM just described. In this approach, detailed in [33], lexical, part-of-speech, and pause information is used to condition extraction of the same features to specific locations believed to behave similarly prosodically. The goal of the conditioning is thus conceptually similar to that for word-constrained cepstral features [10], but for prosodic rather than phonetic similarity. Note that in the case of prosodic features, phonetic content can be normalized out, allowing multiple words (such as lists of backchannels) per wordlist, increasing robustness. Interestingly, although the unconditioned and conditioned systems use the same features and differ only in conditioning, there is a considerable gain by combining them at the feature level in a single SVM.

Prosodic systems comprise a wide range of approaches and results, making it difficult to summarize performance. The best-performing individual system appears to be a feature-level combination of the unconditioned and conditioned syllable-based prosodic sequence model. Combination of this prosodic system with a cepstral system reduces EER by about 20% and 40% for 1- and 8-side training, respectively. An advantage of this system is that it offers significant complementary information when multiple cepstral systems are present (see Section 3).

2.4 Lexical Features

A speaker's distribution of word sequences is historically one of the earliest types of higher-level features explored for speaker recognition, with roots in the analogous task of author attribution in the text classification domain. Early work using lexical N-gram statistics to discriminate speakers is described in [45]. The approach did not produce a significant gain at the time, presumably because of the brief training and test samples used in task definitions at the time. With the advent of the extended data condition, however, it was found that rates of idiosyncratic word N-grams (for example, "how shall") could be used to help discriminate speakers [34]. The study in [34] used likelihood ratios; in [35], the relative frequencies of frequent word unigrams, bigrams, and trigrams are obtained and assembled into a feature vector that is modeled by SVMs.

More recently, the approach has been extended to encode the duration (slow/fast) of frequent word types as part of the N-gram frequencies [36]. This technique represents a true *hybrid* model of lexical and prosodic features, since it explicitly models both N-gram frequencies and word durations. It thereby simultaneously captures lexical, pronunciation, and prosodic characteristics of the speaker. An interesting further line of research in this area is to postprocess lexical features with latent semantic analysis, so that by grouping words similar in semantic space, one may increase the robustness of estimates for less-frequent words [46].

In terms of performance results, word N-gram modeling yields about 25% EER on 1-side and 10% EER on 8-side training for recent NIST SRE data. Despite the poor performance when used individually, combination with a state-of-the-art cepstral system on recent SRE data improves the overall system by about 15-20% for 8-side training.

Table 2. Data Sets Used in Experiments

Test set	SRE-06 Common Condition	
Training	1-side	8-sides
Conversation sides	3,209	6,556
Models	517	483
Trials	24,013	17,547

3 Performance in a Recent System

The preceding section provided an overview of higher-level features, with minimal discussion of performance because of difficulties in comparing across studies. It also did not address the important question of how systems combine with others, beyond simple comparisons with a baseline cepstral system. To this end, it is useful to look at performance on a recent corpus and task. A set of useful results is available from SRI International, which has in-house efforts to develop both systems based on frame-level cepstra and systems using high-level features.

3.1 Task and Data

We will look at the task of speaker verification on the 2006 NIST SRE evaluation data [4,47]. Results are for the primary subset (the "Common Condition"), which consists of English-only conversations. Test data consist of 1 conversation side. Because high-level features are defined at a longer time scale than are frame-level features, it is interesting to ask how high-level systems perform as a function of the amount of training data per speaker. We will thus look at two training conditions: one with 1 conversation side per speaker, the other with 8 conversation sides per speaker (each with a different conversational partner). Data set statistics are provided in Table 2.

Background training data consisted of 1,553 conversation sides from separate data collections (Switchboard-II and Fisher). Background data did not share any speakers with the data in the test set.

3.2 ASR system

All speech was processed by SRI's speech recognition system. None of the test or background data were used in training or tuning of the recognition system. The system is a fast, two-stage version of SRI's conversational telephone speech (CTS) system, as originally developed for the 2003 DARPA Rich Transcription evaluation [48] and later modified for the NIST 2004 speaker recognition evaluation [35]. It performs a first decoding using Mel frequency cepstral coefficient (MFCC) acoustic models and a bigram language model (LM), generating lattices that are then rescored with a higher-order LM. The resulting hypotheses are used to adapt a second set of models based on perceptual linear prediction (PLP) acoustic features. The adapted models are used in a second decoding pass that is constrained by trigram lattices, which generates N-best lists. These are then rescored by a 4-gram LM and by prosodic models to arrive at the final word hypotheses.

3.3 Session Variability Compensation and TNORM

The SRI system employs techniques for reducing the effect of within-speaker variability associated with the speaking context or environment, rather than the speaker. In the speaker verification community, techniques are often referred to as "session variability" compensation techniques, because they were applied to handle the variability found when the same talker speaks in different conversations. As such, the techniques may capture differences in handsets, background noise, topic of conversation, emotion, speaker health, and so on. The idea is to estimate from data the feature space directions along which intersession variability lies, and then project the features onto the remaining directions. The techniques used are factor analysis for GMM-based models [49] and nuisance attribute projection (NAP) for SVM-based models [50]. An interesting aspect of these approaches is that although they were developed for systems based on cepstral features, they also significantly benefit the SRI prosodic SVM system, with error reductions of over 20% for the 8-side condition. Various systems also make use of TNORM [51], a score normalization technique.

3.4 Systems

While the set of SRI systems does not cover all system types reviewed in Section 2, it has the advantage of including five higher-level and three lower-level systems based on frame-level cepstral features. Where applied, systems used the same ASR output and similar methods for session variability compensation [49,50] and score normalization (TNORM [51]). Systems are roughly, albeit not directly, comparable.[1]

Higher-level systems. The higher-level systems represent five approaches from Table 1: (1) the MLLR system based on word recognition (Section 2.1), (2) a combination of constrained and unconstrained syllable-based prosodic feature sequences in a single SVM (Section 2.3), (3) the word-constrained phone duration system (Section 2.3), (4) the phone-constrained state duration system (Section 2.3), and (5) the duration-conditioned word N-gram (Section 2.4). No phonetic system is represented, because earlier work showed little gain from combining such systems with multiple frame-level cepstral systems. This issue should certainly be revisited, however, given the many updates to various approaches since that time. Another missing feature type is text-conditioned cepstral systems, which is obviously important to explore as well.

Frame-level cepstral systems. In addition to the MLLR system, three other systems model frame-level cepstral features: a cepstral GMM, a cepstral SVM, and a Gaussian

[1] For practical reasons, TNORM was applied for the cepstral SVM, duration, and word N-gram systems, and session variability compensation was applied for the cepstral GMM, MLLR, Gaussian supervector, and prosodic sequence systems. Although the latter technique generally produces larger gains, direct comparisons of systems without normalizations indicate that the ordering of systems by individual performance does not depend on normalization.

supervector SVM. The cepstral GMM system is a generative model of the cepstral feature distribution in the form of a mixture of Gaussian densities [1]. It is trained on a large background set of speakers to cover the entire observed distribution of cepstral features. Frames are treated as unordered, independent samples, discarding longer-term sequence information. Given target speaker training data, the GMM is then adapted by reestimating the Gaussian means on the target speaker data (with a mixture of the background data for smoothing). This results in the target speaker GMM. The system computes the likelihood ratio that the test sample was generated by the target speaker model versus the background model, and accepts the sample if the score exceeds an empirically set threshold.

The cepstral SVM system computes polynomials of the cepstral features and averages them over the entire conversation [52]. For example, one feature might be the average product of the first cepstral coefficient times the square of the second. A feature vector consisting of a large number of these polynomial features characterizes the joint distribution of cepstral features. These feature vectors are then modeled by SVMs. SVMs are trained using a large population of diverse (background) speakers as negative samples and a small set of target speaker instances as positive samples. In testing, a feature vector extracted from the test data is classified by the SVM, and the signed distance from the decision hyperplace is used as a score to be thresholded.

The Gaussian supervector SVM is based on the adapted target speaker GMM mentioned above [53]. Instead of modeling the cepstral features directly, it uses the adapted Gaussian means as features, stringing them together into a long "supervector." The supervector is then modeled as an SVM classifier input, similar to the cepstral SVM.

3.5 Results

Performance results for individual systems are summarized in Table 3. As expected from the review in the previous section, systems based on frame-level cepstral or cepstral-derived features show higher accuracy than longer-range systems. Within the set of cepstral-based systems, the MLLR system has best performance, presumably because it takes advantage of linguistic information from ASR. Of the longer-range systems, the conditioned syllable-based prosody sequence system is the most successful, with less than half the error rate of other longer-range systems for the 8-side condition.

As noted earlier, however, the importance of higher-level systems for such tasks is not individual performance but how well they complement standard systems. To answer that question, we examine results for various system combinations. Individual system scores are combined using an SVM with a linear inner product kernel; the combiner is trained using scores for separate data (from the NIST 2005 evaluation). We first look at how well each system combines with the cepstral GMM system. Results for 1- and 8-side training are shown in Figure 1, respectively. For reference, the cepstral GMM system alone and the MLLR system (best single system) alone are also indicated.

As shown, all combinations (triangle and square symbols) improve performance over the baseline cepstral GMM alone, in most cases by a significant degree. For both training conditions, combinations with other frame-level cepstral systems (squares) are better than combinations with some higher-level systems (triangles) and worse than others. The MLLR system alone performs better than the combination of the cepstral GMM

Table 3. Individual System Results for Eight Systems. "(H)" denotes higher-level systems. Performance is given as both equal error rate (EER) and the detection cost function (DCF) used by NIST [4].

System	Time	ASR	EER/DCFx10	EER/DCFx10
(Feature Type and Model)	Range		1-side	8-side
Cepstral GMM	frame	no	4.75 / 0.216	2.79 / 0.107
Cepstral SVM	frame	no	5.07 / 0.242	2.33 / 0.093
Gaussian supervector SVM	frame	no	4.15 / 0.198	3.24 / 0.164
(H) MLLR SVM	frame	yes	4.00 / 0.197	2.14 / 0.073
(H) State-in-phone-duration GMM	longer	yes	16.02 / 0.705	8.07 / 0.423
(H) Phone-in-word-duration GMM	longer	yes	22.22 / 0.874	9.30 / 0.420
(H) Syllable-based prosody sequence SVM	longer	yes	10.41 / 0.461	3.74 / 0.162
(H) Duration-conditioned word N-gram SVM	longer	yes	23.46 / 0.815	9.95 / 0.446

with either the cepstral SVM or the supervector system with 8 sides of training data, but this is not the case for the 1-side training condition—demonstrating that higher-level systems add more value as training data increases. For the 8-side condition, the best two combinations with the cepstral GMM are clearly the MLLR system and the prosodic feature sequence system, both higher-level systems. In this condition even the word N-gram system, which performs poorly on its own, combines about as well with the baseline as does the cepstral SVM system.

To understand how more than two systems combine, we can look to Figure 2. This figure shows which systems are selected when one optimizes an N-way system combination for best performance (in DCF terms). Since there are 8 SRI systems, N ranges from 1 to 8. If the selection pattern is monotonic—i.e., if the list of systems for N+1 includes all systems from the list for N—then the order in which systems are progressively added can be construed as reflecting system importance in the combination. As shown in Figure 2, with one exception, the selection order is monotonic, thus providing information about which systems make the most contribution to the overall result.

We can extract a number of useful observations by comparing Table 3 and Figure 2. As already mentioned, by themselves noncepstral systems perform less well than cepstral systems. Among the systems using higher-level information, the more acoustic information a system models, the better it tends to perform on its own (MLLR-SVM > prosody sequence model > duration > word-duration N-grams), which is not surprising. The MLLR-SVM system, which takes advantage of both high-level constraints and frame-level acoustic information, is also the best single system overall.

What is striking is the finding that of the four systems using frame-level cepstral features (MLLR, cepstral GMM, cepstral SVM, and supervector SVM), only two are actually useful for a particular training condition (1 or 8 sides). Within each condition, only two such systems appear at the left side of the figures; the other two appear at the

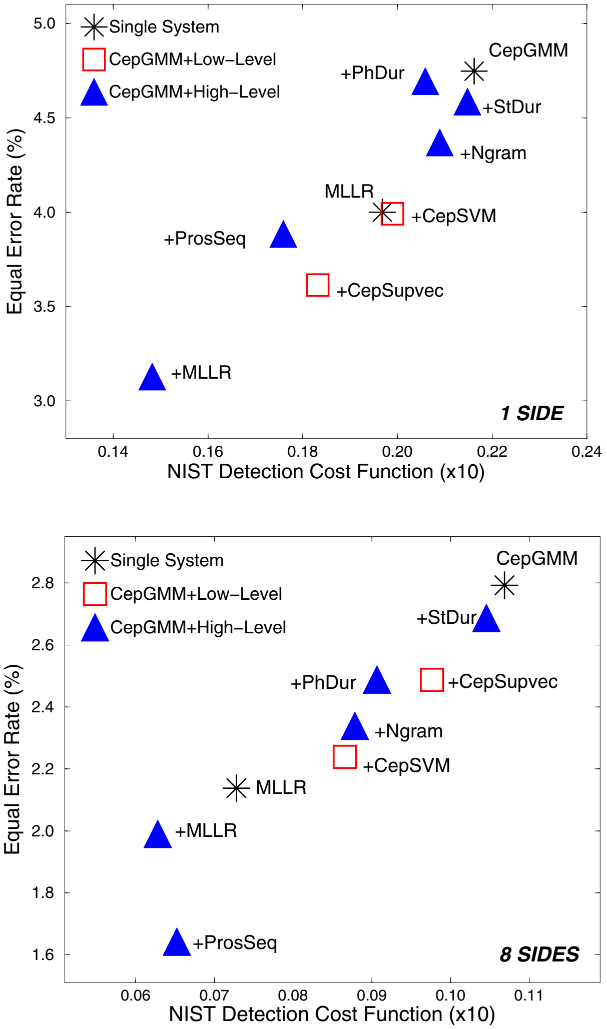


Fig. 1. Two-way combination results (system + cepstral GMM) by type of system. CepGMM = cepstral GMM, CepSVM = cepstral SVM, CepSupvec = Gaussian supervector SVM, MLLR = MLLR SVM, ProsSeq = syllable-based prosodic feature sequence SVM, PhDur = phone-in-word duration GMM, StDur = state-in-phone duration GMM, Ngram = duration-conditioned word N-gram SVM.

right and do not add any real performance improvements. In fact, they can even degrade performance (as can any system added late in the combination) because of overfitting in combiner training. Note that different cepstral systems are useful for different amounts of training data. Complementary information comes from higher-level

	MLLR SVM (H)	Supervector SVM	Prosody Sequence SVM (H)	Phone-in-Word Duration GMM (H)	State-in-Phone Duration GMM (H)	Word+Duration N-gram SVM (H)	Cepstral GMM	Cepstral SVM	%EER	DCF (x10)
1 Best	🟥								4.00	0.197
2 Best	🟥						🟥		3.13	0.148
3 Best	🟥	🟥	🟥						2.86	0.139
4 Best	🟥	🟥	🟥	🟥					2.86	0.137
5 Best	🟥	🟥	🟥	🟥	🟥				2.80	0.136
6 Best	🟥	🟥	🟥	🟥	🟥	🟥			2.86	0.140
7 Best	🟥	🟥	🟥	🟥	🟥	🟥	🟥		2.64	0.141
8 Best	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	2.59	0.144

	MLLR SVM (H)	Prosody Sequence SVM (H)	Cepstral SVM	Word+Duration N-gram SVM (H)	State-in-Phone Duration GMM (H)	Phone-in-Word Duration GMM (H)	Supervector SVM	Cepstral GMM	%EER	DCF (x10)
1 Best	🟥								2.13	0.0728
2 Best	🟥	🟥							1.74	0.0561
3 Best	🟥	🟥	🟥						1.59	0.0501
4 Best	🟥	🟥	🟥	🟥					1.59	0.0488
5 Best	🟥	🟥	🟥	🟥	🟥				1.59	0.0478
6 Best	🟥	🟥	🟥	🟥	🟥	🟥			1.54	0.0483
7 Best	🟥	🟥	🟥	🟥	🟥	🟥	🟥		1.64	0.0480
8 Best	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	1.64	0.0485

Fig. 2. Results for the best N-way combination of systems for the 1-side (top) and 8-side (bottom) training conditions. Filled boxes indicate which systems were selected; corresponding performance is given in both equal error rate (EER) and detection cost function (DCF).

systems, including systems that perform modestly when used alone. Such systems become increasingly useful as training data increases from 1 to 8 sides. It would thus be interesting to look at even larger amounts of speaker training data (more than 20 minutes), to see at which point the higher-level models begin to level off in their performance. Overall, these findings point to a nonobvious strategy for future overall system development. Because systems based on cepstral features tend to be highly correlated, the exploration of complementary systems based on higher-level features should become only more important as lower-level systems continue to improve.

4 Conclusions and Implications for Speaker Classification

Despite the dominance of GMM systems based on frame-level cepstral features, we have seen that higher-level features provide significant complementary information for speaker identification. Higher-level features are increasingly useful as training data increases, and we have not yet witnessed the point at which they level off in relative contribution to performance. Furthermore, because certain higher-level features are inherently more invariant to channel and noise characteristics than are spectral features, they offer the possibility of additional robustness for speaker recognition under degraded acoustic conditions.

For the wider area of speaker classification, higher-level features in speech provide far more information about a talker than only his or her identity. For example, research in [54] reveals that speaker age is reflected not only in acoustic features but also in temporal features such as phone durations. Features based on phone-level, lexical, or prosodic information are correlated with language and dialect classification [55], emotion classification [56], deception detection [57], and perceived charisma [58], as well as a host of other health-related, cognitive, and sociolinguistic factors. Given sufficient data labeled for such characteristics, one might apply some of the features and techniques described here, substituting the new class of interest for speaker identity. Since we know that higher-level features are quite successful at classifying individual speakers, an additional interesting research area in classifying speaker characteristics, rather than individual speakers, would be to apply nuisance attribute projection [59] to project out the variability that is speaker-related. In this way, one might achieve sharper models that can assist speaker classification in other domains.

Acknowledgments

The author thanks Christian Müller; SRI colleagues Andreas Stolcke, Sachin Kajarekar, Luciana Ferrer, Kemal Sönmez, Gokhan Tür, and Martin Graciarena; and SRI visitor Robbie Vogt. This work was supported by NSF Contract IIS-0544682. The views herein are those of the author and do not represent the views of the funding agency.

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 181–202 (2000)
2. Sturim, D.E., Campbell, W.M., Reynolds, D.A.: Classification Methods for Speaker Recognition. In: Müller, C. (ed.) Speaker Classification I. LNCS (LNAI), vol. 4343, Springer, Heidelberg (2007)
3. Markowitz, J.: The Many Roles of Speaker Classification in Speaker Verification and Identification. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007)
4. Martin, A.F.: Evaluations of Automatic Speaker Classification Systems. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007)
5. Carey, M., Parris, E., Lloyd-Thomas, H., Bennett, S.: Robust prosodic features for speaker identification. In: Bunnell, H.T., Idsardi, W. (eds.) Proc. ICSLP. Philadelphia, vol. 3, pp. 1800–1803 (1996)

6. Sönmez, M.K., Heck, L., Weintraub, M., Shriberg, E.: A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition. In: Kokkinakis, G., Fakotakis, N., Dermatas, E. (eds.) Proc. EUROSPEECH, Rhodes, Greece, pp. 1391–1394 (1997)
7. Arcienega, M., Drygajlo, A.: Pitch-Dependent GMMs for Text-Independent Speaker Recognition Systems. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01 – Interspeech), Aalborg, Denmark, pp. 2821–2825 (2001)
8. Kinnunen, T., Gonzalez-Hautamaki, R.: Long-Term F0 Modeling for Text-Independent Speaker Recognition. In: Proceedings of the 10th International Conference Speech and Computer (SPECOM), Patras, Greece, pp. 567–570 (2005)
9. Park, A., Hazen, T.J.: ASR Dependent Techniques for Speaker Identification. In: Hansen, J.H.L., Pellom, B. (eds.) Proc. ICSLP, Denver, pp. 1337–1340 (2002)
10. Sturim, D.E., Reynolds, D.A., Dunn, R.B., Quatieri, T.F.: Speaker Verification Using Text-Constrained Gaussian Mixture Models. In: Proc. ICASSP. vol. 1., Orlando, pp. 677–680 (2002)
11. Baker, B., Vogt, R., Sridharan, S.: Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal, pp. 2429–2432 (2005)
12. Gauvain, J.L., Lamel, L.F., Prouts, B.: Experiments with Speaker Verification Over the Telephone. In: Pardo, J.M., Enríquez, E., Ortega, J., Ferreiros, J., Macías, J., Valverde, F.J. (eds.) Proc. EUROSPEECH, Madrid (1995)
13. Newman, M., Gillick, L., Ito, Y., McAllaster, D., Peskin, B.: Speaker Verification Through Large Vocabulary Continuous Speech Recognition. In: Bunnell, H.T., Idsardi, W. (eds.) Proc. ICSLP. vol. 4, Philadelphia, pp. 2419–2422 (1996)
14. Boakye, K., Peskin, B.: Text-Constrained Speaker Recognition on a Text-Independent Task. In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain (2004)
15. Gillick, D., Stafford, S., Peskin, B.: Speaker Detection without Models. In: Proc. ICASSP. Philadelphia, vol. 1, pp. 757–760 (2005)
16. Aronowitz, H., Burshtein, D., Amir, A.: Text Independent Speaker Recognition Using Speaker Dependent Word Spotting. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea, pp. 1789–1792 (2004)
17. Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., Venkataraman, A.: MLLR Transforms as Features in Speaker Recognition. In: Proc. Interspeech, Lisbon, pp. 2425–2428 (2005)
18. Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J., Hernandez-Cordero, J.: Gender-Dependent Phonetic Refraction for Speaker Recognition. In: Proc. ICASSP. Orlando, vol. 1, pp. 149–152 (2002)
19. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A., Leek, T.R.: Phonetic Speaker Recognition with Support Vector Machines. Advances in Neural Information Processing Systems 16, 1377–1384 (2004)
20. Hatch, A.O., Peskin, B., Stolcke, A.: Improved Phonetic Speaker Recognition Using Lattice Decoding. In: Proc. ICASSP. Philadelphia, vol. 1, pp. 169–172 (2005)
21. Navrátil, J., Jin, Q., Andrews, W.D., Campbell, J.P.: Phonetic Speaker Recognition Using Maximum-Likelihood Binary-Decision Tree Models. In: Proc. ICASSP. Hong Kong, vol. 4, pp. 796–799 (2003)
22. Jin, Q., Navrátil, J., Reynolds, D.A., Campbell, J.P., Andrews, W.D., Abramson, J.S.: Combining Cross-Stream and Time Dimension in Phonetic Speaker Recognition. In: Proc. ICASSP. Hong Kong, vol. 4, pp. 800–803 (2003)
23. Lei, H., Mirghafori, N.: Word-Conditioned Phone N-Grams for Speaker Recognition. In: Proc. ICASSP, Honolulu (2007)

24. Klusáček, D., Navrátil, J., Reynolds, D.A., Campbell, J.P.: Conditional Pronunciation Modeling in Speaker Detection. In: Proc. ICASSP. Hong Kong, vol. 4, pp. 804–807 (2003)
25. Ka-Leung, Y., Man-Mak, W., Kung, S.Y.K.: Articulatory Feature-Based Conditional Pronunciation Modeling for Speaker Verification. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea, pp. 2597–2600 (2004)
26. Sönmez, K., Shriberg, E., Heck, L., Weintraub, M.: Modeling Dynamic Prosodic Variation for Speaker Verification. In: Mannell, R.H., Robert-Ribes, J. (eds.) Proc. ICSLP. vol. 7, pp. 3189–3192, Australian Speech Science and Technology Association, Sydney (1998)
27. Adami, A.G., Mihaescu, R., Reynolds, D.A., Godfrey, J.J.: Modeling Prosodic Dynamics for Speaker Recognition. In: Proc. ICASSP. Hong Kong, vol. 4, pp. 788–791 (2003)
28. Kajarekar, S., Ferrer, L., Sönmez, K., Zheng, J., Shriberg, E., Stolcke, A.: Modeling NERFs for Speaker Recognition. In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain, pp. 51–56 (2004)
29. Peskin, B., Navrátil, J., Abramson, J., Jones, D., Klusáček, D., Reynolds, D.A., Xiang, B.: Using Prosodic And Conversational Features for High Performance Speaker Recognition: Report From JHU WS'02. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong, pp. 792–795 (2003)
30. Ferrer, L., Bratt, H., Gadde, V.R.R., Kajarekar, S., Shriberg, E., Sonmez, K., Stolcke, A., Venkataraman, A.: Modeling Duration Patterns for Speaker Recognition. In: Proc. EUROSPEECH, Geneva, pp. 2017–2020 (2003)
31. Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A.: Modeling prosodic feature sequences for speaker recognition. Speech Communication. (Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation) 46(3-4), 455–472 (2005)
32. Ferrer, L., Shriberg, E., Kajarekar, S., Sönmez, K.: Parameterization of Prosodic Feature Distributions for SVM Modeling in Speaker Recognition. In: Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07), Honolulu, Hawaii (2007)
33. Shriberg, E., Ferrer, L.: A Text-Constrained Prosodic System for Speaker Verification. In: Proceedings of Interspeech, Antwerp, Belgium (2007)
34. Doddington, G.: Speaker Recognition Based on Idiolectal Differences Between Speakers. In: Dalsgaard, P., Lindberg, B., Benner, H., Tan, Z. (eds.) Proc. EUROSPEECH, Aalborg, Denmark, pp. 2521–2524 (2001)
35. Kajarekar, S.S., Ferrer, L., Shriberg, E., Sonmez, K., Stolcke, A., Venkataraman, A., Zheng, J.: SRI's 2004, NIST Speaker Recognition Evaluation System. In: Proc. ICASSP. Philadelphia, vol. 1, pp. 173–176 (2005)
36. Tür, G., Shriberg, E., Stolcke, A., Kajarekar, S.: Duration and Pronunciation Conditioned Lexical Modeling for Speaker Verification. In: Proceedings of Interspeech, Antwerp, Belgium (2007)
37. Scheffer, N., Bonastre, J.F.: Speaker Detection using Acoustic Event Sequences. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal (2005)
38. Reynolds, D., Andrews, W., Campbell, J., Navrátil, J., Peskin, B., Adami, A., Jin, Q., Klusáček, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang, B.: The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong (2003)
39. Titze, I.: Principles of Voice Production. Prentice Hall, Englewood Cliffs (1994)
40. Atal, B.: Automatic Speaker Recognition Based on Pitch Contours. Journal of the Acoustical Society of America 52(6), 1687–1697 (1972)

41. Chen, S.H., Wang, H.C.: Improvement of Speaker Recognition by Combining Residual and Prosodic Features with Acoustic Features. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada (2004)
42. Chen, J., Dai, B., Sun, J.: Prosodic Features Based on Wavelet Analysis for Speaker Verification. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal, pp. 3093–3096 (2005)
43. Chen, Z.H., Liao, Y.F.L., Juang, Y.T.: Eigen-Prosody Analysis for Robust Speaker Recognition under Mismatch Handset Environment. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea (2004)
44. Weber, F., Manganaro, L., Peskin, B., Shriberg, E.: Using Prosodic and Lexical Information for Speaker Identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), Orlando, Florida (2002)
45. Heck, L.: Integrating High-Level Information for Robust Speaker Recognition (2002), http://www.clsp.jhu.edu/ws2002/groups/supersid/
46. Nayeeemulla Khan, A., Yegnanarayanaa, B.: Latent Semantic Analysis for Speaker Recognition. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea (2004)
47. Martin, A., Miller, D., Przybocki, M., Campbell, J., Nakasone, H.: Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004. In: Proceedings 4th International Conference on Language Resources and Evaluation, Lisbon, pp. 587–590 (2004)
48. Stolcke, A., Franco, H., Gadde, R., Graciarena, M., Precoda, K., Venkataraman, A., Vergyri, D., Wang, W., Zheng, J., Huang, Y., Peskin, B., Bulyko, I., Ostendorf, M., Kirchhoff, K.: Speech-to-text Research at SRI-ICSI-UW. In: DARPA RT-03 Workshop, Boston (2003)
49. Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Factor Analysis Simplified. In: Proc. ICASSP. vol. 1, pp. 637–640 (2005)
50. Solomonoff, A., Campbell, W.M., Boardman, I.: Advances in Channel Compensation for SVM Speaker Recognition. In: Proc. ICASSP, Philadelphia, vol. 1, pp. 629–632 (2005)
51. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independent Speaker Verification Systems. Digital Signal Processing 10(1-3), 42–54 (2000)
52. Campbell, W.M.: Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. In: Proc. ICASSP, Orlando, vol. 1, pp. 161–164 (2002)
53. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support Vector Machines Using GMM Supervectors for Speaker Verification. IEEE Signal Processing Letters 13(5), 308–311 (2006)
54. Schötz, S., Müller, C.: A Study of Acoustic Correlates of Speaker Age. In: Müller, C. (ed.) Speaker Classification II. LNCS(LNAI), vol. 4441, Springer, Heidelberg (2007)
55. Schultz, T.: Speaker Characteristics. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007)
56. Devillers, L., Vidrascu, L.: Real-life Emotion Recognition in Speech. In: Müller, C. (ed.) Speaker Classification II. LNCS(LNAI), vol. 4441, Springer, Heidelberg (2007)
57. Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., Kajarekar, S.: Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection. In: Proc. ICASSP, vol. 1, pp. 1033–1036 (2006)
58. Rosenberg, A., Hirschberg, J.: Acoustic/Prosodic Correlates of Charismatic Speech. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal (2005)
59. Solomonoff, A., Quillen, C., Boardman, I.: Channel Compensation for SVM Speaker Recognition. In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain (2004)

Enhancing Speaker Discrimination at the Feature Level

Jacques Koreman[1], Dalei Wu[1], and Andrew C. Morris[2]

[1] Department of Language and Communication Studies
Norwegian University of Science and Technology
NO-7491 Trondheim Norway
jacques.koreman@hf.ntnu.no, daleiwu@gmail.com
[2] SpinVox Ltd.
Wethered House, Pound Lane
Marlow, Bucks, SL7 2AF
United Kingdom
andrew.morris@spinvox.com

Abstract. This chapter describes a method for enhancing the differences between speaker classes at the feature level (feature enhancement) in an automatic speaker recognition system. The original Mel-frequency cepstral coefficient (MFCC) space is projected onto a new feature space by a neural network trained on a subset of speakers which is representative for the whole target population. The new feature space better discriminates between the target classes (speakers) than the original feature space. The chapter focuses on the method for selecting a representative subset of speakers, comparing several approaches to speaker selection. The effect of feature enhancement is tested both for clean and various noisy speech types to evaluate its applicability under practical conditions. It is shown that the proposed method leads to a substantial improvement in speaker recognition performance. The method can also be applied to other automatic speaker classification tasks.

Keywords: Feature enhancement, Speaker basis, Kullback-Leibler distance, GMM, Speech in noise.

1 Introduction

In accordance with the aim of this book, we shall present results from experiments on automatic speaker classification. The experiments investigate the possibility of enhancing the discrimination between target classes by appropriate pre-processing. The target classes in our experiments are the speakers themselves, but the approach is also suitable for automatic classification tasks in which speakers are grouped according to more general characteristics, for instance age classes, gender, language or dialect, among others (cf. several other contributions to this issue).

The spectral variation observed in speech is naturally mainly determined by its linguistic content (message), the communication of which is the main function of speech. The different phones (in context) therefore mainly determine the distribution of the speech in the acoustic space. Nevertheless, the specific distributions observed

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 260–277, 2007.

for individual speakers are not the same. This is due to anatomical and physiological differences between speakers, which set a bound to their acoustic possibilities, but also to differences in learnt behaviour, for instance caused by the speaker's articulatory setting [1] or his/her dialect, which determine the habitual use of the vocal apparatus [2]. The distribution of linguistic and speaker variation in the acoustic space is presented very schematically in Figure 1, where different fill colours represent different phones and the colours of the circle borders represent different speakers (only two speakers are shown for clarity). The figure shows that, despite their variation, the phones (fill colours) are separated fairly well, despite overlap for instance due to the different contexts in which they occur, while there is substantial overlap between different speakers' realisations (border colours) of the same phone despite inter-speaker differences in voice quality.

In speech recognition the variation due to different speakers is not relevant to the aim, and can be dealt with by speaker normalisation techniques such as vocal tract length normalisation. But for speaker recognition, it is exactly these variations in the production of phones which must be relied on to distinguish between speakers – in addition, of course, to prosodic differences, which are not taken into consideration here. It is all the more surprising that the same features are normally used in both speech and speaker recognition. In this chapter, we shall use Mel-frequency cepstral coefficients (MFCCs) to recognize speakers. Instead of normalising for speaker differences as is necessary for speech recognition, we shall *enhance* the differences between speakers, in accordance with the aim of speaker recognition. From Figure 1 it is clear that a non-linear mapping is required to obtain a better separation of the target classes. A multi-layer perceptron (MLP) is optimally suited for this task.

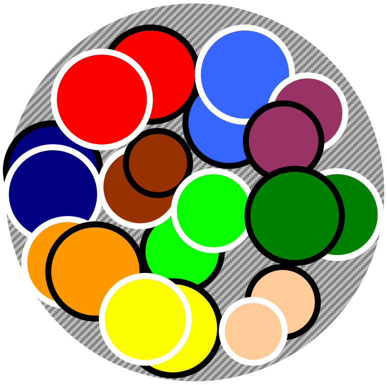


Fig. 1. Schematic representation of the distribution of phones (fill colours) and speakers (border colours) in the acoustic space

The speaker identification problem addressed here is that of large speaker sets. We shall show that better speaker discrimination can be obtained even if only a subset of the speakers is used for feature enhancement. It is important to note that the feature enhancement method proposed here does not need retraining when new speakers are added. This makes it useful for speaker verification, too, where the aim is not to select the most likely speaker from a given set (as in speaker identification or classification), but to evaluate an identity claim, deciding whether the speech signal stems from the claimed client or from an impostor, as for instance used in access systems for buildings or computers. We expect, therefore, that the method can also be applied to verification problems.

The method proposed here is based on that of [3,4]. They showed that by deriving features from a multi-layer perceptron (MLP) trained on 31 speakers in the NIST database who had all used multiple handsets, speaker verification could be improved, in some cases on the basis of the new, enhanced features on their own, in other cases only when these were combined (feature-level fusion) with the original MFCC features. While the aim in this previous work was to deal with variability caused by the use of different handsets by normalisation, the aim of this chapter is to obtain a better discrimination between speakers per se (Section 4). The method will also be applied to noisy data to see whether it can still enhance speaker discrimination when different kinds of background noise (Section 5) or channel characteristics (Section 6) affect the speech signals, as is often the case in normal conditions. The chapter finishes with a discussion in Section 7 and conclusions in Section 8. But before we go into feature enhancement we shall describe the data used in the experiments (Section 2) and the baseline speaker identification system (Section 3).

2 Data

Because the primary aim of our experiments is to investigate whether feature enhancement can increase the discrimination between speakers per se in the acoustic space, rather than to evaluate its ability to deal with noisy data (which is the secondary aim), we selected a clean-speech database for our speaker identification experiments. The TIMIT speech database, which was collected under optimal conditions in a recording studio [5], is particularly suitable for our aims, although we were forced to downsample the database from 16 to 8 kHz (TIMIT-8k), because speaker identification in our baseline system is 100% correct on the original, 16 kHz data, and therefore not suitable for investigating improvements due to other methods. Using 8 kHz signals is also more suitable for comparison to the effects of channel noise using the NTIMIT database (below).

For many practical applications, however, noisy data are the norm. To allow for generalisation to such conditions we also add different noise types to the database to evaluate its effect on feature enhancement. The noise types added to the speech signal consist of (stationary) car as well as (non-stationary) factory-1 noise and babble from the NOISEX-92 database [6]. They were added to the TIMIT-8k database at SNRs of 20, 10 and 0 dB, using the ITU software [7] to determine SNRs. These noises are also used in the Aurora evaluations for speech recognition in noise. Of course, considering only the effect of specific, single noises – although this is a widespread convention – is a relatively simplistic approach, and not really comparable to an actual noisy environment, in which many noises can be present simultaneously and can vary during a recording. Furthermore, the simple addition of noise leaves the speech unchanged, whereas in real noisy conditions the Lombard effect causes the speaker to adapt his/her speech to the noisy environment. Nevertheless, the experiments with noisy speech can show whether the feature enhancement method proposed here stands up to noisy data.

To investigate the effect of channel noise on feature enhancement, experiments are also carried out on the NTIMIT database [8]. NTIMIT was collected by transmitting the original TIMIT recordings through a telephone handset and over various telephone channels, half of them using long-distance carriers, and then redigitizing them. As with the noisy speech data, the aim of the NTIMIT experiments is to see whether our feature enhancement approach can be useful for practical speech signals like telephone speech.

Since the standard TIMIT division does not include a development set, we created our own division into speaker-disjoint training, development and evaluation data, with 300, 168 and 162 speakers, respectively. The three sets are selected such that gender and the eight dialect regions represented in TIMIT have an equal proportional representation in the three sets. TIMIT consists of three sentence types: two SA sentences intended to expose dialectal differences between speakers (spoken by all speakers); 450 phonetically compact SX sentences, intended to give good coverage of all pairs of phones (five sentences per speaker); and 1890 phonetically diverse SI sentences with phones in varying allophonic contexts (three sentences per speaker). To make the speaker identification system text-independent, we used all sentences of type SA_{1-2}, SI_{1-2} and SX_{1-2} for training, sentences of type SX_3 and SI_3 for development and sentences of type SX_4 and SX_5 for evaluation. Whereas SA_1 and SA_2 sentences are always the same for different speakers, SI_n sentences are always different ones and the index *n* only indicates the order in which the sentences were spoken by each speaker as indicated by the numbers in the TIMIT database; each SX_n sentence is spoken by seven speakers. This strict division optimises text-independence of the speaker recognition system.

Using 20-ms frames and a 10-ms step size, 20 Mel-scaled filterbank log power features were extracted from the speech signals, using a Hamming window and a pre-emphasis factor of 0.97. A discreet cosine transform (DCT) was then applied to obtain Mel-frequency cepstral coefficients (MFCCs), from which the c0 energy coefficient was dropped. Time difference features were not appended, because these did not improve performance with TIMIT-8k. Neither silence removal nor cepstral mean subtraction were used, since none of these led to any performance improvement with TIMIT-8k. The probable reason for this is that for clean speech silence removal may also lead to the deselection of low-energy speech sounds which can help to distinguish speakers, while cepstral mean subtraction not only subtracts noise, but also the average speaker characteristics, making the speakers more similar. Unlike for more noisy speech conditions, the advantages of these methods do not weigh up to the disadvantages in the case of clean speech.

3 Baseline Speaker Identification System

The baseline speaker identification system only differs from that used in later sections in the features it uses as input. Whereas the systems used in later sections are trained with features derived from the MFCCs to enhance speaker discrimination, the baseline system uses the MFCCs directly. (For a general description of the features and statistical models which can be used for speaker recognition, see [9].)

The MFCCs for the six training utterances from each speaker are used as input to Gaussian mixture modelling (GMM) of the diagonal covariances using 32 Gaussians [10,11]. With TIMIT-8k (though not with other databases, such as the CSLU speaker recognition database) no gain is found in training speaker models by maximum a-posteriori (MAP) adaptation from a universal background model (UBM) for all 300 training speakers, so that each speaker model was trained from scratch with data for that speaker only.

As in [11], GMMs are trained by *k*-means clustering, followed by expectation maximisation (EM) iteration. This is performed by the Torch machine learning API [12], using a variance threshold factor of 0.01 and minimum Gaussian weight of 0.05 (performance falling sharply if either was halved or doubled). These parameter choices were determined on the basis of the development sentences of the 168 development speakers.

Test results are obtained for 162 test speakers (for two test sentences per speaker, cf. Section 2). Speaker identification for utterance feature data X is performed by selecting the speaker S_j with the largest posterior probability, $P(S_j|X)$ (which corresponds here to the largest data likelihood $p(X| S_j)$, as all speaker priors $P(S_j)$ are equal). With a speaker identification accuracy of 96.60%, the baseline model trained with MFCCs of the six training utterances gives state-of-the-art speaker recognition performance.

4 Feature Enhancement

For feature enhancement, a multi-layer perceptron (MLP) is used with the aim of obtaining better speaker discrimination. An MLP is suitable for this because its training objective is to optimize *separation* between the target classes. It is important to note explicitly that the MLP is not used for classification itself (as it often is), but only for feature enhancement in the preprocessing stage, before the features are fed into the speaker classifier, which is based on GMM (as explained in the previous section).

For speech recognition, the target classes are usually phones. The feature projection is obtained from the pre-squashed outputs of the MLP, which is trained to output a posterior probability for each phone. For speech recognition, a simple MLP with just one hidden layer can provide a feature projection which gives an improvement of the speech recognizer's performance [13].

It is not possible to apply an MLP in the same way to speaker recognition. One reason is that in speaker recognition there are no fixed target classes like phonemes in ASR. Particularly for speaker recognition tasks where a large number of speakers must be identified (or verified), the large number of target classes makes training (convergence) of the MLP difficult, particularly when – as is often the case – there is relatively little training data for each speaker. To counteract this effect, the MLP is trained with a representative subset of speakers *(speaker basis)* as its target classes, comparable to phones as target classes for ASR. It will be shown that the transformation which the MLP learns for the speaker basis is also beneficial for other, unseen speakers from the same population for clean speech.

But an MLP with only a single hidden layer is not able to make the complex mapping (cf. Figure 1) from the speech signal (MFFCs) to speakers. Unlike for ASR, therefore, a one-layer MLP applied to speaker recognition with clean-speech TIMIT-8k data does not lead to an increase in the percentage of correct speaker identifications (as demonstrated in [14]). This may be because speaker data, being clustered around every phoneme, is less easy to partition than speech data. But the separating power of an MLP can be increased by using more hidden layers. In [3,4], an MLP with three hidden layers was trained to recognise 31 speakers, and the internal representation in the outputs from the central, linear bottleneck hidden layer (also called compression layer) were used as discriminative features. The 31 speakers were selected because they had been recorded over multiple handsets. It was found that the features obtained from the MLP's bottleneck layer provide a performance enhancement, although not consistently across all training and test conditions [4] and sometimes only when the feature vector was concatenated with the original MFCC features which were used as input to the MLP [3]. The good results may be due to a better compensation for the different handsets that were used or to a better separation of the speakers in the acoustic space, even if the speakers were not selected with this aim – or a combination of the two.

The same MLP as in [3] was used here. Its structure is shown in Figure 2. Each node in the MLP has a two-stage function. The first stage, the net input function, is a many-to-one linear combination of the neuron's inputs. The second stage is a one-to-one non-linear sigmoid function which squashes the net-input to a value between zero and one.

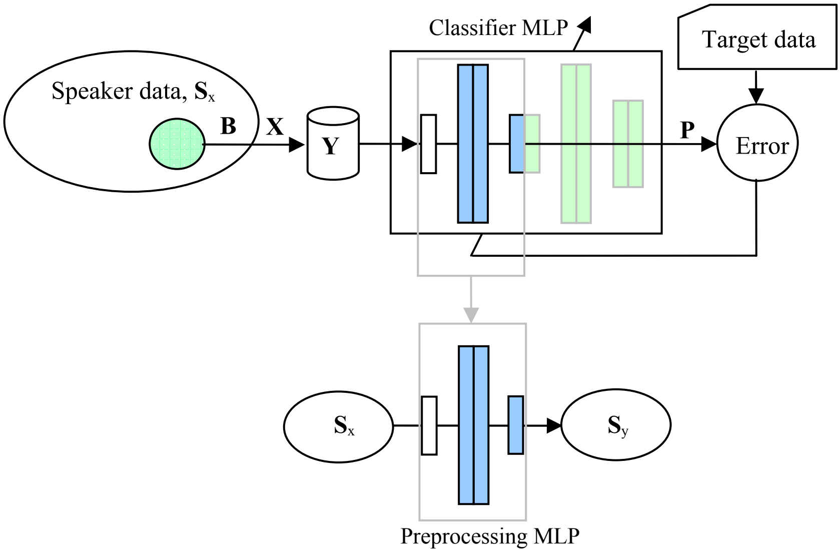


Fig. 2. Feature enhancement procedure

From the point of view of using the MLP-internal feature representation to provide discriminative features, the squashed outputs are not very suitable because they tend to be close to zero or one, thereby not complying with the GMM assumption that all features have an approximately Gaussian distribution. Therefore the net input to the second hidden layer was used as input to GMM modelling.

The MLP was implemented in Torch [12]. Each single frame of the standard MFCC features is preprocessed by a 5-layer MLP. This MLP was found to outperform MLPs consisting of fewer layers [14,15]. Training the MLP with single frames instead of the usual input vector of 9 concatenated frames gives the best results for this particular database. The MLP is trained, by gradient descent, to maximise the cross-entropy objective (i.e. the mutual information between the actual and target outputs). Training was performed in batch mode, with a fixed learning rate of 0.01. The data in each utterance was first normalised to have zero mean and unit variance (z-scores). The MLP was trained with a fixed number of iterations (35), after which the error reduction on the training and development data in the MLP frame-based recognition was very small. Of the 3 hidden layers, the first and last hidden layer, which are both non-linear, have 100 units and the middle, linear hidden layer has 19 (bottleneck or compression layer). The features obtained from the compression layer were used as input to the GMM system described in Section 3. The assumption behind this is that this simple representation, which consists of vectors of the same size as the original MFCC vectors, is an internal representation of the acoustic signal which enhances discrimination between the target speakers and can be generalized to the speakers in the entire population.

In the two following subsections, we shall first investigate the effect of the size of the speaker basis and then present different methods for selecting the speaker basis.

4.1 Size of the Speaker Basis

The feature enhancement proposed here does not only work for fixed sets, but also for incremental sets, i.e. when speakers are added to the speaker identification system. For practical reasons, the MLP and the models for already enrolled speakers should not have to be changed when a new speaker is added. To be able to deal effectively with incremental speaker sets, it is therefore important that the small speaker basis used to train the MLP provides a feature projection which also leads to a better separation of other, unseen speakers. Here, we make random selections of speaker bases with different sizes from the 300 training speakers to train the MLP. The selections are obtained for basis sizes of 30, 50, 60, 70, 100 and 150 speakers. Each speaker basis set of a given size is randomly selected from the same group of 300 training speakers with replacement, i.e. every time a speaker basis is extracted from the training set, they are put back for the second independent random selection. The enhanced features are then used as input to the GMM speaker identification system, as described above. The speaker identification results for three different random selections at each speaker basis size are reported in Table 1, together with the mean and standard deviation across the three randomisations.

Table 1. Speaker identification accuracy (%) for three different random speaker basis selections of sizes varying from 30-150 speakers, with means and standard deviations across the three selections

speaker basis selection	30	50	60	75	100	150
1	96.60	96.91	97.53	98.15	97.22	99.07
2	97.53	96.60	95.99	98.15	97.53	98.15
3	97.22	96.91	98.15	96.91	95.99	98.46
mean	97.12	96.81	97.22	97.74	96.91	98.56
sd	0.47	0.18	1.11	0.72	0.81	0.47

Generally a modest improvement in speaker accuracy can be observed for enhanced features compared to the baseline of 96.60% – only for the second random selection of 60 speakers and the third random selection of 100 speakers is speaker identification accuracy lower than the baseline. Even feature enhancement based on a very small speaker basis of 30 speakers can improve identification of 162 “unseen” test speakers for clean speech. Similar results were found for a larger number of disjunct random selections of the speaker basis from the training speakers for a slightly different division of the TIMIT speakers into training, development and test sets [14].

As in [14], the results here show that the effect of feature enhancement depends on the particular speaker set selected for training the MLP. The variance for the different speaker bases at each given size is quite large compared to the change in accuracy, so that the particular random selection of the speaker basis substantially influences system performance. This indicates how important it is to find a reliable method for speaker basis selection. Two methods will be discussed in the following section, the first one knowledge-based and the second automatic.

4.2 Improving Speaker Basis Selection

Since the results for MLP feature enhancement can vary depending on the selected speakers, as shown in the previous section, we investigate two different approaches to optimize the speaker basis selection. The first approach is knowledge-constrained speaker selection, the second is deterministic.

4.2.1 Knowledge-Constrained Random Speaker Basis Selection

For the TIMIT database on which the experiments reported here are carried out, several speaker properties are known beforehand. Age, height, race, education level, gender and dialect region of the training speakers are known. Some of these properties can be expected to affect the speaker’s voice. Since gender and dialect region are recognisable from the filenames and are likely to cause a large part of the speaker variation, this prior knowledge can be exploited for speaker basis selection. The division of gender and dialect region is not entirely balanced in the database. TIMIT contains speech from 438 male and 192 female speakers. Of the eight dialect regions, speakers from dialect regions 2 (Northern), 3 (North Midland), 4 (South Midland), 5 (Southern) and 7 (Western) are overrepresented compared to the other regions (1=New England, 6=New York City, 8=Army Brat). We therefore selected several speaker basis sets for training the MLP by proportionally balancing gender and dialect region. Speakers within each gender/dialect region group were selected randomly. The aim of this method of speaker basis selection is to use prior knowledge as much as possible. Table 2 shows the results for different non-overlapping speaker bases. This shows that on average the knowledge-constrained selection gives very similar results to random selection.

Table 2. Speaker identification accuracy (%) for non-overlapping, proportionally balanced knowledge-based speaker basis selections of sizes varying from 30-150 speakers, with mean and standard deviation across the selections

speaker basis selection	30	50	60	75	100	150
1	96.91	97.12	98.15	97.53	97.84	98.77
2	96.30	97.84	97.84	95.99	97.53	97.53
3	95.68	97.53	96.91	98.77	97.53	
4	95.68	97.53	95.68	96.60		
5	96.60	96.91	95.68			
6	96.30	96.30				
7	96.60					
8	94.14					
9	95.68					
10	96.60					
mean	96.05	97.12	96.85	97.12	97.63	98.15
sd	0.80	0.55	1.16	1.21	1.18	0.88

For small speaker basis sets (30 speakers), the speaker identification accuracy is mostly lower than for the baseline GMM system (96.60%). When we compare this with the random selection results in Table 1, this is somewhat surprising, since there the speaker identification accuracy is mostly higher than or the same as the baseline. For larger speaker basis sets, the identification accuracy is generally higher than the baseline.

Of course, knowledge-constrained speaker basis selection is only possible if the database is labelled with the relevant properties related to the main sources of variation in the speech signals. Despite the controlled representation of speakers in each of the speaker basis sets, there is still considerable variation in the test results. We are forced to conclude, therefore, that gender and dialect region still leave a lot of the variation in the speech signal unaccounted for. Other variables play an important role for the discrimination between speakers.

Furthermore, test results on the evaluation set (not shown here) show that there is also very little correlation between development and evaluation scores when the random or knowledge-constrained speaker basis selection is used. This means that even if we train an MLP for feature enhancement which gives optimal results for the development speaker set, there is no guarantee that that MLP will also lead to optimal results for a set of unseen test speakers. A method is therefore required which will guarantee improved performance with speakers unknown to the MLP.

4.2.2 Deterministic Speaker Basis Selection

For many databases no speaker information is available which can be used to deal with speaker variation, and it would normally be impracticable to label the speakers with this information (but note that some of this information can be obtained automatically fairly reliably). In any case, as the results in Table 2 show, different speaker bases using knowledge-constrained speaker basis selection can still lead to quite variable speaker identification results. Furthermore, as noted above, evaluation performance cannot be predicted from performance on the development set.

All the above reasons call for an automatic approach to speaker basis selection. In this section, we present such a method. It selects those speakers who differ most from all other speakers. The speaker selection is based on the confusion matrix for the baseline GMM speaker identification task, which shows the log likelihoods for each of the 2 *development* sentences for the *training* speakers (300 speakers). Several other methods are described in [16].

Before the automatic speaker basis selection is carried out, the log likelihoods matrix is converted into a matrix of probabilities by first converting log likelihoods to likelihoods and then dividing each row by the row sum. For any test utterance $X=\{x_t\}$, $t=1\ldots n$, the likelihood for speaker S_j is

$$p(X \mid S_j) = \sum_{t=1}^{n}\sum_{i=1}^{M} w_i \cdot N(x_t, \mu_i, \Sigma_i^{-1}) \tag{1}$$

where M is the number of Gaussians in the speaker GMM. The posterior probability for this speaker is then obtained as follows:

$$p(S_j \mid X) = \frac{p(X \mid S_j)p(S_j)}{p(X)} = \frac{p(X \mid S_j)p(S_j)}{\sum_{k=1}^{N} p(X \mid S_k)p(S_k)} \tag{2}$$

where N is the number of speakers. If we assume the prior probabilities $P(S_j)$ are the same for any speaker model, then we obtain the posterior probability given any test utterance:

$$P(S_j \mid X) = \frac{p(X \mid S_j)}{\sum_{k=1}^{N} p(X \mid S_k)} \tag{3}$$

The table of posterior probabilities, with one probability $P(S_j|X_t)$ for each speaker S_j and test utterance X_t, can now be used to select basis speakers.

This method selects those speakers whose pdf's are as far apart as possible from the pdf's of every other speaker. Let $KL(S_j,S_k)$ denote the symmetric Kullback-Leibler distance [17] between two pdfs.

$$KL(S_j,S_k) = \int \left(p(X|S_j) - p(X|S_k)\right) . \ln\left(\frac{p(X|S_j)}{p(X|S_k)}\right) dX \tag{4}$$

When *p(X*|*S)* is modelled by a GMM, (4) cannot be evaluated in closed form. However, using the fact that speaker priors are equal in our test set and applying Bayes' rule to (4) we can proceed as follows.

$$KL(S_j,S_k) \propto \int p(X)\left(P(S_j \mid X) - P(S_k \mid X)\right) . \ln\left(\frac{P(S_j \mid X)}{P(S_k \mid X)}\right) dX \tag{5}$$

We can rewrite (5) as

$$KL(S_j,S_k) \propto \int p(X) K(S_j,S_k|X) dX = E[K(S_j,S_k|X)] \tag{6}$$

We can obtain an approximation to the expected value in (6) (the distance between any two speaker models) by summing $K(S_j,S_j|X_t)$ over all utterances in the development test set.

$$AK(S_j,S_k) = \sum_{t=1}^{S} K(S_j,S_k \mid X_t) \tag{7}$$

where *S* is the number of utterances in the test data. The distance matrix *AK* is symmetric, with $K(S_j,S_j) = 0$. Further, define the sum of the distances from speaker S_j to every other speaker as $SK(S_j)$:

$$SK(S_j) = \sum_{k=1}^{N} AK(S_j,S_k) \tag{8}$$

where *N* is the number of speakers. Speakers are then selected in order of decreasing $SK(S_j)$, i.e. decreasing average approximated Kullback-Leibler distance from all other speakers.

The speaker identification accuracies of GMM experiments using a feature projection obtained from an MLP trained with the deterministically selected, maximally variable speaker basis using the approach just described are shown in Table 3.

Table 3. Speaker identification accuracy (%) for TIMIT data, using KL speaker basis selection, against speaker basis selections of sizes varying from 30-150 speakers

speaker basis selection	30	50	60	75	100	150
KL	97.12	98.15	97.84	97.84	97.84	98.15

The results in Table 3 show that deterministic speaker basis selection consistently leads to an improvement over the baseline performance (baseline identification accuracy: 96.60%). The size of the improvement is not very different from the *average* result for the random or knowledge-based speaker basis selections. Random and knowledge-based speaker basis selections, however, do not always improve speaker identification accuracy on the test speaker set. Deterministic speaker selection has the important advantage that it guarantees an optimal selection of the speaker basis.

Best results use only 50 speakers to train the MLP. Of course, with 96.60% even the baseline results are very high for the TIMIT-8k data. A speaker identification accuracy of 98.15% is equal to an error reduction of 1.55% points, or a relative error reduction[1] of 46%. But looking at the differences in terms of the number of utterances for which the speaker was not correctly identified, we must realize that the difference is fairly small: the system trained with MLP-enhanced features (for an MLP trained with 50 speakers selected from the training speakers on the basis of the KL distance) misidentifies the speaker in only 6 out the 324 test utterances. Compared to 11 errors in the baseline system, this shows that the margins are minimal. We shall take this issue up again in the discussion.

5 Feature Enhancement for Channel Noise

For many practical applications of speaker recognition it is of interest to evaluate our automatic feature enhancement method for telephone speech, in which channel noise disturbs the speech signal. In [3,4], the MLP which we use for feature enhancement was used to enhance speaker differences for 31 NIST speakers who used different telephone handsets, and in which channel noise was therefore also present. The experiments carried out here concentrate on the speaker discrimination improvement when the handsets are unknown, and show whether feature enhancement can alleviate the effect of channel noise on speaker identification. Comparing the results for feature enhancement using the KL distance on clean-speech, but downsampled TIMIT data (Section 4.2) to those from identical experiments using telephone speech from NTIMIT, we find that the performance of the baseline system using MFCC features drops from 96.60% to 58.95%.

[1] Although the general results in the Tables are shown as percentage of accurate speaker identification, the improvement of one system over another in terms of the change in error is more indicative of the aim of the special preprocessing (with the optimal improvement being a 100% relative error reduction). The relative error is computed as the absolute improvement (or error reduction) divided by the error percentage of the baseline system (which is 100% – percentage accuracy).

Table 4. Speaker identification accuracy (%) for NTIMIT data, using KL speaker basis selection, against speaker basis selections of sizes varying from 30-150 speakers

speaker basis selection	30	50	60	75	100	150
KL	55.56	58.02	57.10	59.88	63.89	59.57

Given a sufficiently large speaker basis, MLP feature-enhancement can improve speaker identification. As Table 4 shows, performance is higher than the baseline for speaker basis sizes of 75 and larger. It may be concluded from this that small speaker bases cannot represent the distributional characteristics of the speech with varying channel characteristics. Best performance is found for a speaker basis size of 100 speakers selected using the KL distance measure. Possibly the selection of a larger subset is not optimal because it includes speakers which are more "average", so that they do not help to make the features more discriminative. The absolute error reduction for a speaker basis of 100 speakers is 4.94% points, which is equal to a drop in relative error of 12%.

6 Feature Enhancement for Added Noise

Gaussian mixture models (GMMs) for speaker recognition can achieve very good performance in clean speech, but performance normally degrades strongly in the presence of noise [18]. The effect of different types and levels of added noise is investigated in this section, using the same GMM system architecture as before.

Since MLPs have been shown to be able to deal well with noisy speech in ASR [13], we expect they may also enhance speaker recognition in the presence of additive noise. The same MLP architecture used for feature enhancement in the previous sections is now applied to enhance speaker discrimination in noise – but note that the confusion matrix for the training speakers is always determined for the noisy data, so that the basis speakers selected are representative of the particular training condition, and may be different in each noise condition. To limit the number of experiments, a fixed speaker basis size of 150 speakers was chosen. The reason for choosing a relatively large speaker basis is that the speech signal are more variable in the different noise conditions, so that a larger speaker basis may be needed to effectively reflect this variation. The speaker identification results can be compared for the matched noise conditions with and without feature enhancement (Section 6.1). But in many practical applications, there will be a mismatch between training and test conditions. The reason for this is that it is often practically impossible for speakers to enrol under the same variety of conditions that must be dealt with when the system is operative, if only because the test conditions often cannot be anticipated – but also because there is a pressure to keep the enrolment sessions short for the sake of user-friendliness. In this case, enrolment may take place in fairly clean speech conditions, while the test conditions may vary, causing a mismatch between training and test data (Section 6.2). We not only investigate the effect of a mismatch, where the training data is clean and the test data contain noise, we also evaluate the effect of simply adding several additive noise types to the training data (multi-condition training) to deal with the presence of different possible noise types in the test data (Section 6.3).

6.1 Matched Noise Conditions

In this section, speaker identification results are compared for different noise types and at different SNRs. The MLP-enhanced features are compared with the baseline system in which the MFCCs are not preprocessed by the MLP and used as input to GMM directly. Table 5 shows the results for clean data, and for car, factory and babble noise at SNRs of 20, 10 and 0 dB. The results shown in Table 5 are for matched noise conditions, in which the speakers' test data were compared with speaker models trained on data of the same noise type and at the same SNR. This condition represents the results that can be obtained when a perfect noise condition detector can be used to select appropriate noise models during the recognition stage.

Table 5. Speaker identification accuracy (%) for training on matched noise type and level

	clean	car			factory			babble			mean
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	
MFCC	96.60	88.58	81.79	61.11	81.17	56.17	12.65	82.10	63.89	12.35	63.64
MLP	98.15	93.52	87.04	73.15	84.26	59.57	17.28	89.20	75.00	18.83	69.60

Conform expectations, the best speaker identification results in Table 5 are found for clean speech. Also as expected, the results show a strong drop in speaker recognition accuracy with decreasing SNR. The absolute reduction is greatest for the stationary car noise, particularly at SNR=0 dB (absolute error reduction: 12.04% points; relative error reduction: 31%), as well as for babble at higher SNRs, especially 10 dB (absolute error reduction: 11.11% points; relative error reduction: 31%). The positive effect, though present in all conditions, is smaller for factory noise, although speaker identification is still well above chance level. Notice that no cepstral mean subtraction (CMS) was performed, even for the noisy data. In unpublished experiments on the same TIMIT data with added noise, in which speaker identification performance for MFCCs with and without CMS was compared with performance on the basis of wavelets for the same noise conditions as presented here (matched noise conditions), the use of CMS always led to a decrease in speaker identification performance. The only reasonable explanation for this finding is that subtraction of the spectral mean across an utterance also filters out part of the speaker characteristics.

Overall we can conclude that preprocessing of the MFCCs by an MLP to enhance speaker discrimination *always* reduces the speaker identification error. The average reduction of the speaker identification error is 5.96% points, which equals a relative error reduction of 16%.

6.2 Mismatched Noise Conditions

In many applications, noise conditions may vary, but the condition in which the speaker must be recognized is not known beforehand or cannot be detected reliably. Two scenarios are possible. In the first scenario, the speaker enrolled in the system in a quiet environment, so that the speaker model (and the MLP) is trained on clean speech. But the actual conditions in which the system is subsequently used may vary from one occasion to the next. In order to evaluate the performance of our system under these *mismatched* conditions, the test data from all previously used noise conditions were scored with the GMM speaker models (and MLP) trained with clean speech only. (The data for clean speech are the same as in Table 5 and does not represent a mismatch. The speaker identification performance is only included in Table 6, because it is used to compare the mean percentage error for correct speaker identification across all possible test conditions, in the right-hand column, to the results in Tables 5 and 7.)

Table 6. Speaker identification accuracy (%) for training on clean speech and testing in various clean and noisy conditions

	clean	car			factory			babble			mean
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	
MFCC	96.60	48.77	27.78	6.48	16.05	1.85	0.93	25.62	11.11	2.47	23.77
MLP	98.15	20.37	11.73	3.40	4.63	1.54	0.62	8.95	5.56	2.16	15.71

As the results in Table 6 show, the mismatch between the noisy test data and the clean training data causes a severe deterioration of the performance of the speaker recognition system. For some of the conditions, recognition accuracy is only just above chance level (p=100/162=0.62%). The effect, which is present for the MFCC features (comparison of first data rows in Tables 1 and 2), is even greater after MLP enhancement, with only chance level speaker identification accuracy for factory noise at 0 dB SNR. The average effect of feature enhancement for mismatched data is an increase in the speaker identification error of 8.06% points, or an 11% increase in relative error.

6.3 Multi-condition Training

In the case of known, but variable additive noise, the noises can easily be used to create “virtual” data containing this noise before the speaker model is trained. By training speaker models across a variety of noise conditions in the training phase, the system is expected to better cope with the variability in real operation conditions. As the results in Table 7 show, the performance in all noisy conditions is substantially better than when the GMM speaker models (and the MLP) are trained on clean speech only (cf. Table 6).

Table 7. Speaker identification accuracy (%) for training and testing across all noise conditions

	clean	car			factory			babble			mean
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	
MFCC	85.80	87.04	80.25	43.52	80.25	63.89	14.81	82.10	74.38	25.31	63.73
MLP	85.49	87.65	88.89	66.98	87.35	66.98	16.98	85.49	81.48	29.63	69.69

In some cases, the speaker identification accuracy is even higher than in the matched noise condition in Table 5, e.g. for factory noise at 10 dB SNR and for babble at 10 and 0 dB SNR, thus showing the advantageous effect of multi-condition training. In all conditions except for clean speech is the speaker identification accuracy higher when MLP-enhanced features are used compared to the baseline system using MFCCs. The average error drop is 5.94% points, which is equal to a relative error drop of 16%. The greatest improvement is found for car noise at an SNR of 0 dB, where the absolute error drop is 23.46% points or 42%.

7 Conclusions and Discussion

The aim of this chapter was to show how differences between speakers can be enhanced, given the strong natural overlap between them in the acoustic space. A multi-layer perceptron can be trained to transform the MFCC features derived directly from the speech signals into a feature representation in a space in which the speakers can be better discriminated. In order for the MLP to successfully learn to discriminate the strongly overlapping MFCC speaker spaces, it is important that an MLP with sufficient layers is used [14]. If the number of target classes is too large to train an MLP with (as is the case when speakers are the target classes), the new features can be obtained by learning the space transformation on the basis of a subset of the training speaker called the speaker basis. In Section 4 it was shown that automatic speaker basis selection based on the Kullback-Leibler distance between speakers in the confusion matrix for the speakers leads to the selection of a speaker basis which is representative for the population. The feature enhancement leads to a relative reduction of the error of 46%.

In Section 5 it was shown that, for the same data, the method leads to a 12% error reduction for telephone speech. In Section 6, it was shown that feature enhancement leads to an average error reduction of 16%, or maximum 31%, for added noise, when the noise in the test condition is the same as during training. Because in many practical applications this cannot be guaranteed, tests were also carried out for mismatched training-test conditions, in which case the MLP feature transformation leads to an increase in the error rate. It is clear why this should be so: the feature enhancement learnt for clean speech is not appropriate for the noisy feature space. To better deal with varying test conditions, it is possible to train multi-condition models. In this case the average error rate reduction is 16%, with a maximum of 42% in the tested conditions. Only for clean-speech test data do we find a small increase in the error rate compared to the un-enhanced MFCC features when multi-condition speaker models are used. It is noteworthy that in some cases the performance for the test data is better when multi-condition training is used even compared to matched training conditions. What cannot be concluded from the experiments presented is how suitable multi-condition speaker models are for unseen noise conditions, and thus how robust the feature enhancement is in real applications.

The improvement in speaker discrimination due to feature enhancement varies for different speech conditions. Comparing the mean log likelihoods of the correct speakers in the confusion matrix for downsampled clean speech (Section 4.2.2), a one-sided *t*-test for matched pairs shows that there is no significant difference between MLP-enhanced features (using a speaker basis of 150 speakers) and MFCCs. For speech containing channel noise (Section 5), a significant difference is obtained ($p << 0.01$). The t-tests for matched pairs, however, only show whether the enhanced features obtained from the MLP have higher log likelihoods than those obtained on the basis of MFCCs. This does not automatically translate into a higher system performance, which is measured in terms of correct speaker identifications. For this, MacNemars tests [19] were carried out. These tests showed a non-significant difference in the number of correct speaker identifications for TIMIT, but note that for this data, the correct speaker identification accuracy for the baseline system using MFCC's as input to the GMM was already very high (96.60%). For NTIMIT, on the other hand, a significant improvement was found when MLP-enhanced features are used instead of MFCCs ($p < 0.05$). This test shows that the actual performance for NTIMIT is better when MLP-enhanced features are used than when speakers are identified using the original MFCC feature space. Although we did not carry out these tests for all conditions, we expect most of the improvements for speaker identification in added noise in Tables 5 and 7 to also be very significant, given that the absolute error reduction is mostly much larger than that for the tested NTIMIT data (0.62%, cf. Table 3 and text).

Feature enhancement on the basis of an MLP has been shown to improve automatic speaker identification. It is expected that the method is also useful for enhanced discrimination between speaker classes, such as gender or age groups, or language and dialect. When the number of target classes is large, as it is in speaker identification, the Kullback-Leibler distance can be used to select an optimal subset of the speakers (representative of the classes) to train the MLP. The applicability of the method is therefore wider than to speaker identification alone. It can also be applied to speaker verification, i.e. to accept or reject the claimed identity of a speaker.

Acknowledgments

This work was supported by the SecurePhone project IST-2002-506883 (www.secure-phone.info) funded by the EU under the FP6 IST programme. It was carried out at Saarland University, Saarbrücken (Germany), where all three authors worked previously.

References

1. Laver, J.: The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge (1980)
2. Dellwo, V., Huckvale, M., Ashby, M.: How is individuality expressed in voice? An introduction to speech production & description for speaker classification. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol 4343 Springer, Heidelberg (2007) (this issue)
3. Konig, Y., Heck, L., Weintraub, M., Sonmez, K.: Nonlinear discriminant feature extraction for robust text-independent speaker recognition. In: Proceedings of RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications, Avignon, France, pp. 72–75 (1998)
4. Heck, L., Konig, Y., Kemal Sönmez, M., Weintraub, M.: Robustness to telephone handset distortion in speaker recognition by discriminative feature design. Speech Communication 31, 181–192 (2000)
5. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium (1993)
6. Varga, A., Steeneken, H.J.M.: Assesment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12, 247–252 (1993)
7. ITU recommendation P.56: Objective measurement of active speech level (March 1993)
8. Fisher, W.M., Doddington, G.R., Goudie-Mashall, K.M., Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J.: NTIMIT. Linguistic Data Consortium (1993)
9. Sturim, D.E., Campbell, W.M., Reynolds, D.A.: Classification methods for speaker recognition. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol 4343 Springer, Heidelberg (2007) (this issue)
10. Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17, 91–108 (1995)
11. Reynolds, D.A., Zissman, M.A., Quatieri, T.F., O'Leary, G.C., Carlson, B.A.: The effect of telephone transmission degradations on speaker recognition performance. In: Proceedings ICASSP'95, Detroit, Michigan, pp. 329–332 (1995)
12. Collobert, R., Bengio, S., Mariéthoz, J.: Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46 (2002)
13. Sharma, S., Ellis, D., Kajarekar, S., Jain, P., Hermansky, H.: Feature extraction using non-linear transformation for robust speech recognition on the Aurora database. In: Proceedings ICASSP2000, Istanbul, Turkey, pp. 1117–1120 (2000)
14. Wu, D., Morris, A.C., Koreman, J.: MLP internal representation as discriminative features for improved speaker recognition. In: Faundez-Zanuy, M., Janer, L., Esposito, A., Satue-Villar, A., Roure, J., Espinosa-Duro, V. (eds.) NOLISP 2005 LNCS, vol 3817 pp. 72–80 (2005)
15. Morris, A.C., Wu, D., Koreman, J.: MLP trained to separate problem speakers provides improved features for speaker identification. In: Proceedings IEEE Int. Carnahan Conf. on Security Technology (ICCST2005), Las Palmas, Spain (2005)
16. Wu, D.: Discriminative Preprocessing of Speech: Towards Improving Biometric Identification. Ph.D. thesis Saarland University, Saarbrücken, Germany (2007)
17. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 2nd edn. Elsevier Academic Press, Amsterdam (2003)
18. Reynolds, D.A.: Experimental evaluation of features for robust speaker identification. IEEE Transactions on Speech and Audio Processing 2, 639–643 (1994)
19. Gillick, L., Cox, S.J.: Some statistical issues in the comparison of speech recognition algorithms. In: Proceedings ICASSP, Glasgow, pp. 532–535 (1989)

Classification Methods for Speaker Recognition[*]

D.E. Sturim, W.M. Campbell, and D.A. Reynolds

Massachusetts Institute of Technology
Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA
{sturim, wcampbell, dar}@ll.mit.edu

Abstract. Automatic speaker recognition systems have a foundation built on ideas and techniques from the areas of speech science for speaker characterization, pattern recognition and engineering. In this chapter we provide an overview of the features, models, and classifiers derived from these areas that are the basis for modern automatic speaker recognition systems. We describe the components of state-of-the-art automatic speaker recognition systems, discuss application considerations and provide a brief survey of accuracy for different tasks.

1 Introduction

The development of automatic speaker recognition systems is one example in the field of speech processing that brings together the areas of speech science for speaker characteristization, pattern recognition and engineering. From speech science comes the insights into how humans produce and perceive speaker-dependent information in the speech signal as well as signal processing techniques for analyzing acoustic correlates conveying this information. The area of pattern recognition provides algorithms for effectively modeling and comparing speaker characteristics from salient features. Finally, engineering is used to both realize working systems based on the above ideas and to handle real-world variability that arise in applications. In this chapter we provide an overview of the features, models, and classifiers derived from these areas that are the basis for modern automatic speaker recognition systems.

In Figure 1, we show the basic framework and components of speaker recognition systems. We are using the general term of speaker recognition to encompass the underlying tasks of speaker identification (which one of a set of speakers is talking?) and speaker detection or verification (is this particular speaker talking?). We will note throughout this chapter when particular comments refer

[*] This work was sponsored by the Department of Justice under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 278–297, 2007.

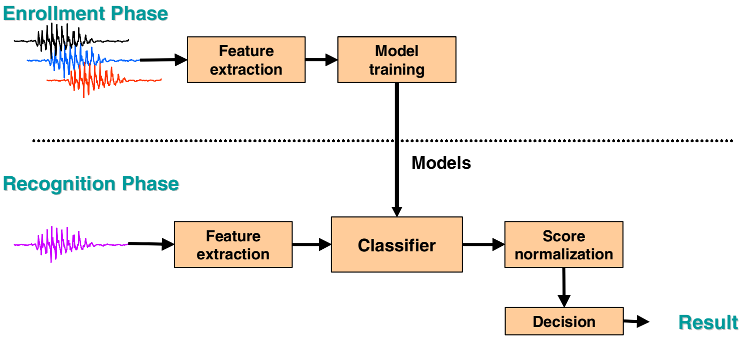


Fig. 1. Structure of a speaker recognition system

to identification or detection. As with any pattern recognition system, speaker recognition systems consist of two distinct phases: enrollment (also called training) and recognition (also called testing).

The first step, common to both enrollment and recognition phases, is the extraction and conditioning of a set of features from the input signal believed to convey information about the speaker. In Section 2, we review some of the commonly used methods for feature extraction.

Features from speech samples by a speaker are used in the enrollment phase to build or train parameters for a model which represents the specific characteristics of that speaker. During the recognition phase, features from the test speech sample are compared to one or more of the speaker models, depending on the task, by the classifier to produce match scores. In Section 3, we review the most successful models and classifiers found in automatic speaker recognition systems.

These scores are optionally normalized to add robustness or to map them to a desired dynamic range (e.g., 0 to 1). This and other forms of normalization and compensation are discussed in Section 3.6.

Finally, the decision component either compares the score to a threshold to decide to accept or reject, in the case of speaker detection, or reports out the highest scoring model, in the case of speaker identification. The decision could also compare the score of the highest scoring model to a threshold and decide to report "none-of-the-above." This is a merger of speaker identification and detection known as *open-set* identification.

2 Feature Extraction

Feature processing for speaker recognition systems consists of extracting speaker dependent information in a form which can be effectively and efficiently used for model building and recognition. Broadly speaking, features used for speaker recognition can be categorized by three key attributes:

- Temporal span
- Discrete vs. continuous values
- Information level

The attributes of features will impact the models and classifiers that are appropriate to use.

The information in speech signals occurs at several different time spans and rates. Thus, features used to capture this information also occur with different time spans and rates. Features that aim to capture information about a person's vocal tract information as seen through the frequency spectrum of speech, will operate using short-time spans (∼20-30ms) so as to analyze quasi-stationary snapshots of the vocal apparatus. Prosodic information, such as a person's average pitch inflection per sentence, is an example of a feature derived by looking at a longer time span (∼1-2 s). Further, the feature time span and rate may be variable, for example, when examining aperiodic, variable duration events like speech pathologies, phonemes, or words.

The value of the speech measurements used in the features can be discrete or continuous. Features consisting of speech frequency spectrum samples are an example of continuous valued measurements. Features counting the number of occurrences of events in speech, such as word usage counts, are an example of discrete values measurements. There is, of course, a continuum between continuous and discrete measurements since one can quantize continuous values for efficiency or use a probability of occurrence that is < 1.0 when counting events.

The third attribute is the information level features represent. Speech conveys many levels of information, from semantic meaning, via the words spoken, to the speaker's physical vocal apparatus, via the acoustic sound of the speech (i.e., bass vs treble). Speaker recognition features can be focused to capture speaker dependent characteristics from these different levels. Features aimed at low-level information tend to extract measurements about the acoustic characteristics related to vocal production, such as frequency spectrum or short time pitch estimates. Features aimed at higher-level information, such as pronunciations and word usage (idiolect), require the output of some other speech recognition tool such as a phone or word recognition system.

We pictorially depict this feature attribute space in Figure 2. Typically, features related to high-level speaker information consist of longer time span, variable rate analysis of discrete events, such as phones or words. Features related to low-level speaker information consist of short time span, fixed rate analysis of continuous phenomenon, such as spectra. We next review some common features used in automatic speaker recognition systems indicating their attributes. Figure 3 shows where these features lie in the attribute space.

Mel Frequency Cepstral Coefficients (MFCCs) [1,2]: MFCCs are the most commonly used features in modern speaker recognition systems[3]. MFCC temporal processing uses a fixed analysis window on the order of ∼20 millisecond. MFCCs are represented by a real valued N-dimensional vector. The coefficients are a parameterization of the spectrum which have some dependency on the

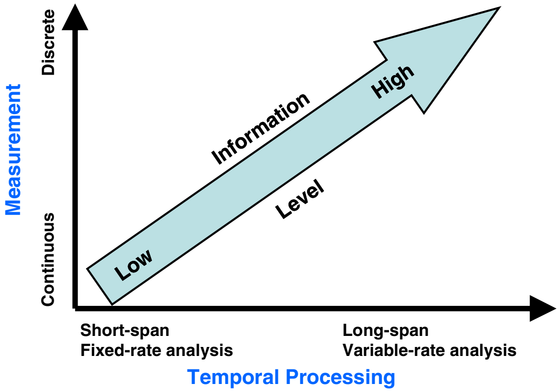


Fig. 2. Relation of attributes for features used in automatic speaker recognition systems

physical characteristics of the speaker. MFCCs are considered to be low-level information.

Linear Prediction-based Cepstral Coefficients (LPCCs) [4,2]: LPCCs are often used in speaker recognition systems, although their susceptibility to noisy environments have made them more undesirable as speaker recognition systems are applied to more challenging channels. Like MFCCs, the LPCC processing uses a fixed analysis window (∼20 millisecond) and are of the continuous measurement type. LPCCs are dependent on the spectral envelope and are considered to be low-level information.

Codebook quantized spectral entries [5]: These features measure the approximate location of the spectrum in acoustic space. Rather than use the continuous representation of cepstral features, the features can be quantized either using a VQ codebook or a Gaussian mixture model (GMM). The feature in this case is the index in the corresponding VQ codebook or the mixture index in the GMM.

Pitch and Energy [6]: The goal is to learn pitch and energy gestures by modeling the joint slope dynamics of pitch and energy. When these features are combined with a short phrases, the analysis window will be variable spanning the duration of the short phrase.

Prosodic Statistics [7]: Are based on various measurements of energy, duration and pitch derived over large speech segment. The goal is to capture the prosodic idiosyncrasies of individual speakers. The feature type will be continuous since the prosodic statistical measures are reported in continuous values. The level of information is considered low-middle since these features are measuring prosodic inflections and patterns.

Word and Phone Tokenization [8,9,10,11,12]: These are a more recent addition to feature sets used in speaker recognition systems. The analysis window is variable, since it is based on the expected duration of the word or phone units.

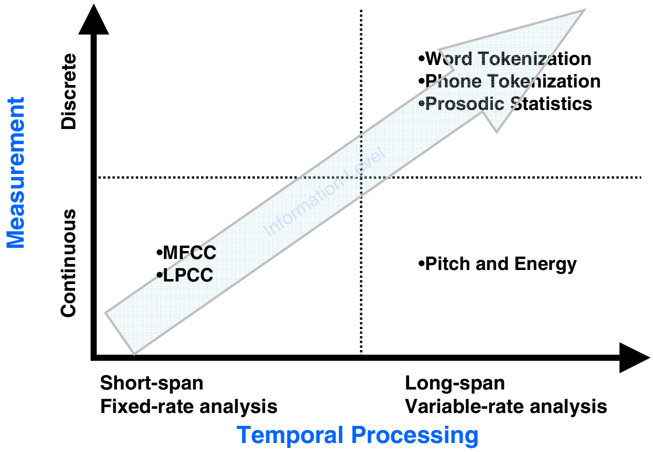


Fig. 3. Approximate location of common feature in the feature attribute space

Further counts of word pairs or triples cover longer time spans. Since counts of discrete words and phone are often used as features, the value type would be discrete. Word and phone models in speaker recognition both try to represent the pronunciation differences of talkers and are considered high-level information.

3 Models and Classifiers

Speaker models and classifiers are tied not only to the features used, but also to the task being addressed. The two tasks of speaker recognition are 1) speaker identification and 2) speaker verification. The speaker identification task is closed-set recognition, where all of the talkers that will be seen by the system are pre-enrolled and known. Figure 4 shows the general structure of a speaker identification system. The applications of closed-set identification are limited since most real-world scenarios must usually handle out-of-set speakers. Performance is a function of the number of speaker in the identification set and the speech used.

The speaker verification task, in contrast, is a binary decision of whether the unknown speaker is the same as the hypothesized (or claimed) speaker. While ostensibly an easier task than classifying among a set on N speakers, verification must potentially be able to effectively reject the open-set of speakers that could act as impostors. This open-set is usually dealt with by using some general impostor model. The general structure of the speaker verification system is presented in Figure 5. Speakers verification addresses a more general problem and has wider application in the speaker recognition community, so it is a more common focus for classifier design and evaluation.

For both the identification and verification structure, there are many types of models and classifiers that have been used. We will mainly focus on those aimed at solving the more general open-set verification task (although they are

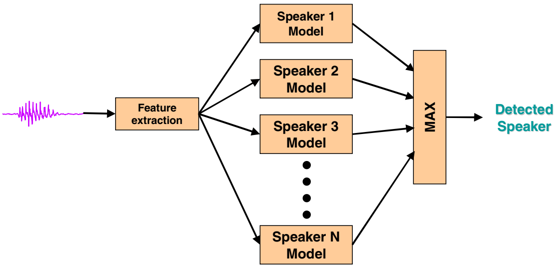


Fig. 4. General classifier structure for speaker identification system

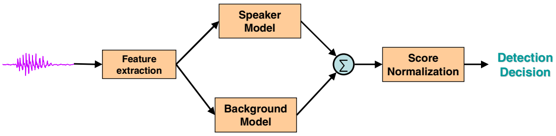


Fig. 5. General classifier structure for speaker verification system

often similarly used for identification). Early methods for speaker recognition included non-parametric techniques (vector quantization and dynamic time warping). Classification methods for speaker recognition in recent years have centered on statistical approaches. The structure and choice of a classifier depends on the application and the features used. In this section we review a subset of classifiers that have been successfully used in automatic speaker recognition systems.

3.1 Gaussian Mixture Modeling (GMM)

The Gaussian mixture modeling (GMM) approach has become one of the mainstay modeling techniques in *text-independent* speaker recognition systems. Consider the verification structure shown in Figure 5. In GMM speaker verification, the impostor model is more commonly known as a background model. In addition, the detection decision or score is normalizated to refine detection decision. The resulting structure is presented in Figure 5.

Figure 5 is realized in the framework of a likelihood ratio detector. In the approach of [3,13,14], we can consider the two hypotheses for a given segment of speech Y:

λ_{hyp}: Speech segment Y is from speaker S
$\lambda_{\overline{hyp}}$: Speech segment Y is not from speaker S

To decide between these two hypotheses we form the following likelihood ratio test:

$$\Lambda(Y) = \frac{p\left(Y|\lambda_{hyp}\right)}{p\left(Y|\lambda_{\overline{hyp}}\right)} \begin{cases} \geq \Theta & \text{Accept hypothesis } \lambda_{hyp} \\ \leq \Theta & \text{Reject hypothesis } \lambda_{hyp} \end{cases} \tag{1}$$

where $p\left(Y|\lambda\right)$ is the probability density function (pdf) of the observed speech segment Y, given the hypothesis λ, or likelihood function. The decision threshold, Θ, determines accepting or rejecting the hypotheses. Let X represent the set of feature vectors generated from the front-end processing of the speech segment Y. The set of features, X, usually MFCCs or LPCCs, are per frame speech-frame vectors: $\{\boldsymbol{x_1}, \cdots, \boldsymbol{x_T}\}$. The frame-based likelihood function can be written as $p\left(\boldsymbol{x}|\lambda\right)$.

In the GMM approach, the choice of the likelihood function is a mixture of M Gaussians:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=i}^{M} w_i p_i(\boldsymbol{x}) \tag{2}$$

where $p_i(\boldsymbol{x})$ is the individual Gaussian density function,

$$p_i(\boldsymbol{x}) = \frac{1}{\left(2\pi\right)^{n/2}\left|\Sigma_i\right|^{1/2}} \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)'\Sigma_i^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_i\right)\right\}. \tag{3}$$

The parameters of the model are: w_i, the mixture weight, $\boldsymbol{\mu}_i$, the N-dimensional mean vector, and Σ_i, the N by N dimensional covariance matrix. The model parameters can be succinctly written as: $\lambda = (w_i, \mu_i, \Sigma_i)$ where $i = [1 \cdots M]$. Equation (2) is just a linearly weighted sum of M individual Gaussians which will be used the likelihood calculation for a detection decision. The weights also satisfy the relation $\Sigma_{i=1}^{M} w_i = 1$. The general form of a Gaussian mixture allows for a fully populated covariance matrix. It has been shown that the diagonal covariance matrix is sufficient for text-independent speaker-verification modeling [3].

Once a model is trained then (2) can be used to evaluate the log-likelihood of model λ for an input test set of feature vectors, X :

$$\log p\left(X|\lambda\right) = \sum_{t=1}^{T} \log p\left(\boldsymbol{x_i}|\lambda\right) \tag{4}$$

Impostor modeling is crucial in producing good speaker recognition performance. Current methods form an universal background model, $p\left(\boldsymbol{x}|\lambda_{\overline{hyp}}\right)$, from a set of background model speakers [15]. The background speakers are chosen from a similarly recorded channel/conditions that will be seen in detection. The number of speakers used to train the background model should be large enough to model the acoustic space of the impostors. There is also a dependency on the number of Gaussians (M) used to model the space. A larger number of Gaussians will require more data to realize the mixture model. The size of M, will depend on channel, application, acoustic variation and amount of speech data seen at each phase. M may range from 64 to 2048. In the telephone speaker-verification task, with 2.5 minutes of enrollment speech and 30 second of verification speech, we have seen good performance with the number of mixtures $M = 512$ and 1-2 hours of background model training speech from over one hundred talkers.

Current state-of-the-art text-independent GMM speaker verification systems obtain background model parameter estimates in an unsupervised manner by using an expectation-maximization (EM) algorithm [16]. Feature vectors generated from a background speaker set provide the training data. The EM algorithm iteratively refines model parameter estimates to maximize the likelihood that the model matches the distribution of the training data. Model parameters converge to a final solution in a few iterations (5-10)[3].

Speaker model training is accomplished by adapting the background model to each enrollment speaker through *Maximum A Posteriori* (MAP) estimation [17,18]. This approach couples the speaker model to the background model and yields better results over the methods using unrelated models. Adapting from the background model utilizes the well trained parameters, $\{w_i, \boldsymbol{\mu}_i, \Sigma_i\}$, from the EM algorithm. The large amount of data used to train the background model allows for a well modeled cepstral space. Speaker models are adapted in turn from this richly populated space. Even though all the parameters of the model can be adapted, it has been shown that best performance results when only the means $(\boldsymbol{\mu}_i)$ are adapted.

The speaker and background models can be applied to the likelihood ratio (1) and (4) to get the likelihood-ratio score,

$$\Lambda(X) = \log p\left(X|\lambda_{hyp}\right) - \log p\left(X|\lambda_{\overline{hyp}}\right) \tag{5}$$

Equation 5 is sufficient to form a detection decision, however better performance is achieved through refinement of the likelihood-ratio score with normalization. We will discuss normalization techniques in Section 3.6.

It should be noted the similarities in the organization of the GMM and the vector quantization (VQ) approach for speaker recognition. In the method of [19,20], the VQ codebook is a partitioning of the cepstral space. The VQ codebook can be weakly considered a quantized version of a Gaussian mixture model.

A support vector machine (SVM) is a versatile classifier that has gained considerable popularity in recent years. An SVM is discriminative and models the boundary between a speaker and a set of impostors. The typical method employed in SVM speaker recognition is based upon comparing speech utterances using sequence kernels. Rather than characterize features from individual frames of speech, these methods model entire sequences of feature vectors. Approaches include the generalized linear discriminant sequence kernel [21], Fisher kernel methods [22,23], n-gram kernels [24], MLLR transform kernels [25], and GMM supervector kernels [26].

Basic SVM Theory. An SVM [27] models two classes using sums of a kernel function $K(\cdot,\cdot)$,

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i t_i K(\mathbf{x}, \mathbf{x}_i) + d, \tag{6}$$

where the t_i are the ideal outputs, $\sum_{i=1}^{N} \alpha_i t_i = 0$, and $\alpha_i > 0$. The vectors $\mathbf{x}_i$ are support vectors and obtained from the training set by an optimization

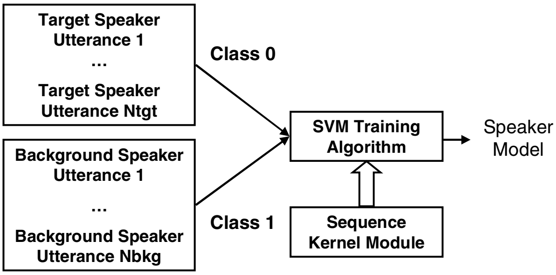

Target Speaker Utterance 1
...
Target Speaker Utterance Ntgt
Background Speaker Utterance 1
...
Background Speaker Utterance Nbkg
Class 0
Class 1
SVM Training Algorithm
Sequence Kernel Module
Speaker Model

Fig. 6. Setup for training an SVM classifier for speaker verification

process [28]. The ideal outputs are either 1 or -1, depending upon whether the corresponding support vector is in class 0 or class 1, respectively. For verification, a class decision is based upon whether the value, $f(\mathbf{x})$, is above or below a threshold.

The kernel $K(\cdot,\cdot)$ is typically constrained to have the Mercer condition, so that $K(\cdot,\cdot)$ can be expressed as

$$K(\mathbf{x},\mathbf{y}) = \mathbf{b}(\mathbf{x})^t\mathbf{b}(\mathbf{y}), \tag{7}$$

where $\mathbf{b}(\mathbf{x})$ is a mapping from the input space (where $\mathbf{x}$ lives) to a possibly infinite-dimensional *expansion space*. Optimization of an SVM relies upon a maximum margin concept. For separable data, the system places a hyperplane in a high dimensional space so that the hyperplane has maximum margin. The data points from the training set lying on the boundaries are the support vectors in equation (6).

Application of Support Vector Machines to Speaker Recognition. Figure 6 indicates the basic training strategy for SVMs using sequence kernels. We train a target model with target speaker utterances and a set of example speakers' utterances that have characteristics of the impostor population—a background speaker set. Each utterance from a target or background speaker becomes a point in the SVM expansion space. We implement a sequence kernel module for comparing two utterances and producing a kernel value. The kernel module is connected into a standard SVM training tool which then produces a speaker model. We keep the background speaker set the same as we enroll different target speakers.

Sequence Kernels for Speaker Recognition—General Structure. To apply an SVM, $f(\mathbf{X})$, in a speaker recognition application, we need a method for calculating kernel values from sequences of features (e.g., MFCC feature vectors). Two general methods have emerged—linearized train/test kernels and adapted model comparison.

The idea of a train/test sequence kernel is shown in Figure 7. The basic approach is to compare two speech utterances, *utt 1* and *utt 2* by training a model on one utterance and then scoring the resulting model on another utterance.

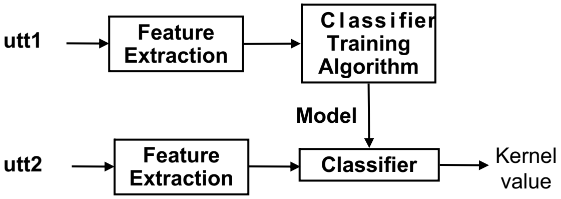


Fig. 7. Constructing a sequence kernel using a train/test strategy

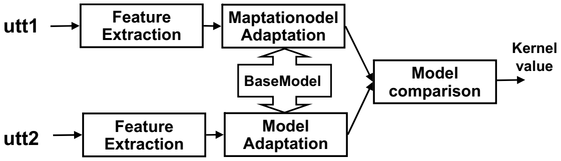


Fig. 8. Constructing a kernel using a base generative model

This process produces a value that measures the similarity between the two utterances. Although in general this comparison is not a kernel, it doesn't satisfy the Mercer condition, in many cases linearization will produce a kernel—see the next section.

The second basic method for constructing sequence kernels is shown in Figure 8. In this setup, we adapt a base model to obtain probability distributions which represent the utterances. We then apply a model comparison algorithm to get a measure of similarity. This approach has the useful property that it is naturally symmetric as long as the comparison calculation is symmetric.

3.2 Sequence Kernels for Speaker Recognition—Specific Examples

For the train/test kernel shown in Figure 7, a typical approach is the generalized linear discriminant sequence (GLDS) kernel [21]. In this method, the classifier is taken to be a polynomial discriminant function. Suppose we have two sequences of feature vectors, $\mathbf{X} = \{\mathbf{x}_i\}$ and $\mathbf{Y} = \{\mathbf{y}_j\}$. If we train a polynomial discriminant using mean-squared error, then the resulting kernel is given by

$$K(\mathbf{X}, \mathbf{Y}) = \bar{\mathbf{b}}_x \bar{\mathbf{R}}^{-1} \bar{\mathbf{b}}_y. \tag{8}$$

In (8),

$$\bar{\mathbf{b}}_x = \frac{1}{N_x} \sum_i \mathbf{b}(\mathbf{x}_i); \tag{9}$$

i.e., $\bar{\mathbf{b}}_x$ is the average expansion over all frames. A similar expansion is used for $\mathbf{Y}$. The matrix, $\bar{\mathbf{R}}$ is the correlation matrix of a background data set; typically, it is approximated with only diagonal terms. For details on the derivation of these equations, we refer to [21]. An interesting generalization of the GLDS kernel is to replace the polynomial expansion by a general kernel using the kernel trick [29,30].

For the generative model sequence kernel, several methods have been proposed. Current methods are based upon adapting from a GMM or HMM base model. In [25], adaptation of an HMM from a speech-to-text system is performed using maximum-likelihood linear regression (MLLR). The MLLR adaptation parameters are then compared with a weighted linear inner product. In [26], the adaptation is performed via MAP adaptation of a GMM. The GMMs are compared using either an approximation to the KL divergence or an integral inner product.

SVMs can also be applied to high-level features [11,24]. A token-sequence comparison kernel can be derived by using the train/test kernel framework in Figure 7. In this case, the classifier in the figure is taken to be the standard language model likelihood ratio using n-gram probabilities. The resulting kernel is of the form

$$K(T_1, T_2) = \sum_k D_k^2 p(d_k|T_1) p(d_k|T_2) \tag{10}$$

where the T_j are token sequences, D_k is a weighting function, $p(d_k|T_j)$ is the probability of a particular n-gram, d_k, occuring in token sequence T_j. A typical choice is something of the form

$$D_k = \min\left(C_k, g_k\left(\frac{1}{p(d_k|\text{background})}\right)\right) \tag{11}$$

where $g_k(\cdot)$ is a function which squashes the dynamic range, and C_k is a constant [24]. The probability $p(d_k|\text{background})$ in (11) is calculated from a large population of speakers. Typical choices for g_k are $g_k(x) = \sqrt{x}$ and $g_k(x) = \log(x) + 1$. The kernel (10) is closely related to methods in information retrieval; we refer to [24] for details.

3.3 Support Vector Machine (SVM)

3.4 Hidden Markov Modeling (HMM)

The GMM-UBM system described in Section 3.1 models the entire acoustic space. However, in text-dependent applications the system has prior knowledge of what will be said and template-matching techniques become advantageous. The first template matching methods were dynamic time warping (DTW) algorithms [31]. However DTW methods proved to be inefficient and methods gave way to a stochastic modeling of each talker's speech where the underlying stochastic processes is not observable of hidden (Hidden Markov Model). Early approaches in applying Hidden Markov Models (HMMs) to text-dependent and text-independent speaker recognition were developed by [15,32,10] and have been continued [33,34,35].

HMMs can efficiently model statistical variations in spectral features. Rather then modeling the entire acoustic space the HMM only models a progression of limited regions of acoustic space. These limited acoustic regions can be defined as states of finite time. These states can be described with a PDF, $p(\boldsymbol{x}_t|s)$ is the probability of per-frame feature vector, $\boldsymbol{x_t}$, given you are in state, s. Transitioning between states, (e.g.: from $t-1$ to t) is defined with a state transition probability, $p(s_t|s_{t-1})$.

The likelihood of T frames of speech occurred given a hypothesis, λ, is:

$$p(X|\lambda) = \sum_{\substack{\text{all}\\\text{states}}} \prod_{t=1}^{T} p(s_t|s_{t-1})p(x_t|s_t) \tag{12}$$

Which is the Baum-Welch decoding [36,37,38]. Equation (12) can be employed in a similar manner as (5). The likelihood ratio can be constructed from a target likelihood $p(X|\lambda_{hyp})$over the an impostor/background likelihood $p(X|\lambda_{\overline{hyp}})$ as in (1).

The first step in HMM modeling is to form a representation of the impostors. Here the concept of the background model is to form a model of the world of all possible speakers. HMM background models can then be trained through the use of a full large vocabulary continuous speech recognition (LVCSR) system as in [35,39]. There are also approaches that use segmental K-means clustering procedure [33] or limited vocabulary phoneme-based methods were implemented in [40].

The speaker model, $p(X|\lambda_{hyp})$, can be formed by Baum-Welch adaptation from the background model [35]. [33] relies on segmental K-means clustering for training of the target model, but utilizes the speaker independent background model for the segmentation. This can be considered a general form of the GMM approach presented in Section 3.1. The GMM can be thought of as a single state hidden Markov model.

The HMM implementation of [35,39] can either be applied in text-independent or text-dependent applications. For text-independent applications, the language model of the LVCSR system has to be broad enough to span the speech that may be seen by the system.

The actual structure of a text-dependent system will depend greatly on the application. Speaker recognition accuracy is dependent on the performance of the system, but can also be controlled by limiting the vocabulary of the domain. Limiting the talkers to alpha-digits is a common domain. System accuracy may also be influenced by gathering more speech from cooperative speaker by prompting them with a series of random phrases.

3.5 Artificial Neural Networks

Artificial neural networks (ANNs) model continuous features using nonlinear modeling inspired by biological neural networks. A typical artificial neural network is a two-layer perceptron, $m(\mathbf{x})$, of the form

$$m(\mathbf{x}) = \tilde{g}\left(\mathbf{w}^t g(\mathbf{A}\mathbf{x} + c) + d)\right) \tag{13}$$

where $\mathbf{x}$ is the input, $g(\cdot)$ and $\tilde{g}$ are squashing functions, $\mathbf{A}$ is a matrix, $\mathbf{w}$ is a vector, and b and c are bias terms. Artificial neural networks were one of the first methods to be successfully used in discriminative speaker recognition [41].

ANNs, when trained with mean-squared or cross-entropy criteria [42], model the posterior probability, $p(\text{spk}|\mathbf{x}_i)$. Here, $\mathbf{x}_i$ is typically a continuous feature vector such as MFCCs. A typical scoring criterion is to take the average weighted posterior (or log posterior) across all frames of an input utterance.

Because an ANN models a posterior rather than a likelihood, typically cohort normalization or background normalization is not needed to achieve good perfromance. This property is expected since the ANN is a discriminative technique. But, as with most speaker recognition methods, techniques such as TNorm can stabilize thresholds.

Training for an ANN is accomplished in a similar manner to the SVM setup shown in Figure 6 except it is performed with frame level features. Feature vectors for the target speaker are extracted and placed in one class (with ideal output 1). Feature vectors for a background speaker set are placed in another class (with ideal output 0). Then, training with a backpropagation algorithm algorithm is performed.

Note that prior balancing is a critical part of ANN training. Because the target speaker training set size is typically significantly smaller than the background training set, the prior of the target is usually small. Since the output of the ANN approximates a posterior, the target prior is a factor in the ANN output. Compensation for this prior can be performed in training via, e.g. random sampling with prior equalization, or in testing by scaling the output by the target prior.

A successful extension of ANNs is the neural tree network [41] (NTN). NTNs are a combination of tree methods (such as CART) and neural networks. At each node in the tree, a neural network is used to determine which branch is taken. Scoring and training are an extension of standard ANN and tree methods. NTNs were successfully used for many years in a commercial system for text dependent speaker recognition.

Other connectionist methods for speaker recognition include radial basis functions (RBF) and elliptical basis functions (EBF), e.g. [43]. These approaches were only moderately successful and are subsumed by the more general training and modeling approach of GMMs.

3.6 Normalization Techniques

Ideally, score variability should only depend on speaker differences. Other factors may contribute to score variability such as transmission channel, environmental background effects, linguistic variation and session variation. There are many methods to stabilize score variation to make the threshold setting, Θ, more robust. Compensation methods have been developed in the feature domain, model domain, and score domain.

Feature Domain Normalization. Feature domain normalization transforms a base set of features, such as MFCCs, to a new set of features that are more robust to channel and noise effects. Typically, these methods have been based on signal processing and data-driven techniques.

Common feature transformations used to remove channel effects are RASTA [44] and cepstral mean subtraction (CMS) [45]. These methods rely on homomorphic signal processing techniques—filtering a signal in the time domain induces an additive bias in the cepstral domain.

Feature transformations that compensate for noise or other nonlinear distortions include cepstral variance normalization (CVN) and feature warping. CVN, in part, is based upon the fact that additive noise reduces the variance of cepstral coefficients [46]; compensation is realized by renormalizing the cepstral coefficients to unit variance. Feature warping [47] further extends this technique by remapping features to fit some predefined distribution.

More recent feature compensation methods have used supervised data-driven methods. For example, feature mapping [48], uses knowledge of channel types to remap features to a channel neutral model.

Model Domain Normalization–GMM. For GMM based classifiers, techniques that treat the undesired variability as a bias to the mean vectors have been successful. If we stack the means from a GMM into a *supervector* this can be written as

$$\boldsymbol{m}_j(s) = \boldsymbol{m}(s) + \boldsymbol{c}(s) \tag{14}$$

where $\boldsymbol{m}_j(s)$ is the supervector from speaker s's j-th enrollment session, $\boldsymbol{m}(s)$ is the desired compensated supervector for speaker s and $\boldsymbol{c}(s)$ is the undesired variability supervector.

The main difference in the compensation techniques is in how they estimate and remove the variability vector $\boldsymbol{c}(s)$. In Speaker Model Synthesis (SMS) [49], the difference between bias vectors from a set of pre-defined channel types is used to synthetically generate a library of channel-dependent speaker models so as to allow matched-channel likelihood ratio scoring during recognition. More recent latent factor analysis (LFA) based techniques [50,51], model the supervector bias as a low-dimensional normally distributed bias,

$$\boldsymbol{c}(s) = U\boldsymbol{n}(s) \tag{15}$$

where U is the low-rank session loading matrix. The LFA techniques are aimed specifically at compensation of session variability and do not require prior channel detectors or parameters.

Model Domain Normalization–SVM. As with the GMM, compensations with SVM classifiers can also be applied directly in the model domain. The SVM nuisance attribute projection (NAP) method [52] works by removing subspaces that cause variability in the kernel. NAP constructs a new kernel,

$$\begin{aligned} K(\{\mathbf{x}_i\},\{\mathbf{y}_j\}) &= \left[\mathbf{P}\bar{\mathbf{b}}_x\right]^t \left[\mathbf{P}\bar{\mathbf{b}}_y\right] \\ &= \bar{\mathbf{b}}_x^t \mathbf{P}\bar{\mathbf{b}}_y \\ &= \bar{\mathbf{b}}_x^t (\mathbf{I} - \mathbf{v}\mathbf{v}^t)\bar{\mathbf{b}}_y \end{aligned} \tag{16}$$

where $\mathbf{P}$ is a projection ($\mathbf{P}^2 = \mathbf{P}$), $\mathbf{v}$ is the direction being removed from the SVM expansion space, $\mathbf{b}(\cdot)$ is the SVM expansion, and $\|\mathbf{v}\|_2 = 1$. NAP can be applied to both low-level and high-level features.

Score Normalization. Typically, score normalization techniques remap target speaker scores based on some reference set of models, utterances, or channels. One of the most effective score normalization techniques, TNorm (test-normalization) was introduced in [53]. TNorm transforms a target model score, s, to

$$\frac{s - \mu}{\sigma} \tag{17}$$

where μ and σ are the mean and standard deviation of scores from a set of reference speakers' models scored on the input utterance. Other score normalization techniques include Z-Norm [54] (based on normalizing to a reference set of utterances) and H-Norm (based on normalizing to a reference set of channels) [55].

4 Classifier Choice

The choice of classifier to be used is greatly dependent on the application. Examples of application constraints that influence the classifier choice and configuration include the following.

- *Level of user cooperation*
- *Required recognition/detection accuracy*
- *Expected channels*
- *Amount of speech available for enrollment and detection*
- *Available compute and memory resources*
- *How the output is used*

User cooperation will determine whether or not you can field an active or passive system. If the user is cooperative the system can actually prompt the user for additional input speech. The additional input speech will boost performance while at the same time verify that the incoming user is "live". However if the users are uncooperative the system has take to more of a passive role. In these applications the systems have no control over the data they process.

High *recognition/detection accuracy* may be a requirement in areas such as banking account access. Here, it is desirable to be very accurate in who gets access to a user's account. A text-dependent system is applicable in this case since it offers higher performance then text-independent techniques.

The *channel* consists of, the type of microphone used to record the speech, the way the speech is encoded/transmitted, as well has ambient noises. If the application has to deal with a wide variety of channel conditions the classifier could employ some form of channel compensation to boost performance.

The *amount of speech data available for enrollment and detection* will also help determine the classifier. If more data is available then classifiers that key off of high level information become feasible.

Applications may also be limited in *computation and memory resources*. Embedded devices have limited amounts of processing power and available memory. A cell phone will have very limited capabilities that will uniquely constrain the speaker recognizer.

The consumer of the *output of the system* will determine what information is presented to the end user. Certain forensic applications require that systems return word usage and phonotactic information. In this application a word or phone based recognition systems, as described in Section 3.3, may be required to generate the information needed by the user. Further the type of output may need to be a hard decision, a human interpretable score, or a relative score to used by another automatic process.

It is quite difficult to characterize the accuracy of speaker verification systems in all applications due to the complexities and differences in the enrollment/detection scenarios. Figure 9 attempts to provide a range of performance for some of the cases mentioned above. These numbers are not meant to indicate the best performance that can be obtained, but rather a relative ranking of some different scenarios. In Figure 9, we depict a detection error trade-off (DET) plot, which shows the trade-off between false-rejects, f_r, and false-accepts, f_a, as the decision threshold changes in a verification system. On this DET we show four equal error rate points (EER is a summary performance indicator where $f_r = f_a$) for four different verification application scenarios. One thing to note is that system performance improves as more constraints are placed on the application conditions (e.g., text-dependent vs. test-independent, increased speech for enrollment and verification, more benign channels).

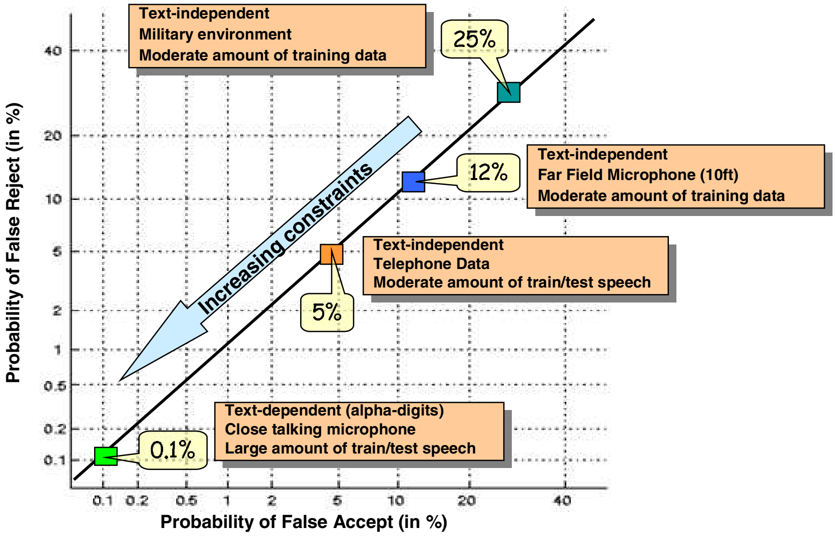


Fig. 9. Range of speaker verification performance

To examine some differences in classifiers, Figure 10 shows EER performance for a few of the text-independent systems described in section 3 for two conditions of enrollment data [56,57]. In the first condition about 2.5 minutes of speech is available for both enrollment and detection. In the other condition about 20 minutes of speech is available for enrollment and 2.5 minutes is available for detection. As expected, the trend is for performance to get better when more enrollment data is available. Further we see that spectral systems (GMM-LFA and SVM-GSV) perform better than high-level feature systems (SVM Word), but fusion of high and low level systems can produce some performance gains.

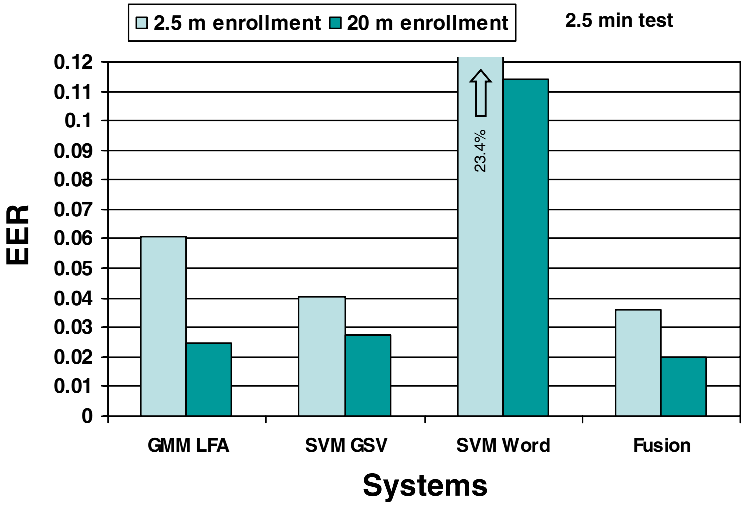


Fig. 10. The performance measure equal error rate for text-independent speaker verification systems

5 Conclusions

In this chapter, we have provided a brief overview of the classification methods used in speaker recognition. In Section 2, we presented some of the common feature extraction techniques that are currently being used in speaker recognition systems. In Section 3, we described classification methods that are representative of those currently being studied in research and used in application. We introduced common approaches for text-dependent and text-independent applications, as well as offering some historical evolution of how these classifiers came to be used.

Future work in speaker recognition will continue to exploit advances in speech science, classification, and engineering. Speech science continues to give insight into feature that characterize speakers—speaker idiolect, speaker dialect, as well as vocal characteristics (roughness, breathiness, etc.). More precise measurements and techniques for extracting these features will lead to more diverse and accurate speaker recognition systems.

Classification continues to be a strong component of the speaker recognition problem. Specialization of classification techniques to deal with speaker recognition challenges will no doubt lead to significant improvements. Current trends are methods that deal with channel variability, the continuum of feature types, and general mismatch.

Finally, engineering provides a feedback to all of the design techniques. Implementing and deploying technologies to different application domains—forensic, security, etc.—gives insight into robustness, computation, and fusion of speaker characterization techniques.

References

1. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech, Signal Processing, ASSP 28(4), 357–366 (1980)
2. Quatieri, T.: Discrete-Time Speech Signal Processing: Principles and Practice. Prentice-Hall, Englewood Cliffs (2001)
3. Reynolds, D.A., Quatieri, T.F., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10(1-3), 19–41 (2000)
4. Tierney, J.: A study of LPC analysis of speech in additive noise. IEEE Trans. Acoust., Speech, Signal Processing, ASSP 28(4), 389–397 (1980)
5. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs (1993)
6. Adami, A., Mihaescu, R., Reynolds, D.A., Godfrey, J.J.: Modeling prosodic dynamics for speaker recognition. In: Proc. ICASSP, pp. IV–788–IV–791 (2003)
7. Peskin, B., Navratil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D., Xiang, B.: Using prosodic and conversational features for high-performance speaker recognition: Report from JHU workshop. In: Proc. ICASSP (2003)
8. Doddington, G.: Speaker recognition based on idiolectal differences between speakers. In: Proc. Eurospeech, pp. 2521–2524 (2001)
9. Navrátil, J., Jin, Q., Andrews, W.D., Campbell, J.P.: Phonetic speaker recognition using maximum-likelihood binary-decision tree models. In: Proc. ICASSP, pp. IV–796–IV–799 (2003)
10. Matsui, T., Furui, S.: Concatenated phoneme models for text-variable speaker recognition. In: Proc. ICASSP, vol. II, pp. 391–394 (1993)
11. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A., Leek, T.R.: Phonetic speaker recognition with support vector machines. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16, MIT Press, Cambridge (2004)
12. Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J., Hernandez-Cordero, J.: Gender-dependent phonetic refraction for speaker recognition. In: Proc. ICASSP, pp. I149–I153 (2002)
13. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Meignier, S., Merlin, T., Ortega-Garc, J., Magrin-Chagnolleau, I., Petrovska-Delacretaz, D., Reynolds, D.A.: A tutorial on text-independent speaker verication. EURASIP Journal on Applied Signal Processing 4, 430–451 (2004)

14. Reynolds, D.A.: Speaker identification and verification using gaussian mixture speaker models. Speech Commun. 17(1-2), 91–108 (1995)
15. Carey, M., Parris, E., Bridle, J.: A speaker verification system using alpha-nets. In: Proc. ICASSP (1991)
16. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39, 1–38 (1977)
17. Gauvain, J.-L., Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains. IEEE Trans. Speech and Audio Processing 2(2), 291–298 (1994)
18. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley and Sons, New York (1973)
19. Soong, F., Rosenberg, A., Rabiner, L., Juang, B.: A vector quantization approach to speaker recognition. In: Proc. ICASSP, pp. 387–390 (1985)
20. Rosenberg, A., Soong, F.: Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. In: Proc. ICASSP, pp. 873–876 (1986)
21. Campbell, W.M.: Generalized linear discriminant sequence kernels for speaker recognition. In: Proc. ICASSP, pp. 161–164 (2002)
22. Fine, S., Navrátil, J., Gopinath, R.A.: A hybrid GMM/SVM approach to speaker recognition. In: Proc. ICASSP (2001)
23. Wan, V., Renals, S.: SVMSVM: support vector machine speaker verification methodology. In: Proc. ICASSP, pp. 221–224 (2003)
24. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A., Leek, T.R.: High-level speaker verification with support vector machines. In: Proc. ICASSP, pp. I–73–76 (2004)
25. Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., Venkataraman, A.: MLLR transforms as features in speaker recognition. In: Proc. Interspeech, pp. 2425–2428 (2005)
26. Campbell, W.M., Sturim, D.E., Reynolds, D.A., Solomonoff, A.: SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proc. ICASSP, pp. I–97–I–100 (2006)
27. Cristianini, N., Shawe-Taylor, J.: Support Vector Machines. Cambridge University Press, Cambridge (2000)
28. Collobert, R., Bengio, S.: SVMTorch: Support vector machines for large-scale regression problems. Journal of Machine Learning Research 1, 143–160 (2001)
29. Louradour, J., Daoudi, K., Bach, F.: SVM speaker verification using an incomplete cholesky decomposition sequence kernel. In: IEEE 2006 Odyssey: The Speaker and Language Recognition Workshop (2006)
30. Mariéthoz, J., Bengio, S.: A max kernel for text-independent speaker verification systems. In: Second Workshop on Multimodal User Authentication (2006)
31. Soong, F.K., Rosenberg, A.E.: On the use of instantaneous and transitional spectral information in speaker recognition. In: Proc. ICASSP, pp. 877–880 (1986)
32. Matsui, T., Furui, S.: Speaker recognition using concatenated phoneme models. In: Proc. ICSLP (1992)
33. Rosenberg, A.E., Parthasarathy, S.: Speaker background models for connected digit password speaker verification. In: Proc. ICASSP, pp. 81–84 (1996)
34. Corrada-Emmanuel, A., Newman, M., Peskin, B., Gillick, L., Roth, R.: Progress in speaker recognition at dragon systems. In: Proc. ICSLP (1998)
35. Weber, F., Peskin, B., Newman, M., Corrada-Emmanuel, A., Gillick, L.: Speaker recognition on single- and multispeaker data. Digital Signal Processing 10, 75–92 (2000)

36. Rabiner, L.R., Juang, B.H.: An introduction to hidden markov models. IEEE ASSP Mag. 3 (1986)
37. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 77(2), 257–285 (1989)
38. Campbell, J.P.: Speaker recognition: A tutorial. Proc. of the IEEE 85(9), 1437–1462 (1997)
39. Newman, M., Gillick, L., Ito, Y., McAllaster, D., Peskin, B.: Speaker verification through large vocabulary continuous speechrecognition. In: Proc. ICSLP (1996)
40. Matsui, T., Furui, S.: Likelihood normalization for speaker verification using phoneme- and speaker-independent model. In: Speech Communication (1995)
41. Farrell, K.R., Mammone, R.J., Assaleh, K.T.: Speaker recognition using neural networks and conventional classifiers. IEEE Trans. on Speech and Audio Processing 2(1), 194–205 (1994)
42. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
43. Oglesby, J., Mason, J.: Radial basis function networks for speaker recognition. In: Proc. ICASSP, pp. 393–396 (May 1991)
44. Hermansky, H., Morgan, N., Bayya, A., Kohn, P.: Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP). In: Proc. Eurospeech, pp. 1367–1371 (1991)
45. Atal, B.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America 55(6), 1304–1312 (1974)
46. Mansour, D., Juang, B.: A family of distortion measures based upon projection operation for robust speech recognition. IEEE Trans. Acoust., Speech, Signal Processing, ASSP 37, 1659–1671 (1989)
47. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proc. of Speaker Odyssey Workshop, pp. 213–218 (2001)
48. Reynolds, D.A.: Channel robust speaker verification via feature mapping. In: Proc. ICASSP, vol. 2, pp. II–53–56 (2003)
49. Teunen, R., Shahshahani, B., Heck, L.: A model-based transformational approach to robust speaker recognition. In: Proc. ICSLP (2000)
50. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Trans. Speech and Audio Processing 13(3), 345–354 (2005)
51. Vogt, R., Baker, B., Sriharan, S.: Modelling session variability in text-independent speaker verification. In: Proc. Interspeech, pp. 3117–3120 (2005)
52. Solomonoff, A., Campbell, W.M., Boardman, I.: Advances in channel compensation for SVM speaker recognition. In: Proc. ICASSP (2005)
53. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing 10, 42–54 (2000)
54. Reynolds, D.A.: Comparison of background normalization methods for text independent speaker verification. In: Proc. Eurospeech, pp. 963–966 (1997)
55. Heck, L., Weintraub, M.: Handset-dependent background models for robust text-independent speaker recognition. In: Proc. ICASSP, pp. 1071–1074 (1997)
56. Campbell, W.M., Navratil, J., Reynolds, D.A., Shen, W., Sturim, D.E.: The MIT/IBM 2006 speaker recognition system:High-performance reduced complexity recognition. In: ICASSP (2007)
57. Reynolds, D.A., Campbell, W., Gleason, T., Quillen, C., Sturim, D., Torres-Carrasquillo, P., Adam, A.: The 2004 MIT Lincoln Laboratory speaker recognition system. In: ICASSP (2005)

Multi-stream Fusion for Speaker Classification

Izhak Shafran

Computer Science & Electrical Engineering
OGI School of Science & Engineering
Oregon Health & Science University (OHSU)
20000 NW Walker Rd, Beaverton, OR 97006
zakshafran@cslu.ogi.edu

Abstract. Accurate detection of speaker traits has clear benefits in improving speech interfaces, finding useful information in multi-media archives, and in medical applications. Humans infer a variety of traits, robustly and effortlessly, from available sources of information, which may include vision and gestures in addition to voice. This paper examines techniques for integrating information from multiple sources, which may be broadly categorized into those in feature space, model space, score space and kernel space. Integration in feature space and model space has been extensively studied in the context of audio-visual literature, and here we focus on score space and kernel space. There are large number of potential schemes for integration in kernel space, and here we examine a particular instance which can integrate both acoustic and lexical information for affect recognition. The example is taken from a widely-deployed real-world application. We compare the kernel-based classifier with other competing techniques and demonstrate how it can provide a general and flexible framework for detecting speaker characteristics.

Keywords: Mutli-stream Fusion, Rational Kernels, Affect Recognition, Speaker Recognition, Language Recognition, Score Combination.

1 Introduction

Humans infer a number of important traits about a speaker from his or her voice, apparently without any effort and as a matter of routine. These features may include gender, age, dialect, ethnicity, affect, level of education, and even state of intoxication. Automation of this ability to infer traits can provide clear benefits in the design of more natural human-machine speech interfaces, the extraction of useful information from large quantities of speech data and in medical applications related to speech and cognitive skills.

Given the widespread deployment of automated speech interfaces, the knowledge of these traits could be exploited to increase their acceptability in society. For example, an analysis of the performance of automatic speech recognition (ASR) with respect to age, gender, affect and dialect or accent of the speaker revealed that dialect or accent was the most influential trait in predicting the word

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 298–312, 2007.

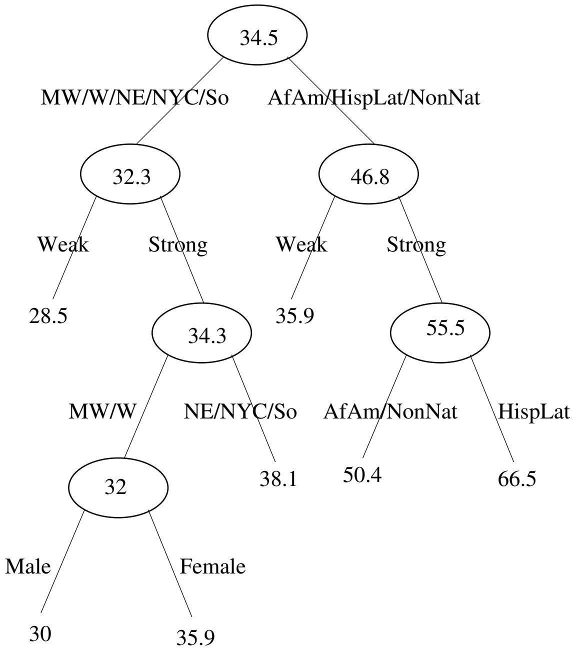


Fig. 1. The decision tree obtained by automatic clustering of utterances to predict WER using age, gender, affect and accent/dialect. Accent or dialect categories were derived from TIMIT and were marked as Mid-Western(MW), Western (W), North-Eastern (NE), New York City (NYC), Southern (So), African-American (AfAm), Hispanic or Latino (HispLat) and other non-natives (NonNat). Age was grouped into three categories and affect into two categories as described in [28].

error rate (WER) of AT&T's "How May I Help You" spoken-dialog system [1]. Figure 1 shows the decision tree obtained by automatic clustering of utterances with respect to WER, where the splits were evaluated using multi-fold cross-validation. The WER varies from 28.5% for a person with a weak mid-western accent to 68.5% for someone with a strong Hispanic or Latino accent. This means that the spoken-dialog system at a call center failed to recognize two in three words from an entire segment of the population, rendering it practically useless to them. Equipped with the knowledge of speaker traits, systems can be designed to tackle a wider range of scenarios.

For the most part, speaker characteristics are static, or they vary slowly over the course of an utterance. Thus, the problem of detecting speaker characteristics can be formulated as that of assigning a class label to each utterance. For traits that remain constant, certain applications may permit sufficient latency to utilize multiple utterances for inference. The actual classification may be performed by *maximum aposteriori* (MAP) decision rule, support vector machines (SVM), or even generalized linear models (GLM).

In addition to cues from voice, we humans seamlessly exploit other modalities such as facial features or expressions and body language or gesture, whenever they are available. Researchers have examined a variety of models for integrating different modalities. The modalities may be audio and visual (*multi-modal*), or even different streams of features from audio (*intra-modal*), such as in language recognition. The *maximum aposteriori* decision rule for speaker classification is similar to audio-visual speech recognition, and equivalent to recognizing isolated words or digits. As a result the generative models, such as the variants of HMMs, developed for audio-visual speech recognition are applicable in speaker classification. The parameters of these models can be learned by maximizing likelihood, accuracy or even equal error rate (EER), a metric that is popular in biometric evaluation.

This chapter will focus on techniques for integrating multiple streams. They can be broadly categorized into integration in feature space or *early integration*, in score space or *late integration*, in model space using variants of HMMs, and in kernel space. The integration in feature space and in model space have been investigated extensively in the context of audio-visual speech recognition. Here we include a brief review of both for completeness.

2 Feature Space

When streams have the same data rate, the observations can be concatanated. This increases the dimension of the feature space, requiring more parameters to model them. The redundancies in feature space can be removed either by projecting them into a lower dimension or by selecting only the components or dimensions that are relevant for classification or recognition. In this context, the most popular scheme for linear projection is the linear discriminant analysis (LDA), which this has been used in several published results for affect recognition (e.g. [2]). LDA inherently assumes that within-class covariances are equal, and this assumption is not necessary. By casting LDA as a problem of constraint maximum likelihood estimation, it has been generalized to heteroscedastic LDA (HLDA) [3]. A variant of HLDA, using diagonalizing linear transformation (MLLT), has been shown to be effective in automatic speech recognition [4], as well as in integrating streams for audio-visual speech recognition [5]. Alternative linear projections such as latent semantic analysis and canonical correlation analysis have also been investigated for feature integration [6]. Feature selection of components using greedy search has also been employed to discard redundant dimensions [7].

3 Score Space

Speaker and language recognition often employ *late integration*, where the integration of information from intra-modal streams is delayed till the last stage. Scores are computed for each class, C, and each stream, X_i, and then summed to decide the output class, $S(X, C) = \sum_{i=1} S(X_i, C)$.

Empirical studies have shown that score normalization has a significant impact on EER [8]. The two popular schemes for normalization are: T-norm and Z-norm. The Z-norm operates on the score distribution using target-specific statistics:

$$S_{z-norm}(X_i, C) = \frac{S(X_i, C) - \mu_C}{\sigma_C} \tag{1}$$

where μ_C and σ_C are the mean and the standard deviation of the scores, $S(X_i, C)$ of the target class (e.g. a speaker). Alternatively, the normalization can be performed over the test input, X_i. Thus, the T-norm computes the normalization based on statistics of the competing classes (e.g. cohort speakers) for a given input:

$$S_{t-norm}(X_i, C) = \frac{S(X_i, C) - \mu_{X_i}}{\sigma_{X_i}} \tag{2}$$

T-norm and Z-norm can be regarded as variants of a unified scheme for score normalization [9].

The normalized scores may be fused with *order statistics* such as maximum, minimum and median, logic operators such as AND and OR, simple summation, neural networks or support vector machines [10,11,12,13]. The resulting EER performance depends on the diversity and correlation between different streams. Often a higher degree of correlation can be expected between intra-modal experts than between extra-modal experts, while different degrees of performance can occur in both. Thus, four scenarios come into play.

1. Combining uncorrelated experts with very different performance
2. Combining highly correlated experts with very different performance
3. Combining uncorrelated experts with very similar performance
4. Combining highly correlated experts with very similar performance

The common intuition that best performance will be obtained in the third case has also been observed empirically. This intuition can be quantified with a theoretical analysis.

A theoretical analysis can be performed by modeling the score from each stream, i, for a target class, k, as a sum of two terms – a bias term, μ_i^k, and a Gaussian random noise component, $w_i^k \sim \mathcal{N}(0, \sigma_i^k)$ [14]. Consider a two class problem, say $C \in \{c, i\}$. Equal error rate (EER), a popular evaluation measure for speaker characteristics, is the error rate when false alarm is equal to miss. For a fusion operation such as summation, the combined score will be Gaussian for each target. Therefore, the EER is given by the error function [15], which can be parameterized in terms of a certain F_{ratio}:

$$EER = \frac{1}{2} - \frac{1}{2} erf(\frac{F_{ratio}}{\sqrt{2}}) \tag{3}$$

$$F_{ratio} = \frac{\mu^{k=c} - \mu^{k=i}}{\sigma^{k=c} + \sigma^{k=i}} \tag{4}$$

where μ^k and σ^k are the means and variances for each class respectively. The scores from the base-experts for each stream can be rescaled so that the numerators are the same and this scales the variances σ^k accordingly. In most experts the variance of the two classes are proportional and so for the sake of simplifying the illustration, let $\sigma_1^{k=c} = \sigma_1^{k=i}$, a reasonable assumption that allows the class subscript,k, to be dropped. Thus, the EER has a nonlinear inverse-dependency on the F_{ratio} which in turn inversely depends on the variance of the combined score. The variance of the combined score can be written in terms of the variance of base-experts and their correlation coefficient, ρ.

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2 \tag{5}$$

Now, the EER for the four cases can be analyzed in terms of the impact on the combined score, σ. Without loss of generality, let one base-expert be better than the other, $\sigma_1 \leq \sigma_2$. Substituting Eqn. 5, an improvement over base-experts can be observed only if $\sigma < \sigma_1^2$. In other words, the requirement for the combined system to perform better than the best base-expert is:

$$\sigma_2^2 < 3\sigma_1^2 - 2\rho\sigma_1\sigma_2 \tag{6}$$

This provides a minimum requirement on variance of the base-experts when they are uncorrelated. When the base-experts are uncorrelated, $\rho \to 0$, the fusion will improve performance only when the variance in the scores of the poorer system, σ_2^2, is strictly less than three times that of the better system, σ_1^2. In the third case, $\sigma_1^2 \approx \sigma_2^2$ and Eqn. 6 reduces to $\rho\sigma_2^2 < \sigma_1^2$, which is always true and hence such a combination will always improve performance. Positive correlation usually reduces performance, while negative correlation helps. Negative correlation, however, is not a sufficient condition to improve performance; additional constraints on variance are important as well, as explained in [14]. In a battery of empirical tests, these theoretical predictions were shown to hold even when the Gaussian assumptions were not satisfied [14].

4 Model Space

Standard HMMs can be written in terms of composite states and two streams, which will be instructive in understanding the different trade-offs involved in integrating streams.

$$\pi(i,j) = P(q_0^1 = i, q_0^2 = j) \tag{7}$$
$$a_{i,j|k,l} = P(q_t^1 = i, q_t^2 = j | q_{t-1}^0 = k, q_{t-1}^1 = l) \tag{8}$$
$$b_t(i,j) = P(O_t^1, O_t^2 | q_t^1 = i, q_t^2 = j) \tag{9}$$

where (i,j) denotes the composite tuple. Following the usual notation for HMMs, π, a, b, O, and q stand for probability of initial state configuration, the transition matrix, the observation probability, the observations from the streams and the hidden states respectively. The number of parameters involved are $O(\pi) = NM$,

$O(a) = N^2M^2$, and $O(b) = NMAB$, where N and M are the size of the state space of the two streams and A and B are the feature dimensions. Such a detailed model comes at the cost of an explosion in the number of parameters. Consequently, parameters estimated from limited training data may be unreliable.

The number of parameters in the model can be reduced by applying conditional independence assumptions between the streams on hidden or observed variables. Viewed from this perspective, there are two broad scenarios for integration – factorial HMMs and coupled HMMs.

4.1 Factorial HMMs

In factorial HMMs, the hidden transitions of one stream, q_t^1, are not directly dependent on the hidden states of the other, q_t^2. However, the interaction between the two streams occurs through a joint observation distribution, $b_t(i,j)$ [16].

$$\pi^c(i) = P(q_0 = i) \tag{10}$$

$$a_{i|j}^c = P(q_t^c = i|q_{t-1}^c = j) \tag{11}$$

$$b_t(i,j) = P(O_t^1, O_t^2|q_t^1 = i, q_t^2 = j) \tag{12}$$

Multi-stream HMMs. Multi-stream HMMs were formulated in the context of multi-stream speech recognition and can be regarded as a variant of factorial HMMs [17, 18]. The joint observation is factored into two components, which are weighted with exponents,λ_c. The exponents are often chosen empirically to avoid the dynamic range of one stream overwhelming the others, similar to the exponential deweighting of acoustic scores in ASR.

$$\pi^c(i) = P(q_0^c = i) \tag{13}$$

$$a_{i|j}^c = P(q_t^c = i|q_{t-1}^c = j) \tag{14}$$

$$b_t(i) = \prod_{c=1}^{C} P(O_t^c|q_t^c = i)^{\lambda_c} \tag{15}$$

The exponent also provides a mechanism to change the reliance on a feature stream based on other measurements such as voicing or signal-to-noise ratio [19, 20]. It can be estimated to minimize classification error (MCE) or maximize entropy of the resulting model [5].

Asynchronous HMMs. Asynchronous HMMs allows the two observation streams to have different lengths [21]. As in factored HMMs, the two streams are observed jointly, but the shorter stream is skipped (or marginalized over) with certain state dependent probabilities, $\eta_t(i)$. The identity of the shorter stream is assumed to be known and constant across training and testing conditions. This formulation allows the model to operate with only one hidden state sequence. As a result, the extra stream does not change the search space and the computational complexity significantly.

$$\pi(i) = P(q_0 = i) \tag{16}$$

$$a_{i|j} = P(q_t = i|q_{t-1} = j) \tag{17}$$

$$\eta_t(i) = P(\tau_t = s|\tau_{t-1} = s-1, q_t = i, O^1_{1:t}, O^2_{1:s}) \tag{18}$$

Additional care needs to be exercised while decoding with this model. The consumption of observations from the two streams should be consistent with the path constraints. Thus, the partial accumulators for Viterbi and forward-backward algorithms can be written as follows.

$$V_{t,s}(i) = \max_{\tau_{1:t-1}, q_{1:t-1}} P(q_t = i, \tau_t = s, O^1_{1:t}, O^2_{1:s}) \tag{19}$$

$$= \max \left\{ \begin{array}{l} \eta_t(i) P(O^1_t, O^2_s|q_t = i) \max_j P(q_t = i|q_{t-1} = j) V_{t-1,s-1}(j) \\ (1 - \eta_t(i)) P(O^1_t|q_t = i) \max_j P(q_t = i|q_{t-1} = j) V_{t-1,s}(j) \end{array} \right. \tag{20}$$

$$\alpha(i,t,s) = \eta_t(i) P(O^1_t, O^2_s|q_t = i) \sum_j P(q_t = i|q_{t-1} = j) V_{t-1,s-1}(j)$$

$$+ (1 - \eta_t(i)) P(O^1_t|q_t = i) \sum_j P(q_t = i|q_{t-1} = j) V_{t-1,s}(j) \tag{21}$$

4.2 Coupled HMMs

In contrast to factorial HMMs, coupled HMMs encode direct dependence of hidden states across streams. The observations of a stream are conditionally independent given the hidden states of that stream [22].

$$\pi^c(i) = P(q^c_0 = i) \tag{22}$$

$$a^c_{i|j,k} = P(q^c_t = i|q^0_{t-1} = j, q^1_{t-1} = k) \tag{23}$$

$$b^c_t(i) = P(O^c_t|q^c_t = i) \tag{24}$$

Coupled HMMs can be generalized to dynamic Bayesian network such as [23], where there may be additional hidden states between the two streams.

Multi-rate HMMs. Multi-rate HMMs can be regarded as a variant of coupled HMMs that also allow streams with different lengths or rates. This is achieved by mapping a state of one stream to two or more states in the other stream. The rate factor may be fixed or variable [24,25].

$$\pi^c(i) = P(q^c_0 = i) \tag{25}$$

$$a^1_{i|j} = P(q^1_t = i|q^1_{t-1} = j) \tag{26}$$

$$a^2_{i|j,k} = P(q^2_t = i|q^2_{t-1} = j, q^1_{\lfloor t/L \rfloor} = k) \tag{27}$$

$$b^c_t(i) = P(O^c_t|q^c_t = i) \tag{28}$$

The hidden states of the second stream, q^2_t, is dependent on that of the first, q^1_t, and not directly dependent in the reverse direction. In the above equations, a fixed decimation ratio, L, is maintained between the two streams. Multi-tape finite state transducers can be designed to efficiently implement this model [26].

5 Kernel Space

An alternative viewpoint decouples the task of integrating different modalities into stream-specific distance measures and a general classifier. The task of designing a classifier can be considerably eased when stream-specific distance measures are kernels, i.e., they obey Mercer conditions [27]. The Mercer conditions require that the distance measure be *positive definite symmetric* (PDS) and can be written as an inner product in some space. In other words, kernels are functions, K, that return the inner product of the images of the operands in some space, Φ. That is, $K(X,Y) =< \Phi(X)\Phi(Y) >$. Choosing K is equivalent to choosing Φ, but provides considerable computational savings when $\Phi(X)$ has a larger dimension than X.

Kernels can be defined on general data types, provided they obey Mercer condition. This allows them to embed sequences, trees, graphs and general structures which may provide natural distance measures for each stream. Additionally, kernels can be easily combined by operations such as sum, product and Kleene closure, and the resultant function is guaranteed to be a kernel. Once the kernels are defined, they can be utilized in any algorithms that operate on inner products such as ridge regression, Fisher discriminant, principal component analysis, canonical correlation analysis, spectral clustering and support vector machines. For further discussion on the properties and utilities of kernels, see [27]. The decoupling of the distance from the classifier allows the exploration of a large number of combinations of the two.

5.1 Affect Recognition

For spoken utterances, we illustrate a general framework for integrating different modalities by exploiting kernels. The example is drawn from the task of recognizing affect from speech, which can be viewed as a classification task, consisting of assigning, out of a fixed set, an affect category (e.g., *joviality*, *anger*, *fear*, or *satisfaction*) to a speech utterance.

Affect detection classifiers can use diverse information sources. Specifically, in this example, we show the integration of two streams of information – “what was said” and “how it was said” [28, 29]. While the former is represented by discrete lexical items, the latter is encoded through a variety of continuous features. Acoustic and prosodic features are comprised of standard Mel frequency cepstral coefficient (MFCC) features, as in automatic speech recognition (ASR), augmented with pitch measurements, as normalized in [30]. When the training data is limited, a simple model for the manner of speaking consists of speech/non-speech HMMs, where the speech portions are tagged with the label of affect. The observation densities in both HMMs are Gaussian-mixture distributions.

A *baseline acoustic-only classifier* can be designed using the MAP decision rule. The MAP rule can be implemented as a Viterbi search over a weighted finite-state transducer representing the state space constraints of the HMM [31]. The state space constraints are shown in Figure 2. The finite-state transducer representing the constrained state space allows insertion of non-speech states.

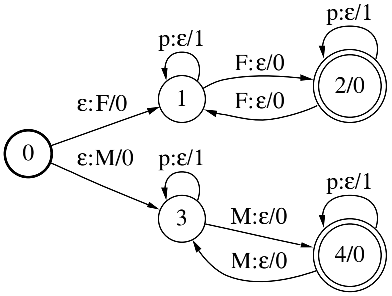


Fig. 2. Search space constrained to allow only paths that contain one type of tagged speech state (M or F), optionally, with non-speech (p) states

The time-synchronous search is pruned using beam widths as in ASR decoding. For experiments with unbalanced test sets, the priors are incorporated as unigram weights in the weighted finite-state transducers (WFSTs).

This baseline classifier can easily be turned into a tokenizer, whose output is a weighted finite-state automata. The labels of the resulting output lattice are tags for affect and weights are the corresponding posterior probabilities. Similary, the lexical information can also be represented as weighted finite-state automata, where the labels correspond to words, hypothesized by an ASR, and the cost encodes their posterior probabilities. Thus, the representation encodes the uncertainty in the ASR hypotheses. After making sure the acoustic and lexical automata do not have overlapping symbols, the two can be combined using a union operation to provide an integrated input for the classifier.

Now, we need a kernel that measures similarity between two weighted finite-state transducers. Extending the notion that two graphs are similar when they share many common n-gram subsequences, a *rational kernel* can be defined over weighted automata [32]. A word lattice L can be viewed as a probability distribution P_L over all strings $s \in \Sigma^*$ with alphabet Σ. Let $|s|_x$ denote the number of occurrences of a sequence x in the string s. The expected count or number of occurrences of an n-gram sequence x in s for the probability distribution P_L is:

$$C_L(x) = \sum_s P_L(s)|s|_x \tag{29}$$

Two lattices output by a speech recognizer can then be viewed as similar when the sum of the product of the expected counts they assign to their common n-gram sequences is sufficiently high. Thus, an n-gram kernel K can be defined for two lattices L_1 and L_2 by:

$$K(L_1, L_2) = \sum_{|x|=n} C_{L_1}(x)\, C_{L_2}(x) \tag{30}$$

K is a rational kernel and it can be computed efficiently. There exists a simple weighted transducer T that can be used to compute $C_{L_1}(x)$ for all n-gram

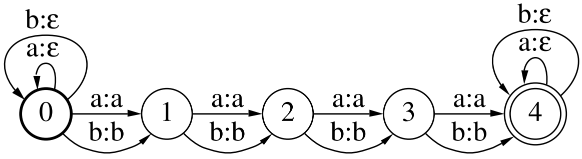


Fig. 3. Weighted transducer T defining a 4-gram kernel

sequences $x \in \Sigma^*$. Figure 3 shows that transducer in the case of 4-gram sequences $(n = 4)$ and for the alphabet $\Sigma = \{a, b\}$.

The general definition of T is:

$$T = (\Sigma \times \{\epsilon\})^* \ (\sum_{a \in \Sigma} \{a\} \times \{a\})^n \ (\Sigma \times \{\epsilon\})^* \tag{31}$$

The kernel K can thus be written in terms of the weighted transducer T as:

$$K(L_1, L_2) = w[(L_1 \circ (T \circ T^{-1}) \circ L_2)] \tag{32}$$

where $w[T]$ represents the sum of the weights of all paths of T. This shows that K is a rational kernel whose associated weighted transducer is $T \circ T^{-1}$ and thus that it is *positive definite symmetric* (PDS), or equivalently that it satisfies Mercer's condition [32]. This condition guarantees the convergence to a global optimum of discriminant classification algorithms such as SVMs. K can be further used to construct other families of PDS kernels, e.g., *polynomial kernels* of degree p defined by $(K + a)^p$. Thus, for a given n, an *n-gram kernel* can be defined as the sum of the k-gram kernels, $k = 1, \ldots, n$.

5.2 Empirical Results

On a binary affect classification task, the performance of n-gram kernels has been compared with several popular classifiers. The comparison was carried out on real-world data extracted from a deployed customer-care system, the AT&T "How May I Help You" system (HMIHY 0300) [28,29].

The corpus consisted of 5147 utterances from 1854 speakers, with an average utterance length of 15 words. The affect for each utterance was grouped into two categories – negative and non-negative, in a manner similar to [2]. While creating the reference utterances, the human annotators had the advantage of knowing the context beyond the utterance being labeled. A subset of 448 utterance was used for testing on which two human labelers were in full agreement. For further details on the corpus and the consistency of annotations, see [28]. The automatic transcripts or lattices were generated by an ASR system, whose word error rate on the corpus was 37.8%.

Classifiers using acoustic and lexical features were compared, including the baseline acoustic-only classifier, described earlier. A MAP classifier based on an interpolated language model (LM) was trained, using only lexical features,

Table 1. Comparison of classifiers on the task of binary affect classification

	Classifier	Accuracy
(a)	MAP w/ acoustic HMMs	76.8
(b)	MAP w/ interpolated LM ($\lambda = 0.8$)	70.1
(c)	MAP w/ MI-filtered LM	78.8
(d)	SVM w/ n-gram kernels on ASR one-best hypothesis	79.9
(e)	SVM w/ n-gram kernels on ASR word lattices	80.6
(f)	SVM w/ n-gram kernels on ASR word lattices and acoustic lattices from (a)	81.7

as in [33]. The interpolation was performed to smooth the class-specific LM. A similar language model-based classifier was also evaluated, where the lexical features were selected using mutual information (MI) and the model was not smoothed [2]. These three classifiers from the literature, specifically designed for affect detection task, were compared with an n-gram kernel-based classifier. The n-gram kernels made it possible to exploit both acoustic and lexical features seamlessly using a support vector machine.

Table 1 summarizes the results of the comparison. The MI-based feature-selection classifier yielded significantly better results than the interpolated language model classifier: an improvement of the classification accuracy by 7.7% absolute. This suggests that feature selection plays a crucial role for affect detection with a classifier based on an n-gram model since that is the key difference between the interpolated language model classifier and the MI-based feature-selection classifier. This result was obtained when infrequent words below 8 occurrences were ignored in computing mutual information, and when the size of the selected vocabulary was 350 words. While selected vocabulary included words such as *disconnect*, *good*, *yes*, *correct* and *cancel* that could be viewed by humans as indicative of an affect category for the corpus used, it also contained a number of seemingly uninformative words such as *hi*, *couple*, *see* and *name*.

The classifier based on rational kernels combined with SVMs outperformed the previous two classifiers with an accuracy gain of 1.1% absolute over the best one of them. This was further improved by using the full word lattices generated by the speech recognition system (80.6% accuracy), which is only about 1% short of the accuracy that can be attained with reference transcripts (81.7%). The best result was obtained with an n-gram kernel of order four (n=4). Applying the n-gram kernel to the combined acoustic and lexical weighted finite-state automata improves the accuracy further to 81.7%.

The design of the kernel-based classifier does not rely on the definition of a specific subset of words since that can introduce a bias. Moreover, the generalization bounds for SVMs do not depend on the dimension of the feature space. The n-gram kernels have also been found to be useful in other utterance classification tasks such as call classification and language identification [34,13]. Other candidates that operate on discrete sequences include convolution kernels and spectrum kernels [35,36]. Similarly, Fisher kernels and continuous rational kernels provide alternatives for measuring distances in acoustic space [37,38]. Although it is possible, heterogenous sources of information, such as the state of the dialog in a spoken-dialog system, can not be easily incorporated in traditional HMM-based classifiers [39]. Such information can easily be incorporated in a large-margin classifier based on kernels.

So far, research on integration has been limited to combining kernels using simple operators such as sum, product and Kleene closure. More complex parametric combinations of kernels with learned weights have been formulated in the context of genomic data fusion and they are applicable for detecting speaker characteristics as well [40].

6 Discussion

Information from multiple sources can be integrated in feature space, model space, score space or in kernel space. The benefit of integrating multiple sources depends on how closely the streams are correlated. When the observation rates are comparable, the integration can be performed effectively in feature space without incurring too much additional computation cost per stream. A number of HMM variants can model mutliple streams effectively and they include coupled HMMs, factorial HMMs, and dynamic Bayesian networks. The optimal choice depends on factors such as the need to model asynchrony, the hidden relation between the streams,and the available amount of training data. The scores from base-experts can improve performace when they are uncorrelated and have comparable performances. Theoretical analysis demonstrates the trade-offs when the scores are correlated or the performances differ significantly. Emiprical evalauation in several studies have shown the importance of score-normalization such as T-norm before fusion. Kernels decouple the task of measuring distances between objects in a stream from the classifier. This makes it easy to integrating streams containing diverse structures. Using an example from affect recognition, we show how acoustic and lexical features can be easily integrated to provide performance superior to other popular algorithms developed specifically for the task. This framework is general enough to be applicable for detecting other speaker traits and has already been shown to be effective in language identification [13].

References

1. Gorin, A.L., Abella, A., Alonso, T., Riccardi, G., Wright, J.H.: Automated natural spoken dialog. IEEE Computer Magazine 35(4), 51–56 (2002)
2. Lee, C.M., Narayanan, S.S., Pieraccini, R.: Combining acoustic and language information for emotion recognition. In: Proc. Int'l Conference on Spoken Language Processing (2002)

3. Kumar, N., Andreou, A.G.: Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. Speech Communication (4), 283–297 (1998)
4. Saon, G., Padmanabhan, M., Gopinath, R., Chen, S.: Maximum likelihood discriminant feature spaces. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1129–1132 (2000)
5. Gravier, G., Axelrod, S., Potamianos, G., Neti, C.: Maximum entropy and MCE based HMM stream weight estimation for audio-visual asr. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 853–856 (2002)
6. Chetty, G., Wagner, M.: Audio-visual multimodal fusion for biometric person authentication and liveness verification. In: Proc. ACM NICTA-HCSNet Multimodal User Interaction Workshop, pp. 17–24 (2006)
7. Gunes, H., Piccardi, M.: Affect recognition from face and body: Early fusion vs. late fusion. In: Proc. of IEEE Int'l Conference on Systems, Man and Cybernetics, pp. 3437–3443 (2005)
8. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text independent speaker verification system. In: Digital Signal Processing, pp. 42–54 (2000)
9. Mariethoz, J., Bengio, S.: A unified framework for score normalization techniques applied to text independent speaker verification. IEEE Signal Processing Letters, 532–535 (1997)
10. Poh, N., Bengio, S.: EER of fixed and trainable fusion classifiers: A theoretical study with application to biometric authentication tasks. In: Multiple classifier systems, pp. 74–85 (2005)
11. Hong, L., Jain, A.K., Pankanti, S.: Can mutli-biometrics improve performance. In: Proc. IEEE Workshop on Automatic Identification Advanced Technologies (WAIAT), pp. 59–64 (1999)
12. Ferrer, L., Sonmez, K., Kajarekar, S.: Class-dependent score combination for speaker recognition. In: Proc. European Conference on Speech Communication and Technology, pp. 2173–2176 (2005)
13. White, C., Shafran, I., luc Gauvain, J.: Discriminative classifiers for language recognition. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 213–216 (2006)
14. Poh, N., Bengio, S.: How do correlation and variance of base-experts affect fusion in biometric authentication tasks? In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4384–4396 (2005)
15. Trees, H.L.V.: Detection, Estimation, and Modulation Theory: Part I. Wiley, Chichester (2001)
16. Nefian, A.V., Liang, L., Pi, X., Liu, X., Murphy, K.: Dynamic Bayesian networks for audio-visual speech recognition. EURASIP Journal of Applied Signal Processing (11), 1–15 (2002)
17. Poh, N., Bengio, S.: Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks? In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2004)
18. Poh, N., Bengio, S.: Noise-robust multi-stream fusion for text-independent speaker authentication. In: The Speaker and Recognition Workshop (2004)
19. Glotin, H., Vergyri, D., Neti, C., Potamianos, G., Luettin, J.: Weighting schemes for audio-visual fusion in speech recognition. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 173–176 (2001)

20. Heckman, M., Berthommier, F., Kroschel, K.: Noisy adaptive stream weighting in audio-visual speech recognition. Journal of Applied Signal Processing, Special Issue of Audio-Visual Signal Processing, 1260–1273 (2002)
21. Bengio, S.: An asynchronous hidden Markov model for audio-visual speech recognition. In: Neural Information Processing System (NIPS) (2002)
22. Nefian, A.V., Liang, L., Fu, T., Liu, X.: A Bayesian approach to audio visual speaker identification. In: IEEE International Conference on Audio-and Video-based Biometric Person Authentication (June 2003)
23. Gowdy, J.N., Subramanya, A., Bartels, C., Bilmes, J.: Dbn-based multi-stream models for audio-visual speech recognition. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2004)
24. Cetin, O., Ostendorf, M.: Multi-rate hidden Markov models and their application to machining tool-wear classification. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (2004)
25. Cetin, O., Ostendorf, M.: Multi-rate and variable-rate modeling of speech at phone and syllable time scales. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 665–668 (2005)
26. Hetherington, I.L., Han, S., Glass, J.R.: Flexible multi-stream framework for speech recognition using multi-tape finite-state transducers. In: ICASSP. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 417–420. IEEE Computer Society Press, Los Alamitos (2006)
27. Schölkopf, B., Burges, C.J.C., Smola, A.J.: Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge (1999)
28. Shafran, I., Riley, M., Mohri, M.: Voice signatures. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 31–36 (2003)
29. Shafran, I., Mohri, M.: A comparison of classifiers for detecting emotion from speech. In: ICASSP. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 341–344. IEEE Computer Society Press, Los Alamitos (2005)
30. Ljolje, A.: Speech recognition using fundamental frequency and voicing in acoustic modeling. In: ICASSP. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE Computer Society Press, Los Alamitos (2002)
31. Mohri, M., Pereira, F.C.N., Riley, M.: Weighted finite-state transducers in speech recognition. Computer Speech and Language 16(1), 69–88 (2002)
32. Cortes, C., Haffner, P., Mohri, M.: Rational kernels: Theory and algorithms. Journal of Machine Learning Research (JMLR), 1035–1062 (2004)
33. Devillers, L., Vasilescu, I., Lamel, L.: Emotion detection in task-oriented dialog corpus. In: Proc. IEEE Int'l Conference on Multimedia, IEEE Computer Society Press, Los Alamitos (2003)
34. Cortes, C., Haffner, P., Mohri, M.: Lattice kernels for spoken-dialog classification. In: ICASSP. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 628–631. IEEE Computer Society Press, Los Alamitos (2003)
35. Haussler, D.: Convolution kernels on discrete structures. In: UC Santa Cruz Technical Report UCS-CRL-99-10 (1999)
36. Leslie, C., Eskin, E., Cohen, A., Weston, J., Noble, W.S.: Mismatch string kernels for svm protein classification. In: Proc. conference on Advances in Neural Information Processing Systems (NIPS) (2003)
37. Jaakkola, T.S., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Proc. conference on Advances in Neural Information Processing Systems (NIPS), pp. 487–493 (1999)

38. Layton, M., Gales, M.: Acoustic modeling using continuous rational kernels. Journal of VLSI Signal Processing (2007)
39. Reg, J.M., Murphy, K.P., Fieguth, P.W.: Vision-based speaker detection using Bayesian networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 110–116. IEEE Computer Society Press, Los Alamitos (1999)
40. Lanckriet, G.R.G., Bie, T.D., Cristianini, N., Jordan, M.I., Stafford, W.: A statistical framework for genomic data fusion. Bioinformatics (16), 2626–2635 (2004)

Evaluations of Automatic Speaker Classification Systems

Alvin F. Martin

National Institute of Standards and Technology
100 Bureau Drive Stop 8940
Gaithersburg, MD 20899-8940
alvin.martin@nist.gov

Abstract. The annual NIST Speaker Recognition Evaluations (SREs) from 1996 to 2006 have been internationally recognized as the leading source or performance evaluation of research systems in the speaker classification field. We discuss how these evaluations have developed and been conducted and the performance measures used. We consider the key factors that have been studied for their effect on performance, including training and test durations, channel variability, and speaker variability. We examine the extent to which progress has been observed in state-of-the-art performance. We also consider how the technology has changed over the past decade, other evaluations that have been conducted or planned, and where the field may be headed in the future.

Keywords: speaker recognition, speaker detection, speaker classification, speaker identification, speaker evaluation, NIST evaluations, NIST SRE, DET curves.

1 The Challenge

We consider the challenge of developing effective procedures for testing and evaluation of automatic speaker classification systems. This is a developing field of technology, and one with significant commercial potential. Such a field does not readily lend itself to objective technical evaluation, particularly in its early development.

Speaker recognition has developed somewhat in the shadow of the field of automatic speech recognition, where the objective is to transcribe the words (and perhaps understand their meaning as well) of a particular, or preferably of any, speaker. The development of evaluation in this area may be instructive.

In the 1970's and 1980's a number of speech recognition companies were offering products and anticipating a growing market for their offerings. And how good were their products. Each company recognized the need to quantify their performance and, invariably, each reported a correct word recognition rate in the range of 95–100 %. Yet potential users of the technology soon came to realize that in real world application scenarios of interest to them, they were likely to find far lower word recognition rates.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 313–329, 2007.

Aside from telling outright lies about their performance, which may have occurred, each vendor would collect test data under ideal conditions for the speech recognition application of interest to them. And each would make very sure that a high recognition rate was achieved with this data; they couldn't hope to compete if they reported otherwise.

Potential users of the technology were in a difficult position. Each vendor claimed superior performance, and presumably had achieved it for its own proprietary data. But since the data was not shared, the performance of the different vendors' systems could not be meaningfully compared. Insightful users would recognize that with their own data and their own application scenarios they would not achieve the kind of results being reported, but until they acquired systems and used them in-house, they would not know which system was likely to be best for them, and how well it might do. This made it difficult to decide if the new technology would be cost effective compared with existing procedures or competing technologies.

George Doddington perhaps made the first efforts to test the performance of then existing speech recognizers on a common database [1]. He collected a database of spoken digits at Texas Instruments and invited vendors to supply a version of their systems to be used in in-house testing.

Soon after that, interest in such evaluation of speech recognition technology was taken up at the National Institute of Standards, which later became the National Institute of Standards and Technology (NIST), in Gaithersburg, Maryland. NIST has conducted a series of evaluations of speech recognition on different types of speech data, concentrating in recent years on broadcast news and conversational telephone speech. These evaluations have typically initially reported rather high word error rates, which have been reduced as a particular type of evaluation has been continued over several years. Indeed, when such error rates have been reduced below 10 % or so, NIST has shifted its evaluation focus to more difficult types of speech.

Speaker recognition lacked such independent evaluation into the 1990's. Each research site would choose its own data to use. This sometimes involved the use of proprietary corpora not available to other systems. But at least a few common speech corpora were becoming available, and a popular choice was the TIMIT Corpus [2]. This was a corpus of high quality phonetically transcribed speech including multiple sessions from a number of speakers (as needed for speaker recognition) that had been collected at Texas Instruments.

In 1994 the first of series of international workshop on speaker recognition was held in Martigny, Switzerland. It was followed by a similar workshop in Avignon, France in 1998. The third such workshop, in Crete in 2001 was dubbed "2001: A Speaker Odyssey". The subsequent workshops, in Toledo Spain in 2004 and San Juan, Puerto Rico in 2006 have continued the Speaker Odyssey name. The first two pre-Odyssey workshops, however, were dominated by researchers reporting results, generally very good results, on proprietary data sets or on the TIMIT data. This was viewed as frustrating by those who wanted to see meaningful performance comparisons on more real-world type data.

It was in this context that in 1996 NIST initiated its series of annual speaker recognition evaluations. These have concentrated on the use of conversational telephone data from corpora collected by the Linguistic Data Consortium (LDC) [20]. The central speaker detection task has remained the same throughout the evaluations. A system is given speech data (training data) known to be from a given target speaker, and given a separate test segment of speech data. It must then determine whether the test data was spoken by the target speaker. An evaluation test consists of a (long) sequence of trials of this type. For each trial, the given target speaker, defined by the training data, is the only speaker "known" to the system.

The NIST speaker recognition evaluations are described in greater detail in further sections of this chapter. Their history encapsulates the progress and problems encountered in this area over the past decade. They document the level of performance of state-of-the-art systems for speaker detection involving text independent conversational speech transmitted over public telephone channels and the degree of performance improvement over the period. But the evaluations have changed over the years, with the variety of test conditions increased, and the problems addressed sometimes made harder due to changes in general telephone technology and to greater interest in more challenging conditions as the technology has improved.

2 The NIST Evaluations

As noted, the basic task in all of the NIST speaker recognition evaluation has been speaker detection. This means that each test consists of a sequence of trials, where each trial is defined by a target speaker and a test segment of speech. The target speakers are defined by training data provided for each such speaker. This training data may consist of one or several speech segments guaranteed to contain speech of the speaker. The test segment contains unknown speech. The system must determine if in fact this speech was spoken by the target.

For each trial the system must supply both a hard decision ('T' or 'F') in answer to this question. In addition a likelihood score is required that quantifies the decision. Higher scores should indicate greater probability that the test speech is by the target.

Trials where the target is speaking, those for which the correct decision is 'T', are target trials. Trials where the target is not speaking are non-target (or impostor) trials. System errors in target trials are misses, while those in non-target trials are false alarms. Thus a system has two basic error rates, the percentage of target trials that are misses (miss rate) and the percentage of non-target trials that are false alarms (false alarm rate).

The basic error metric in the NIST evaluations has been a linear combination of these two rates that has been denoted C$_{DET}$. It is defined as

$$C_{DET} = Norm_{Fact} * ((C_{Miss} * P_{Miss|Target} * P_{Target}) + (C_{FA} * P_{FA|NonTarget})) \quad (1)$$

Table 1. The cost function that has served as the primary metric in the NIST evaluations is based on assigned relative costs for each miss and each false alarm and an assumed target richness chosen for possible applications of interest

Cost of a miss	$C_{Miss} = 10$
Cost of a false alarm	$C_{FA} = 1$
Probability of a target	$P_{Target} = 0.01$
Probability of a non-target	$P_{Non-Target} = 1 - P_{Target} = 0.99$
Normalization factor ($Norm_{Fact}$) is defined to make 1.0 the score of a knowledge-free system that always decides "False" It detection cost $C_{default} = 10 * 100\ \% * 0.01 + 1 * 0.99 = 0.1$ So $Norm_{Fact} = 10$	

C$_{DET}$ can be viewed as a cost function based on assigned costs for misses and false alarms and an assumed target richness. But the assigned cost and assumed target richness are essentially arbitrarily chosen parameters. (Note that PTarget need not, and does not, correspond to the actual percentage of target trials in the evaluation test sets.) The values selected are believed to be reasonable ones for some applications of interest. The low target richness may be particularly applicable to text-independent applications. For some other applications a higher value may be appropriate, but so may a higher relative cost for false alarms, so these may cancel each other out to some extent.

There has, however, been recent work on developing a more application independent type of metric that allows after evaluation examination of performance for any specific parameters of interest. This requires that the confidence scores provided be actual probabilities, or better, actual log likelihood ratios. The metric Cllr, and the ways it may be utilized, are discussed in [3]. Such scores, and the use of this metric, was an option for participants in the 2006 evaluation and will probably receive attention in future evaluations.

3 Evaluation Parameters

Having defined the evaluation task, choices need to be made about the data to be collected and utilized. Evaluations are heavily dependent upon the collection of appropriate and sufficient data. Each evaluation test is defined by a sequence of trials, and time and cost for collection is likely to be the limiting factor determining the number of trials to be included.

The most basic evaluation parameters defining the trials are the duration of the training and test speech segments, and the timing of their collection. The training data for each target speaker may be collected in one or more different sessions. The amount of training data (duration of training speech) is typically the same or greater than the amount of test speech used in a given trial. (At least for single session training, the training and test speech used in each trial may be viewed as playing symmetric roles.)

NIST has used a speech activity detector to determine the approximate durations of speech in training and test segments. In earlier evaluations considerable effort was made to be fairly precise about the speech durations in each trial. In later evaluations interest shifted in large part to using longer speech durations (in particular whole conversation sides) with less precision. Also in earlier evaluations the training and test segments consisted of concatenated segments of speech (as determined by the speech activity detector) with non-speech portions of the signal excised. In later evaluations continuous segments without excision were used, though estimates were still made of actual speech duration.

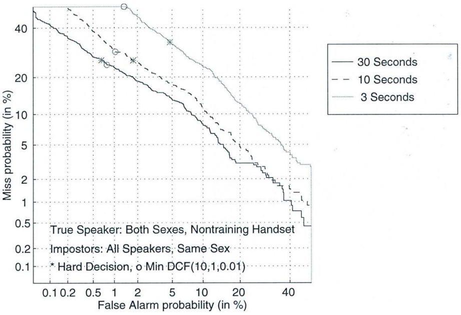


Fig. 1. Effect of test segment duration on performance, fixed durations

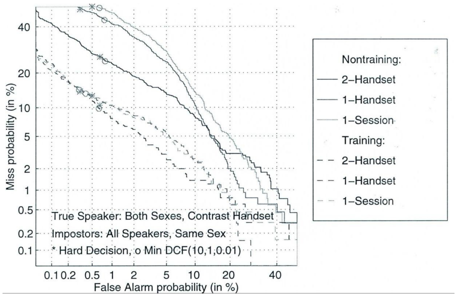


Fig. 2. Effect of match or non-match of training and test handsets, and of multiple training sessions with same or different handsets

Figures 1 and 3 show the effects of test segment duration on performance for a typical system in three different NIST evaluations. In all cases, we see the expected result of better performance with longer durations. In the early evaluations (Figure 1) the test segments had fixed approximate speech durations of 3, 10, or 30 seconds each. Later variable durations of up to a minute were used (Figure 3). Here it may be noted that the only strong effect on performance is seen for durations of less than 15 seconds.

With respect to training data, early NIST evaluations examined the effect of the number of training sessions, their diversity with respect to the telephone handsets used, and their relationship to the test segment handset for target trials. Figure 2 shows results for a system both where the test handset was the same as (one of) the training handsets and where it was not. (The duration of training speech is approximately the same for all six DET curves.) Most notable is the better performance when the same handset in used in training and test. (This is for target trials only; nontarget trials invariably involve different handsets.) Subsequent evaluations have emphasized different handsets, at least for landline transmission data. Examining the three curves where the test handset is different, it may be seen that having two training sessions yields better performance

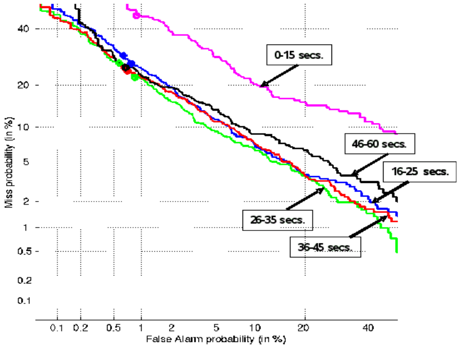


Fig. 3. Effect of test segment duration on performance, variable durations

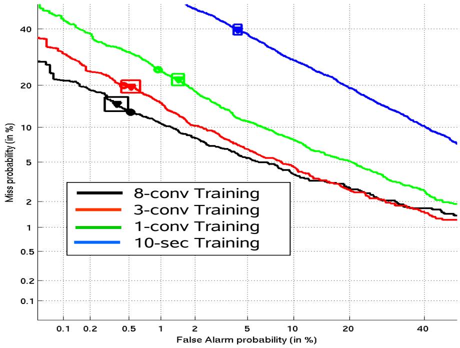


Fig. 4. Effects of varying amounts of training data on performance, all using the same handset

than one, and having these two sessions use different handsets further improves performance.

More recent evaluations have concentrated the effects of offering much larger amounts of training data. In figure 4 the curves show results when training consisted of 1, 3, or 8 whole conversation sides (each averaging about 2.5 minutes of speech). Also included is a 10-second training condition, which certainly remains of interest, particularly for some commercial applications. (In all cases the test segments consist of one conversation side of speech data.) The advantage of increased training data, where applications will support this is seen. It may also be seen that there is still a long way to go to achieve equivalent performance with very short segments of training data.

4 Channel Variability

Speaker recognition performance may be greatly enhanced by using a constant high-quality wideband channel, but the primary application interest of the technology is in its use over telephone channels, and perhaps over various types of differing and varying quality microphone channels. Thus the handling of channel variability is one of the key challenges to be overcome by the system designer and a key factor to be considered by the system evaluator.

The NIST evaluations, as noted previously, have until the last few years concentrated on telephone channels. But the nature of public telephone channels in the United States has changed considerably in recent years. The quality of traditional landline channels has improved. A decade or so ago carbon-button and electret microphones were both common in telephone handsets, and the early NIST evaluations considered the effects of handset microphone type on performance. Carbon-button microphones have become less common in recent years, but a bigger change has been the widespread use of cellular phones in the U.S. in recent years. Thus the recent evaluations have examined the performance effects of cellular as opposed to landline transmission.

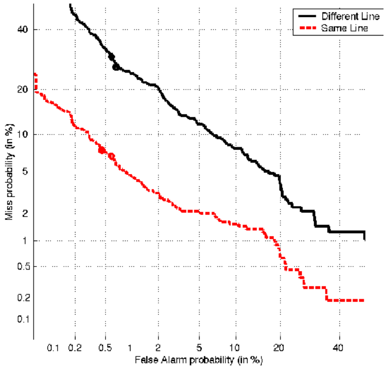


Fig. 5. Effect of same or different phone line (and presumably handset) in training and test

Figure 5, involving one system from an early evaluation, shows the effect of using a fixed or a variable telephone line, and presumably handset, in target trials. Clearly, having the same handset used in each speaker's training and test segments makes the problem far easier. But the use of caller id is simpler and more effective. (Note that non-target trials invariably involve the use of different phone lines and handsets unless special arrangements are made to do otherwise.) The situation of practical interest is where training and test phone lines differ, and later evaluations focused only on such cases, as least for landline trials.

Figure 6 shows the effects of microphone handset types for five different systems in an early evaluation. Two different effects are convolved to different overall effect in the different systems. In general performance is better with electret than with carbon-button handsets (the fourth system is something of an exception). But performance is also generally superior when the training and test handset types are the same. So the black curves generally show relatively good performance, and the red and blue curves relatively poor performance, while the green curves (all carbon-button) show variable performance.

Figure 7, from a recent evaluation, presents a similar type of plot for one system showing the effect of cellular or landline transmission in training and test. Perhaps not surprisingly, performance appears to be considerably better for landline data.

The most recent NIST evaluations have included some telephone conversations where the speech of one of the conversants was simultaneously recorded over a (cellular) telephone channel and over eight different microphone channels. Figure 8 shows performance results for one system involving the nine different representations of the same test conversations. (The training is fixed and recorded over a telephone channel.) The main point to be noted is that the telephone results are far superior to those of all the microphones. It should be noted that this was the first such NIST evaluation, and that cross-channel performance may be expected to improve in future evaluations.

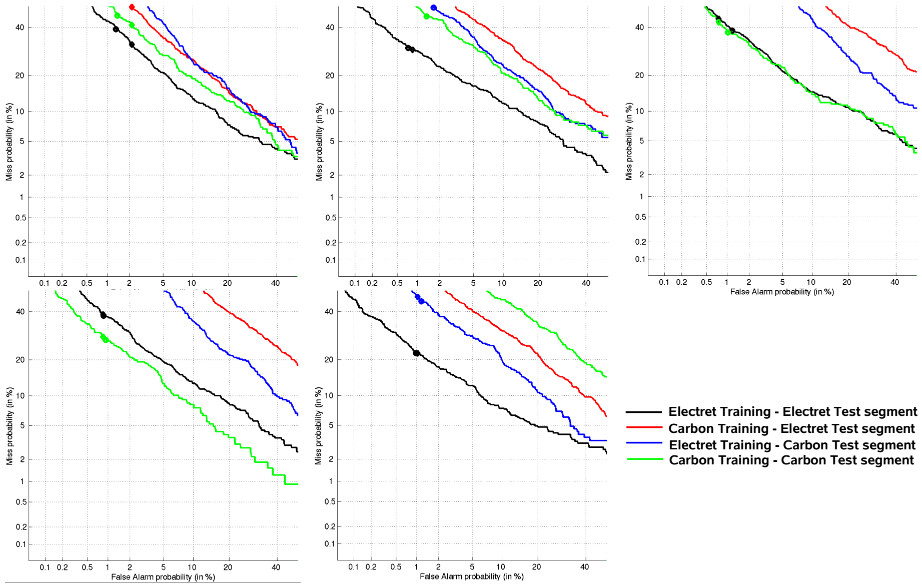
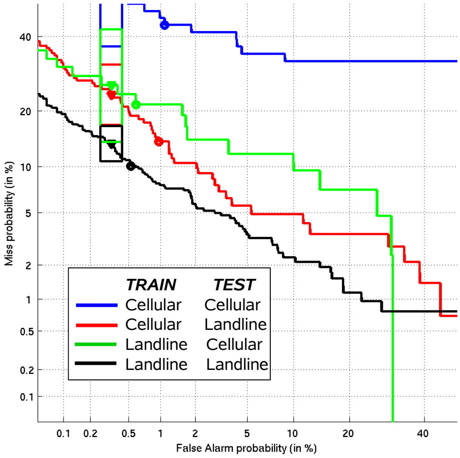


Fig. 6. Effect of using different combinations of handsets with carbon-button or electret microphones in training and test. These effects vary for the five different systems shown.



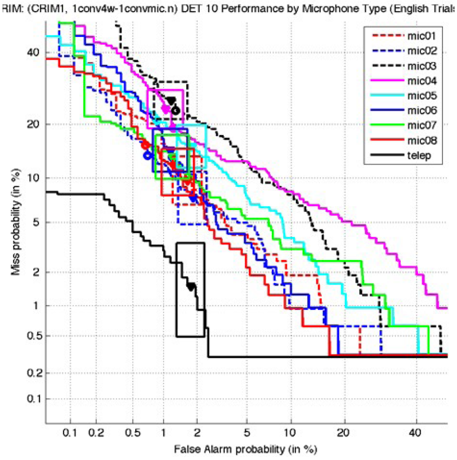


Fig. 7. Effects of using cellular or landline data in training and test on performance

Fig. 8. Effect on performance of using any of eight different microphone channels or telephone data in the test segment, with training always on telephone data

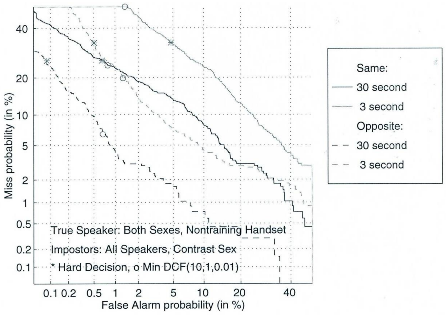


Fig. 9. Performance by whether non-target trials involve speakers of same or opposite sex for two durations

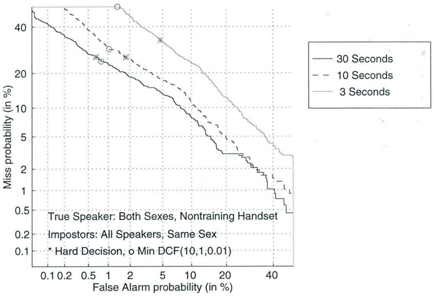


Fig. 10. Performance by language (English or non-English) in the training and test data for all trials

5 Speaker Variability

Variability between different speakers, a key problem for speaker independent word recognition, is the characteristic that makes speaker classification technology possible. A major division of speakers into two classes is by sex. Figure 9, from the first NIST evaluation, shows performance (for both 30-second and 3-second test segments) when the non-target trials involve speakers of same sex or of opposite sex. Since gender recognition tends to be highly accurate, the results are as might be expected. Including cross-sex trials in evaluations is one way to show better results. Subsequent NIST evaluations have excluded such trials.

The variability of individual speakers, on the other hand, is a major challenge to speaker classification technology. Speaker consistency is a highly desirable attribute for successful recognition, but in the real world speakers often do not maintain consistency for a variety of reasons. Voices change because of health problems (such as colds) and because of stress and emotional conditions. And in the long run they change as people age.

Measuring speaker variability in evaluation is not easy to do, as people cannot readily be instructed to demonstrate variability in their voices on demand. Creating stress conditions is not something that committees on the use of human subjects look fondly upon. And data collection sessions far enough apart in time to reveal the effects of aging are not readily arranged.

Figure 11 explores one way of examining the effect of speaker variation. For one system in a particular evaluation, we estimated the speaker's average pitch in the training and in the test data. The figure shows the large performance difference between the quarter of the target trials where the speaker was most consistent in average pitch between training and test and the quarter of the trials (perhaps involving one session with a cold) where the speaker had the greatest relative pitch differences.

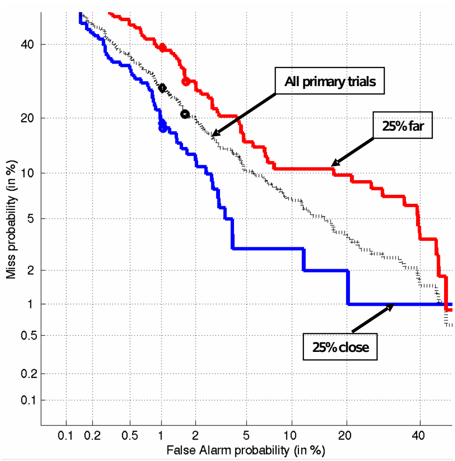


Fig. 11. Performance by relative closeness of training and test average pitch differences in target (same-speaker) trials

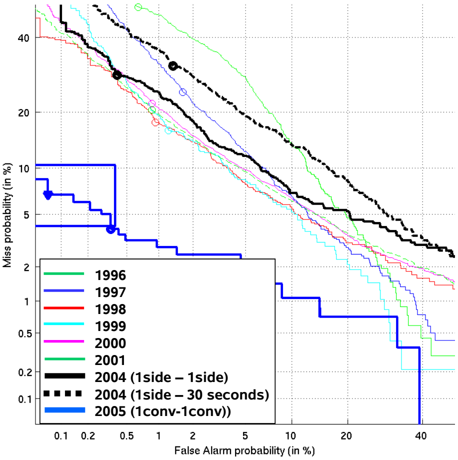


Fig. 12. Best-system performance history for landline trials 1996-2005. Prior to 2004 training generally consisted of two minutes of speech, and test of 30 seconds (an average of 30 seconds) of speech.

Another, more controllable way in which a speaker may vary, is in language. In recent NIST evaluations a number of bilingual speakers (of English and another language) were included. Figure 10 show performance results based on whether the training and the test speech were in English (E) or a non-English language (N). Clearly language consistency matters, at least for this system and others tested in this evaluation.

6 Measuring Progress

The primary purpose of evaluation of research systems in a developing field of technology such as speaker recognition is to encourage progress in the field. It is therefore of key concern to determine the degree of progress that has occurred over a period of years.

But there are difficulties in doing this. It can be hard to ensure that different test sets present equal task difficulty, even if they are chosen in substantially the same way. But evaluations do not remain constant from year to year. They change to reflect the changing interests and priorities of those who are sponsoring and organizing the evaluations. Improving system performance may be a reason to choose to make the task harder, thus appearing to suppress further performance improvement. And in the case of speaker recognition over telephone lines, changes in the public phone system affect the evaluation results. In particular, the increasing use of cellular telephones, which we have seen have an adverse effect on performance, has made comparisons more difficult.

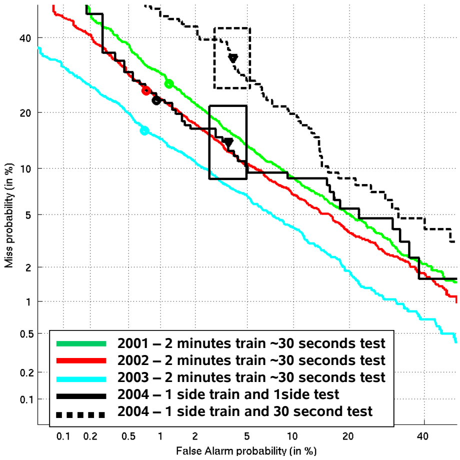


Fig. 13. Best-system performance history for cellular trials 2001-2004

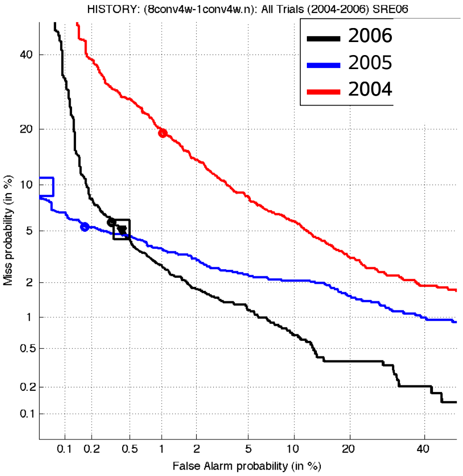


Fig. 14. Best-system performance history for eight conversation training (extended training data) 2004-2006

The NIST evaluations may be divided into three phases. From 1996 to 2001 the data was primarily landline, selected from the conversations of the several Switchboard corpora collected by the LDC. The primary test condition involved two minutes of training and thirty second test segments (variable averaging thirty seconds in 2001). For 2001 to 2003 testing similar but on the Switchboard cellular corpora (both landline and cellular were used in 2001). Since 2004 the LDC Mixer Corpus [16,17] has been used, with a different collection protocol, a mix of landline and cellular data, and some calls in languages other than English. Table 2 summarizes these three phases in the data used in the evaluations.

For each evaluation an effort is made to assess the overall level of performance improvement (or the lack thereof) between the best performing systems of the current and preceding years, matching test conditions of interest to the extent possible, and NIST has regularly sought to do this. Figures 12 – 15 attempt to suggest the degrees of progress that have been observed over the course of the NIST evaluations.

Figure 12 presents best system results on trials involving landline data between 1996 and 2005. (2002 and 2003 are omitted because the great majority of trials those years involved cellular data.) The results tend to divide between those for years prior to 2002 and those for years after 2003. For the earlier years, there was clear progress from 1996 to 1998, and then somewhat of a plateau until 2001. The Mixer data used starting in 2004 resulted in an apparent adverse performance effect, even with increased training and test durations. Two different test conditions in 2004 show better performance with longer duration test data, as expected. The number of all landline trials was limited in 2005, but a considerable performance improvement is observed.

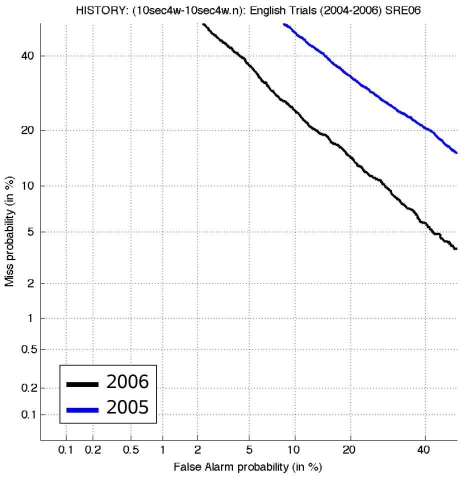


Fig. 15. Best-system performance history for short duration (10-second) training and test data trials 2005-2006

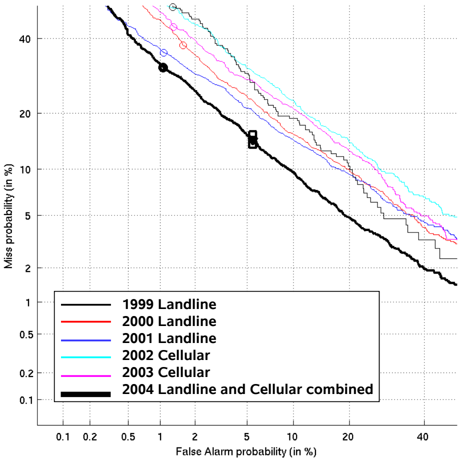


Fig. 16. Best-system performance history for two-speaker training and test trials 1999-2004

Figure 13 gives a similar history plot of best performing systems on cellular data from 2001 to 2004. There is general progress from 2001 to 2003. For 2004 there are two different test conditions, both of which are different from the conditions of the preceding years, and the number of cellular trials was smaller than before, making the curves less smooth. Moreover, 2004 was the first year in which Mixer data was used, adapting to which may have been a challenge for systems. In any case, the best 2004 performance did not match the best of 2003.

Since 2001, when George Doddington demonstrated the potential gains from exploiting high level idiolectal type information for speaker recognition from longer durations of speech [4,5], a major focus of the evaluations has been on the level of performance that may be achieved by the use of "extended duration" speech, particularly for training. Recent evaluations have included a condition on training on eight different conversation sides of each target (averaging about 2.5 minutes of speech each, while testing on single whole conversation sides. The previous discussion on duration has noted the effect of extended training on performance. Figure 14 shows results for the best performing system for the past three years. Results for earlier years are not comparable, because only with the Mixer data of these recent years was it possible to assure that the test handsets were distinct from those used in training. There was a considerable improvement in 2005 over 2004, and a more mixed result in 2006 compared with 2005. It is believed that the shape of the 2006 DET curve may be due to the presence of more trials involving non-English speech in 2006 than in 2005. This is another confounding factor in judging performance improvement.

Short duration training and test has been included in the NIST evaluations largely by popular demand. While performance is much inferior when training

Table 2. Corpora used and primary tests in three phases of the NIST SREs

1996-2001	Switchboard-1, Switchboard-2 Phases I, II, III	2-minute training (1 or 2 sessions), 3, 10, 30 second test segments, variable duration test segments in 2001 (averaging 30 sec.)
2001-2003	Switchboard Cellular Parts 1, 2	2-minute single session training, variable (15-45 sec.) duration test segments
2004-2006	Mixer (including some non-English conversations and multi-channel microphone data in 2005-2006	8, 3, or 1 conversation side training, 1 conversation side test segments (also 10 sec. training and test)

and test are limited to ten second speech durations, there is considerable commercial potential in being able to achieve good results in this case. Figure 15 shows that considerable improvement was seen in the best evaluation systems between 2005 and 2006, but that there remains a long way to go to achieve performance acceptable for most applications.

7 Multi-speaker

Speaker recognition in a multi-speaker environment, a subject perhaps outside the mainstream of work in speaker classification, has been a part of the recent NIST evaluations. They have focused on the summed channel situation where the input consists of the combined two channels of a phone conversation between two persons. The target speaker training data may be single channel, but the recent NIST evaluations have included a training condition consisting of three conversations involving the target speaker with three different people, requiring systems to find and segment the target speech in the training conversations.

Figure 16 shows a history plot of best systems for the two-speaker condition involving both landline and cellular data from 1999 to 2004. It shows a rather satisfying record of improvement for each type, with the best results occurring in 2004 on data involving both landline and cellular calls.

Earlier NIST evaluations also had tasks specifically for speaker segmentation and tracking within multiple speaker speech [23]. This kind of task has since been pursued in other in other evaluations, including the speaker diarization task of the NIST Rich Transcription Meeting Room evaluations [6,7,8] and the internationally (U.S. and Europe) based CLEAR (Classification of Events, Activities, and Relationships) [9] evaluations.

8 Other Evaluations

The author's perspective is oriented toward the NIST evaluations, and these have certainly assumed the leading role in the field to date, but there have been other evaluations, and there will undoubtedly be further ones.

In 2003 TNO, a Dutch applied scientific research organization, sponsored an evaluation of forensic speaker recognition. They were able to obtain, for limited evaluation use, appropriate audio data from actual police investigations. There were a variety of test conditions, involving different durations and types of data, and participant were asked for decisions in a sequence of trials using a format based on that used in the NIST evaluations. Some of its results are described in [10].

Another, if somewhat less successful evaluation, was held in conjunction with the Odyssey 2001 workshop in Crete. A couple of evaluation tracks were offered to participants in connection with the workshop. One involved a subset of the previous year's NIST evaluation. NIST analyzed submitted results much as in its regular evaluations. See [11,12]. The other track involved text-dependent speaker verification, where the enrollment and verification data consisted of speakers saying one of 17 specified passwords. This track is discussed in [13,14]. Participation was limited and, with respect to the second track involving spoken passwords, this perhaps may show the difficulty of creating text-dependent evaluations of general interest that can attract participants from commercial companies.

The use of speaker recognition as a biometric that may be used for secure verification of people's identities in light of recent word events is attracting increasing interest on both sides of the Atlantic. In Europe, however, there has been greater interest in using multiple biometrics, including speech, in combination to achieve increased performance. A major project denoted BioSecure, a part of the 6th Framework Programme of the European Community, is coordinating a multi-year interdisciplinary research program in support of this. It includes a "2007 BioSecure Evaluation Campaign" involving the use of voice, face, signature, fingerprint, hand, and iris data in a multi-faceted effort that is to launch in March, 2007 [15].

9 Future of Speaker Evaluation

After annual NIST evaluations from 1996 to 2006 it was decided, for a variety of reasons not to hold an evaluation in 2007. The evaluations have become larger over the years, both in test set size and number of participants, and more complicated in terms of the variety of tests included. The hiatus will provide additional time for data collection, always the key limiting factor in evaluation planning. This will allow the next evaluation to include considerably more data corresponding to cross-channel evaluation conditions. The hiatus is also intended to allow time to recruit an additional person to support the evaluation, but it remains to be seen whether continuing annual evaluations will be seen as feasible.

But speaker detection is an area of growing interest, and future evaluations, coordinated by NIST and perhaps other organizations appears quite certain.

There is a likelihood of growing government funding to support research in the area both in the United States and the European Union. This is expected to result in expanded evaluations in the United States while, as noted previously, there are plans in Europe for expanded evaluation of the fusion of biometric technologies including speaker.

The development of the technology may also produce increased demand for more product oriented evaluation. Very high performance, as noted, can be achieved for somewhat limited conditions, and systems to support these will become more visible in the commercial marketplace. But for the more challenging aspects of the task, with full text-independence and the use of the public telephone network or across multiple channels, there remain considerable performance limitations and a continuing need for ongoing evaluation of research systems.

References

1. Doddington, G.: Speech Recognition: Â turning theory to practice. IEEE Spectrum 18(9), 26–32 (1981)
2. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallet, D.S., Dahlgren, N.L.: The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. Technical report, National Institute of Standards and Technology, Gaithersburg (1993)
3. Brümmer, N., Du Preez, J.: Application-independent evaluation of speaker detection. Computer Speech & Language 20(2-3), 230–275 (2006)
4. Doddington, G.: Speaker recognition based on idiolectal differences between speakers. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001). Aalborg, Denmark, Vol. 4, pp. 2521–2524 (2001)
5. Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J.: Phonetic, idiolectal, and acoustic speaker recognition. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 55–63 (2001)
6. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Woves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98) (1998)
7. Fiscus, J.: The NIST Rich Transcription Evaluation Series, NIST web-site (2007), `http://nist.gov/speech/tests/rt/`
8. Fiscus, J., Radde, N., Garofolo, J.S., Le, A., Ajot, J., Laprun, C.: The Rich Transcription 2005 Spring Meeting Recognition Evaluation. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, Springer, Heidelberg (2006)
9. CLEAR2007: Classification of Events, Activities and Relationships, Evaluation and Workshop (2007), `http://www.clear-evaluation.org/`
10. Van Leeuwen, D.A., Martin, A.F., Przybocki, M.A., Boutenc, J.S.: NIST and NFI-TNO evaluations of automatic speaker recognition. Computer Speech & Language 20(2–3), 128–158 (2006)
11. Hansen, E.G., Slyh, R.E., Anderson, T.R.: Formant and F0 Features for Speaker Verification. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 25–29 (2001)

12. Przybocki, M.A., Martin, A.F.: Odyssey Text Independent Evaluation Data. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 21–23 (2001)
13. Higgins, A.L., Bahler, L.G.: ITT SpeakerKey Evaluation. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 31–32 (2001)
14. Toledo-Ronen, O.: Speech Detection for Text-Dependent Speaker Verification. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 33–36 (2001)
15. BioSecure: BioSecure Evaluation Campaign (2007), http://www.biosecure.info/eval/
16. Campbell, J.P., Nakasone, H., Cieri, C., Miller, D., Walker, K., Martin, A.F., Przybocki, M.A.: The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04), Lisbon, Portugal, [Alvin: wasn't this one published at the Odyssey 04 workshop raher than LREC?] (2004)
17. Cieri, C., Andrew, W., Campbell, J.P., Doddington, G., Godfrey, J., Huang, S., Libermann, M., Martin, A., Nakasone, H., Przybocki, M., Walter, K.: The Mixer and Transcript Reading Corpora: Resources for Multilingual Crosschannel Speaker Recognition Research. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06), Genoa, Italy (2006)
18. Reynolds, D.A., Doddington, G., Przybocki, M., Marin, A.: The NIST speaker recognition evaluation - overview, methodology, systems, results, perspectives. Speech Communication 31(2-3), 225–254 (2000)
19. Fiscus, J., Ajot, J., Michel, M., Garofolo, J.S.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In: Renals, S., Bengio, S., Fiscus, J.G. (eds.) MLMI 2006. LNCS, vol. 4299, Springer, Heidelberg (2006)
20. Linguistic Data Consortium: Catalog of Speaker Recognition Corpora (2007), http://www.ldc.upenn.edu/Catalog/
21. Martin, A.F., Przybocki, M.A.: The NIST 1999 Speaker Recognition Evaluation - An Overview. Digital Signal Processing 10, 1–18 (2000)
22. Martin, A.F., Przybocki, M.A.: The NIST Speaker Recognition Evaluations: 1996-2001. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 39–43 (2001)
23. Martin, A.F., Przybocki, M.A., Doddington, G.: Speaker Recognition in a Multi-Speaker Environment. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, vol. 2, pp. 787–790 (2001)
24. Martin, A., Miller, D., Przybocki, M., Campbell, J.: Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04), Lisbon, Portugal (2004)
25. Martin, A.F., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET Curve in Assessment of Detection Task Performance. In: Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997). Rhodes, Greece, vol. 4, pp. 1985–1988 (1997)
26. Martin, A.F., Przybocki, M.A., Campbell, J.P.: The NIST speaker recognition evaluation program. In: Wayman, J., Jain, A.K., Wayman, D.M. (eds.) Biometric Systems: Technology, Design and Performance Evaluation, pp. 241–262. Springer, Heidelberg (2005)

27. Martin, A.F., Przybocki, M.A., Le, A.N.: The NIST Speaker Recognition Evaluation Series, NIST web-site (2007), http://www.nist.gov/speech/tests/spk/
28. Philipps, P.J., Martin, A., Wilson, C., Przybocki, M.: An introduction to evaluating biometric systems. IEEE Computer 33(2), 56–63 (2000)
29. Przybocki, M.A., Martin, A.F.: NIST speaker recognition evaluation. In: Proceedings of the Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C '98), Avignon, pp. 120–123 (1998)
30. Przybocki, M.A., Martin, A.F.: NIST Speaker Recognition Evaluation Chronicles. In: Proceedings of the ODYSSEY Speaker and Language Recognition Workshop (Odyssey 04), Toledo, Spain (2004)
31. Przybocki, M.A., Martin, A.F.: NIST's Assessment of Text Independent Speaker Recognition Performance. In: The Advent of Biometrics on the Internet: Proceedings of the COST 275 Workshop, Rome, Italy, pp. 25–32 (2000)
32. Przybocki, M.A., Martin, A.F., Le, A.N.: NIST Speaker Recognition Evaluation Chronicles Part 2. In: Proceedings of the ODYSSEY Speaker and Language Recognition Workshop (Odyssey '06), San Juan, Puerto Rico (2006)

An Introduction to Application-Independent Evaluation of Speaker Recognition Systems

David A. van Leeuwen[1] and Niko Brümmer[2]

[1] TNO Human Factors,
Soesterberg, The Netherlands
david.vanleeuwen@tno.nl
[2] Spescom DataVoice,
Stellenbosch, South Africa
nbrummer@za.spescom.com

Abstract. In the evaluation of speaker recognition systems—an important part of speaker classification [1], the trade-off between missed speakers and false alarms has always been an important diagnostic tool. NIST has defined the task of *speaker detection* with the associated *Detection Cost Function* (DCF) to evaluate performance, and introduced the DET-plot [2] as a diagnostic tool. Since the first evaluation in 1996, these evaluation tools have been embraced by the research community. Although it is an excellent measure, the DCF has the limitation that it has parameters that imply a particular *application* of the speaker detection technology.

In this chapter we introduce an evaluation measure that instead *averages* detection performance over application types. This metric, $C_{\rm llr}$, was first introduced in 2004 by one of the authors [3]. Here we introduce the subject with a minimum of mathematical detail, concentrating on the various interpretations of $C_{\rm llr}$ and its practical application.

We will emphasize the difference between *discrimination* abilities of a speaker detector ('the position/shape of the DET-curve'), and the *calibration* of the detector ('how well was the threshold set'). If speaker detectors can be built to output well-calibrated log-likelihood-ratio scores, such detectors can be said to have an *application-independent* calibration. The proposed metric $C_{\rm llr}$ can properly evaluate the discrimination abilities of the log-likelihood-ratio scores, as well as the quality of the calibration.

Keywords: speaker recognition, speaker detection, speaker evaluation, speaker calibration, log-likelihood-ratio, $C_{\rm llr}$, DET-curve, APE-curve.

1 Introduction

Formal evaluations have played a major role in the development of speech technology in the past decades. The paradigm of formal evaluation was established in speech technology by the National Institute of Standards and Technology (NIST) in the USA. By providing the research community with a number of essential ingredients, such as new speech data, tasks and rules, and a concluding workshop,

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 330–353, 2007.

these regular evaluations have led to significant improvements in all these evaluated technologies. It is therefore not strange that the evaluation paradigm has been adopted by other research and standards organizations around the world in various technology areas.

One of the most regularly held evaluations in the area of speech research is that of *text-independent speaker recognition* [1]. This Speaker Recognition Evaluation (SRE) series has been organized yearly since 1996 by NIST [1], and has had its 11th edition in the first quarter of 2006. Despite the many factors that have varied along the various editions, a few key aspects have remained essentially constant. One of these is the primary evaluation measure, namely the *detection cost function* (DCF). It is specified in terms of the cost of misses and the cost of false alarms, as well as the prior probability for the target speaker hypothesis. In addition to the DCF, NIST compares the discrimination abilities of systems in *Detection Error Trade-off*[1] (DET)-curves [2], which researchers have embraced almost emotionally. In retrospect it can be concluded that it was quite an important insight of NIST to define DCF and the presentation of the error trade-off curves as they did, for it has become the standard in speaker recognition and is also gradually finding its way into other areas of research.

In the workshop concluding the most recent (2006) NIST SRE, an exciting new development became apparent. It was announced that NIST would in future employ a new primary evaluation measure. This measure, which we call C_{llr}, is the subject of this chapter. It was proposed in a conference paper in 2004 [3] and followed in 2006 by an extended journal paper [4]. The purpose of this chapter is to be a more accessible tutorial introduction to the topic. (Apart from the two above references, interested readers may want to see various other papers which have since appeared on the same or closely related topics [5,6,7,8,9])

In the following, we will first review the problem of speaker detection and the traditional evaluation techniques. This will be followed by motivation for and introduction to some aspects of the new C_{llr} evaluation methodology and the analysis thereof.

1.1 Recognition, Verification, Detection, Identification

In the past, researchers have studied various forms of speaker recognition problems. Most notably, the problem of *speaker identification* has been studied extensively. It seems quite intuitive to see speaker recognition as an identification task, because that appears the way humans perceive the problem. When you hear the voice of somebody familiar, you might immediately recognize the identity of the speaker. However, if we try to measure the performance of an automatic speaker identification system, we find a number of questions hard to answer. How many speakers should we consider in my evaluation? What is the distribution of speakers in the test? If we think about it deeper, we can see that performance measures such as identification accuracy will depend on the choice of these numbers in the evaluation. What if a speaker identification system is

[1] Originally termed PROC in the 1996 evaluation plan.

exposed to an ‘unknown’ speaker in the test? People have introduced ‘open set identification’ as alternative to ‘closed set identification,’ but really the latter is quite an unrealistic situation.

The solution to these undesirable questions lies in the proper statement of the speaker recognition task: in terms of *speaker detection*. Formally, the question is: *Given two recordings of speech, each uttered by a single speaker, do both speech excerpts originate from the same speaker or not?*[2] By developing technology that can answer this question for a broad range of speakers, many different applications are possible. Speaker verification is a direct implementation of the detection task, while open or closed set identification problems can be formulated as repeated application of the detection task.

The succinct statement of the speaker recognition problem in terms of *detection* has several advantages. The analysis of the evaluation can be performed in a standard way, which is the subject of Sect. 2. The evaluation measures do not intrinsically depend on the number of speakers or the distribution of so-called target and non-target trials. The true answer of the detection task can, if the evaluation data collection is carefully supervised, be known by the evaluator with very high confidence. Patrick Kenny summarized these positive aspects of the detection approach by saying: “I’ve never come across a cleaner problem [in speech research]”.[3]

2 The Traditional Approach of the Evaluation of Speaker Recognition Systems

2.1 The Errors in Detection

In order to evaluate a speaker detection system, we can subject the system to two different kinds of *trial*. In each trial, the system is given two recordings of speech, originating either from the *same* speaker or from two *different* speakers. The former situation is called a *target trial* and the latter a *non-target trial*. The evaluator has a truth reference to tell the two types of trial apart, but the system under evaluation has only the speech recordings as input. It is therefore the purpose of the speaker detector to distinguish target trials from non-target trials. In classifying the trials, there are two possible errors a system can make, namely

- false positives, or *false alarms*, classifying a non-target trial as a target trial, and
- false negatives, or *misses*, classifying a target trial as a non-target trial.

[2] One might call this a one-speaker open set identification task.

[3] This is how the statement is recalled as perceived by the authors in a salsa-bar during the week of the 2006 Speaker Odyssey Workshop. However, the extremely high noise levels made proper human perception very hard, which is indicative of the fact that Automatic *Speech* Recognition cannot be stated as such a clean problem.

We observe that the speaker detection problem gives rise to *two* types of error, the rates of occurrence of which are to be measured in an evaluation. Having two different error-rates complicates things because it makes it hard to compare the performance of one system with another, or to observe an improvement in one system when it is adjusted. Since comparison is the essential goal of evaluation, it is important to find a way to do this. It is therefore the purpose of this chapter to examine the question: how do we combine these two error-rates into a single performance measure that is representative of a wide range of applications?

2.2 The DET-Plot: A Measure of Discrimination

In order to continue, we need to introduce some of the basic concepts of how speaker detectors work. There are many sources of variability in speech signals and therefore a speaker detection system cannot be based on exact matching of two patterns. Instead, it works with (statistical) models, and it calculates some form of *score*[4] which represents the degree of support for the target speaker hypothesis rather than the non-target hypothesis. The higher (more positive) the score, the more the target hypothesis is supported and the lower (more negative) the score, the more the non-target hypothesis is supported. It can be shown that all the information which is relevant to making decisions between the two hypotheses and which can be extracted from the two speech inputs of a trial, can be distilled into a single real-valued score. Decisions as to which hypothesis is true can now be based on whether or not the score exceeds a well chosen threshold. Setting this threshold (a process known as *calibration*) is the next challenge.

If we now look at the scores that a speaker detector typically yields for the two types of trials, target and non-target trials, we may plot score distributions as in Fig. 1. These score distributions, obtained from a real speaker detector evaluated on NIST SRE 2006 data [1], has typical behaviour: the distributions overlap, the target scores having higher values on average than non-target scores, and the variance of the distributions is different. The threshold-based decision leads to the error-rates P_{FA} and P_{miss}, that can be read from the figure as the proportion of the non-target scores exceeding the threshold and the proportion of target scores below the threshold. From the figure you may also appreciate the fact that if the threshold were chosen differently, the values of P_{FA} and P_{miss} would change. More specifically, they would change in opposite directions. Thus, there is an inherent trade-off between lowering P_{FA} against lowering P_{miss}.

This trade-off is most spectacularly shown in a graph that is known as the Detection Error Trade-off or *DET-plot* [2], where a parametric plot of P_{miss} versus P_{FA} is made, an example is shown in Fig. 2. The axes of a DET-plot are warped according to the *quantile function of the normal distribution*, or using another name, the probit function,

$$Q(p) = \text{probit}(p) = \sqrt{2}\,\text{erf}^{-1}(2p-1). \tag{1}$$

[4] Often called a *likelihood ratio*, but we will not use this term for reasons that will become clear later.

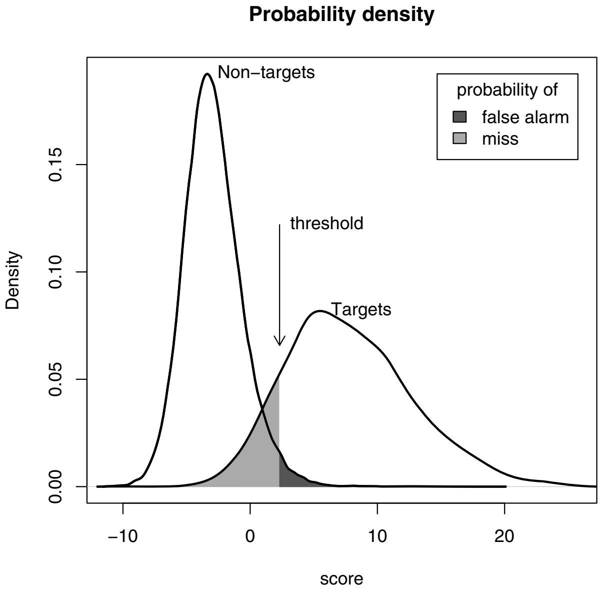


Fig. 1. The score distributions for non-target (left) and target (right) trials. The grey areas left and right of the threshold represent P_{miss} and P_{FA}, respectively.

where p is P_{FA} or P_{miss}, and ‘erf^{-1}’ is the inverse of the *error function*. There are several effects of the warping of axes. Firstly, if the target and non-target score are distributed normally, the detection error trade-off will be a straight line,[5] with a slope $-\sigma_{\text{non}}/\sigma_{\text{tar}}$, where $\sigma_{\text{tar,non}}$ are the standard deviations of the target and non-target distributions, respectively [10,11]. Secondly, the warping has the advantage that several curves plotted in the same graph gives rise to less clutter than if the probability axes were linear, as in ROC-curves (Receiver Operating Characteristic, which is the traditional way of plotting false alarms versus misses, or hits).

The DET-plot shows what happens as the decision threshold is swept across its whole range, but on the curve one can also indicate a fixed *operating point* as obtained when making decisions at a fixed threshold. It has been customary in NIST evaluations to require not only scores, but also hard decisions. The P_{miss} and P_{FA} measured for these hard decisions correspond to such an operating point on the curve.[6] It is good practice to draw a box around this point, indicating the 95 % confidence intervals of P_{FA} and P_{miss}, assuming trial independence and binomial statistics [12].

[5] The reverse is not true, however. Note, that even though the underlying distributions deviate noticeably from normal distributions (see Fig. 1), the DET-curve is straight over a reasonably large range of probabilities.

[6] Provided these hard decisions were indeed made by thresholding the same score that was used to generate the DET-plot.

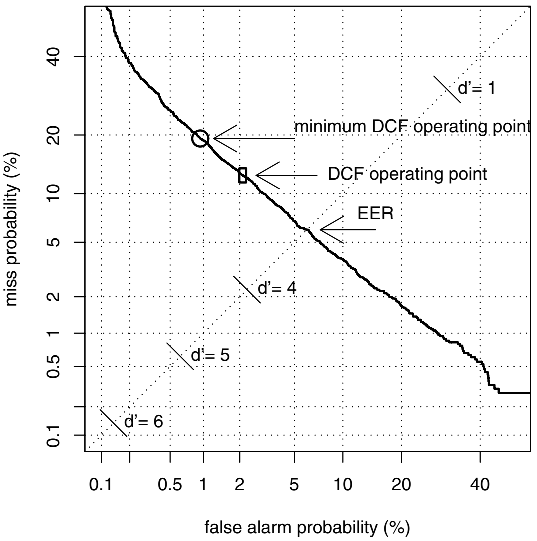


Fig. 2. A DET-plot, obtained from Fig. 1. The line shows the trade-off of false alarm against miss probability as the threshold increases from the lower-right to upper-left corner. The rectangle indicates the operating point of the decisions made, corresponding to the surface of the grey areas in Fig. 1. Further, the Equal Error Rate (EER) and the *operating point* the 'minimum DCF' (see Sect. 2.3) are indicated. For d', see the text.

The DET-plot very clearly shows how the two error types can be traded off against each other. For a given DET-performance the false alarm rate can be reduced to an almost arbitrary low level by setting the detection threshold high enough, if one is prepared to accept a high miss rate. And vice versa; it all depends on the application of the system: if the costs of a false alarm are very high, or the prior probability of a target event is very low, we set the threshold high and we 'operate' in the upper-left corner of the plot. If the application sets different demands, we can operate at the opposite end. This trade-off is not new, a theory of signal detection was developed for radar signals midway the 20th century, and later used by psychophysicists to model human perception of stimuli in the sixties [13,14]. We experience the same trade-off in everyday life, such as in trying to separate spam e-mails from serious messages, and in trying to create laws in society that can convict criminals while guaranteeing freedom for citizens. In fact, in understanding the DET or ROC curves it becomes apparent that striving for 'zero tolerance' or any other form of perfect filtering will backfire immediately by resulting in unreasonable high costs at the flip side of the coin.

Returning now to speaker recognition, researchers have grown very fond of DET-curves because they indicate the discrimination potential of their system at a glance. DET-curves more towards the lower-left indicate better discrimination ability between the target and non-target trials, and hence better algorithms. Tiny improvements in the detector will show noticeable displacement in the

DET-curve, which stimulates the researcher to think of even more clever things. A DET-plot is a great diagnostic tool: if the curve deviates far from a straight line, or shows unexpected cusps or bends, this is usually an indication that there is something wrong in the detector or in the evaluation data or its truth reference. As a final goody, plotting a DET-curve does not require setting a threshold.

The Equal Error Rate. We went from decisions and P_{FA} and P_{miss} to no decisions and a whole *curve* that characterizes our detector. Can we somehow summarize the DET-curve as a single value? Yes, we can, in several ways.

Firstly, noticing P_{FA} and P_{miss} move in opposite directions if the threshold is changed, there always is a point where $P_{\text{FA}} = P_{\text{miss}}$. This joint value of the error rates is called the *Equal Error Rate* or EER. In the DET-plot it can be found as the intersection of the DET-curve and the diagonal. The EER is a concise summary of the discrimination capability of the detector.[7] As such it is a very powerful indicator of the *discrimination* ability of the detector, across a wide range of applications. However, it does *not* measure calibration, the ability to set good decision thresholds.

It may be interesting to compare the EER to a related measure from signal detection theory. Here the task is to detect a signal in Gaussian noise, and hence the two distributions to be separated are normal and have equal variance. In this case, the DET-curve is completely characterized by the single parameter 'd-prime,' the distance between the means of the distributions measured in units of the standard deviation: $d' = (\mu_{\text{tar}} - \mu_{\text{non}})/\sigma$. In Table 1 the relation between d' and the EER is shown, in order to give an idea what the separation of the target and non-target distributions means in terms of EER. Another way of seeing d' is in the DET-plot (see Fig. 2), where it represents straight lines of slope -1. The value of d' determines where the diagonal is crossed, starting at the upper-right corner for $d' = 0$ moving down linearly to the lower-left corner where $d' \approx 6$.

Table 1. Relation between d', the separation of distribution in terms of standard deviations, and the EER

d'	0	1	2	3	4	5
EER (%)	50.0	30.9	15.8	6.7	2.27	0.62

2.3 The Detection Cost Function: Simultaneous Measure of Discrimination and Calibration

In calculating the DET-plot and EER, the evaluator effectively chooses optimal decision thresholds, with reference to the truth. These evaluation procedures therefore do not measure the actual decision-making ability of the detector on

[7] It can be shown [15,4] that if decision thresholds are always set optimally, then the EER is the *upper bound* of the average error-rate of the detector as P_{tar} is varied. By average error-rate, we mean $P_{\text{tar}}P_{\text{miss}} + (1 - P_{\text{tar}})P_{\text{FA}}$, where P_{tar} is the prior probability of a target event.

unseen data. The canonical solution is a direct one—simply require the detector to make decisions and then count the errors. Now how do we now combine these error counts (of two types of error) into a scalar measure of goodness of decision-making ability?

At a first glance, one could simply use the total number of errors as a performance measure. Indeed, this solution is routinely practised by the machine learning research community. However, reflecting on real applications there are at least two important complications:

- The proportion of targets and non-targets may be different from the proportions in the evaluation database.
- The two types of errors may not have equally grave consequences. For example, for a fraud detection application the costs of a missed target (cross customers) can be higher than the cost of a false alarm (a fraudulent action not observed), while for access control the cost of a false alarm (security breach) may outweigh the cost of a miss (annoyed personnel).

It therefore makes sense to weight the two normalized error-rates with (i) the prior probability of targets in the envisaged application and (ii) the estimated *costs* of the two error types. Applying these weightings, one then arrives at a scalar performance measure, namely the *expected cost of detection errors*,

$$C_{\mathrm{det}}(P_{\mathrm{miss}}, P_{\mathrm{FA}}) = C_{\mathrm{miss}} P_{\mathrm{miss}} P_{\mathrm{tar}} + C_{\mathrm{FA}} P_{\mathrm{FA}} (1 - P_{\mathrm{tar}}). \qquad (2)$$

This function has become known as the *detection cost function.* Here the normalized error-rates P_{miss} and P_{FA} are determined by the evaluator by counting errors. The application dependent cost parameters C_{miss} and C_{FA} are discussed above, and the parameter P_{tar} is the prior probability that a target speaker event occurs in the application. This prior must be assigned to correspond to some envisaged application of the speaker detector.

Given prescribed values for the parameters of C_{det}, the onus now rests on the designer of a speaker recognition system under evaluation, to choose a score decision threshold that minimizes C_{det}. For this purpose the evaluee may use a quantity of development data with a known truth reference. Minimizing C_{det} on the development data may or may not give a C_{det} that is close to optimal on new unseen evaluation data. This is an important part of the art of designing a speaker detector: to calculate scores that are well-normalized so that thresholds set on development data still work well on unseen data.

In summary, the three application-dependent parameters C_{miss}, C_{FA} and P_{tar}, form the detection cost function $C_{\mathrm{det}}(P_{\mathrm{miss}}, P_{\mathrm{FA}})$, which gives a single scalar performance measure of a speaker detection system.

The detection cost function is a simultaneous measure of *discrimination* and *calibration.* This error measure of a detector will have a low value provided that both (i) EER is low *and* (ii) the threshold has been set well.

C_{det} has been used since the first NIST speaker recognition evaluation in 1996 as the primary evaluation measure, and with it, the three application-dependent cost parameters have been assigned values $C_{\mathrm{miss}} = 10$, $C_{\mathrm{FA}} = 1$ and $P_{\mathrm{tar}} = 1\,\%$.

These values have never changed in the evaluations, and occasionally a researcher wonders how these values were chosen. The long tradition and fixed research goals have caused these choices to fade from our collective memory, but in a recent publication [12] an example of an application with these cost parameters is given.

'Minimum Detection Cost.' Minimum C_{det} is similar, but not identical to EER. It is a measure of discrimination, but not of calibration. It is defined as the optimal value of C_{det} obtained by adjustment of the detection threshold, given access to the truth reference. Unlike EER it is dependent on the particular application-dependent parameters of C_{det}.

In the context of the NIST SRE, it is customary to indicate $C_{\text{det}}^{\text{min}}$ on DET-curves, as is shown by the circle in Fig. 2. Note that this circle does not show the numerical value of $C_{\text{det}}^{\text{min}}$, rather it shows the values of P_{miss} and P_{FA} at which C_{det} is minimized. This is in contrast to the APE-curve, which we introduce below, which does directly show the numerical value of $C_{\text{det}}^{\text{min}}$.

Discussion. So we've found two more performance metrics, EER and $C_{\text{det}}^{\text{min}}$, that each summarize the DET-plot in their own way. Both are used extensively in literature, the former in a 'general application' context and the latter in a 'NIST evaluation' context. They are very important performance metrics, but they circumvent one major issue: setting the threshold. In fact, EER and $C_{\text{det}}^{\text{min}}$ are *after the fact* error measures. They imply that the threshold can not be set until all trials have been processed and, moreover, the truth about the trials is known. Summarizing, EER and $C_{\text{det}}^{\text{min}}$ are great for indicating the discrimination potential, but they do not fully measure the capability of making hard decisions.

Is this really a problem? For many researchers it is not. Setting the threshold, as is necessary for submitting results to a NIST evaluation, is simply based on last year's evaluation data, for which the truth reference has been released.[8] This usually results in a C_{det} that is not too much above $C_{\text{det}}^{\text{min}}$, and everything is fine. Sometimes, the evaluation data collection paradigm has changed or the recruitment of new speakers has been carried out in a different way, and the calibration turns out wrong. A real shame, but usually most participating systems 'get hurt' in the same way, and there is always a next year to do better.

So let us recapitulate our quest for a single, application independent performance measure for speaker recognition systems. We started with a clear and unambiguous statement of the task of a speaker recognition system. This lead to two types of error which are interrelated by means of a trade-off. By using a cost function C_{det}, we could reduce the two error measures to a single metric, at the cost of having to define application-dependent parameters. Postponing the setting of a threshold gave us a beautiful DET-plot and a powerful EER summary, at the cost of not measuring calibration.

In the previous section we have introduced several measures characterizing the performance of a speaker recognition system. Although they each have their

[8] Often, the calibration happens just before the results are due. The present authors are in this respect not different from other researchers.

merits and their use is quite widespread, we will show in this section that we can demand more information from a speaker detector than just a score and a decision, and that there exists a metric that says how good this information is. It combines the concept of expected costs, like C_{det} does, with soft decisions and application-independence, like the DET-curve suggests. Before we introduce it, we are going to have a closer look at the interpretation of scores.

2.4 The Log-Likelihood-Ratio

So far, we have learnt that a speaker detection system produces a score for every trial. The only thing we have required of the score is that a higher score means that the speech segments are more alike. A set of scores is sufficient to produce a DET-curve, and with an additional threshold we can also calculate C_{det}. But there is a *lot* of freedom in the values of the scores. First, there is an arbitrary offset that can be added to all scores (and the threshold) and nothing in the evaluation will change. Or the score can be scaled; in fact, the whole score-axis can be warped by any monotonic rising function, and everything in the DET-plot will stay exactly the same. There is no meaning in the scores, other than an ordering.

We can use this freedom in score values to fix the problem of application dependence. To see how this works, we examine how a score s for a given trial can be used to make an optimal decision for that trial. The expected cost of making an *accept* decision is $(1 - P(\text{target trial}|s))C_{\text{FA}}$, while the expected cost of making a *reject* decision is $P(\text{target trial}|s)C_{\text{miss}}$. Here $P(\text{target trial}|s)$ is the *posterior* probability for a target trial, given the score s. The minimum-expected-cost decision is known as a Bayes decision.[9] To make a Bayes decision, we need the posterior, which may be expressed, via Bayes' rule, as

$$\operatorname{logit} P(\text{target trial}|s) = \mathcal{L}(s) + \operatorname{logit}(P_{\text{tar}}) \tag{3}$$

where[10]

$$\mathcal{L}(s) = \log \frac{P(s|\text{target trial})}{P(s|\text{non-target trial})} \tag{4}$$

is known as the *log-likelihood-ratio* of the score. Putting this all together, we get a concise decision rule:

$$\text{decision}(s, \theta) = \begin{cases} \text{accept} & \text{if } \mathcal{L}(s) \geq -\theta, \\ \text{reject} & \text{if } \mathcal{L}(s) < -\theta, \end{cases} \tag{5}$$

where the decision threshold θ is a function of the application-dependent cost and prior parameters,

$$\theta = \log \left(\frac{P_{\text{tar}}}{1 - P_{\text{tar}}} \frac{C_{\text{miss}}}{C_{\text{FA}}} \right) \tag{6}$$

[9] It is easily shown that if one makes a Bayes decision for every trial, this will also optimize the expected error-rate over all the trials, which is just our evaluation objective C_{det}.

[10] We use the function: $\operatorname{logit} p = \log \frac{p}{1-p}$, which re-parametrizes probabilities as *log odds*, because for binary hypotheses, it transforms Bayes' rule to the elegant additive form of (3).

Equation (5) forms a neat separation between $\mathcal{L}(s)$ and θ. The purpose of the score, s, is to extract relevant information from the given speech data of the trial. The purpose of $\mathcal{L}(s)$ is to shape, or *calibrate*, this information into a form that can be used in a standard way to make good decisions. The information, $\mathcal{L}(s)$, extracted from the speech data is application-independent, because all the application-dependent parameters have been separated and encapsulated into the single *application parameter* θ.

Notice that $\mathcal{L}(s)$ may also be called a *score*. It has the same look and feel[11] as s, where more negative scores favour the non-target hypothesis and more positive scores favour the target hypothesis. The difference is that $\mathcal{L}(s)$ is calibrated so that minimum-expected-cost decisions may be made with the standard threshold θ.

In fact $\mathcal{L}(s)$ may be interpreted as expressing the degree of support that the raw score s gives to one or the other hypothesis. When $\mathcal{L}(s)$ is close to zero, the score does not strongly support either hypothesis, but as the absolute value of $\mathcal{L}(s)$ grows there is more support for one or the other hypothesis. The hypothesis that is favoured is indicated by the sign of $\mathcal{L}(s)$.

If a speaker detector can produce $\mathcal{L}(s)$ instead of the raw s, this has obvious advantages for users. The *same* system can now be used by different users having *different* applications (i.e., different θ), and still the calibration is right. The user does not have to ask the system developer: "My application parameters have changed. Could you please re-calibrate your detector?" Now the user can easily calculate the threshold θ and indeed change it at will as circumstances dictate.

So what is new here? Nothing in fact. The theory of making Bayes decisions has been known for a long time. The catch is that even if your DET-curve is good it may also be difficult to calculate well-calibrated soft decisions in log-likelihood-ratio form, just like it used to be difficult to set good hard decision thresholds for C_{det}. The key to this problem is that until quite recently it has not been known in the speaker recognition community how to *evaluate the quality* of detection log-likelihood-ratios. The purpose of this chapter is therefore to introduce the reader to how this may be done. Once we know how to measure, half the battle towards improving performance has been won.

2.5 Log-Likelihood-Ratio Cost Function

At a first glance, evaluation of log-likelihood-ratio scores may be accomplished by a small adjustment of the NIST SRE protocol:

> Instead of having evaluees submit hard decisions for evaluation via C_{det}, they are now required to submit soft decisions in log-likelihood-ratio form. Then instead, the *evaluator* makes the decisions by setting the threshold at $-\theta$. These decisions may then be plugged into C_{det} as before, to get a final evaluation result.

[11] This is why we prefer to work with a log-likelihood-ratio, rather than a likelihood-ratio. The (non-negative) likelihood-ratio has the uncomfortable asymmetry where smaller scores are compressed against 0.

In principle this is a very good plan, but it has the flaw of not really changing anything. If the value of θ is known to participants, then they may calibrate their scores to work well only at the specific point on the log-likelihood-ratio axis that is 'sampled' by evaluation at θ. Intuitively, sampling the log-likelihood-ratio at a single point can show that scores have been shifted to have log-likelihood-ratio interpretation, but it still leaves the scale of the evaluated scores completely arbitrary.

Once we have realized that a single sampling point is the problem, it is conceptually easy to fix: just sample the decision-making ability of the log-likelihood-ratio scores under evaluation at more than one value of θ. The evaluator may now calculate a C_{det} at each of these operating points. This leaves the questions of (i) how many points do we need to sample, (ii) which points do we choose and (iii) how do we combine the different C_{det} results over these points in order to get a single metric?

Of course there are many good answers to these questions. Here we discuss the particular solution which has been motivated in detail in [4]. This solution proposes to sample C_{det} over an infinite 'spectrum' of operating points and to then simply integrate over them, thus:

$$C_{\text{llr}} = C_0 \int_{-\infty}^{\infty} C_{\text{det}}\big(P_{\text{miss}}(\theta), P_{\text{FA}}(\theta), \theta\big)\ d\theta \tag{7}$$

where C_{llr} is the new metric, which we call the *log-likelihood-ratio cost function* and where $C_0 > 0$ is a normalization constant. Some notes are in order:

- The error-rates P_{miss} and P_{FA} are now functions of θ, because $-\theta$ is just the decision threshold. By sweeping the decision threshold, the evaluator is effectively sweeping the whole DET-curve of the system under evaluation. This effectively turns C_{llr} into a summary of *discrimination* ability over the whole DET-curve, somewhat similar to EER.
- Equally important is the fact we have now also made C_{det} dependent on θ. Since C_{det} implies making actual decisions, we are also incorporating the evaluation of calibration into our metric. Moreover, since C_{det} varies with θ, we are also measuring calibration over the whole θ-spectrum. Recall from (2) that C_{det} is parameterized by the triplet $(P_{\text{tar}}, C_{\text{miss}}, C_{\text{FA}})$. We may parametrize C_{det} equivalently[12] by $(\tilde{P}_{\text{tar}}, \tilde{C}_{\text{miss}} = 1, \tilde{C}_{\text{FA}} = 1)$, where $\tilde{P}_{\text{tar}}$ 'incorporates' the cost parameters. This single parameter $\tilde{P}_{\text{tar}}$ can be expressed in terms of θ,

$$\begin{aligned}\tilde{P}_{\text{tar}} &= \frac{P_{\text{tar}} C_{\text{miss}}}{P_{\text{tar}} C_{\text{miss}} + (1 - P_{\text{tar}}) C_{\text{FA}}} \\ &= \frac{1}{1 + e^{-\theta}} = \operatorname{logit}^{-1} \theta \end{aligned} \tag{8}$$

  If we parameterize like this, then $\theta = \operatorname{logit}(\tilde{P}_{\text{tar}})$ has the interpretation of *prior log-odds*. The interested reader may consult [4] for further motivation of

[12] By *equivalent*, we mean that identical decisions, DET-curves and comparisons between systems are made. The DCF itself is scaled down by a factor $P_{\text{tar}} C_{\text{miss}} + (1 - P_{\text{tar}}) C_{\text{FA}}$, which is 1.09 for the NIST parameters.

> this parametrization. In short, although specifying cost and prior are necessary when making decisions in real applications, having both costs and prior as evaluation parameters is redundant. Since the cost and prior multiply to form the parameter θ, we may arbitrarily assign fixed costs and parametrize the entire spectrum of applications by the single parameter $\tilde{P}_{\text{tar}}$, or equivalently by θ. By assigning unity costs we gain the advantage that now C_{llr} may be interpreted as an integral over *error-rates*. Finally, since we are making actual decisions and evaluating them via C_{det}, we are not only measuring discrimination, but we are also at the same time measuring *calibration*.

Realizing that the new measure C_{llr} is a measure of both *discrimination* and *calibration*, we see that C_{llr} for a detector will be good provided that both (i) EER is low *and* (ii) $\mathcal{L}(s)$ is reasonably well calibrated over all operating points of the θ-spectrum.

To recapitulate, C_{det} is a measure of discrimination and calibration suitable for evaluating *hard* (application dependent) detection decisions, while C_{llr} is a measure of discrimination and calibration suitable for evaluating *soft* (application-independent) detection decisions in log-likelihood-ratio form.

Practical Calculation. Equation (7) is a derivation and an interpretation of our new metric C_{llr} but how do we practically calculate this integral? The good news is that it has an analytical closed-form solution:

$$C_{\text{llr}}\big(\{\mathcal{L}'_t\}\big) = \frac{1}{2\log 2}\left(\frac{1}{N_{\text{tar}}}\sum_{t\in\text{tar}}\log(1+e^{-\mathcal{L}'_t}) + \frac{1}{N_{\text{non}}}\sum_{t\in\text{non}}\log(1+e^{\mathcal{L}'_t})\right). \quad (9)$$

where $\mathcal{L}'_t$ is the attempt of the system under evaluation to calculate the log-likelihood-ratio (of (4)) for trial t; and where 'tar' is a set of N_{tar} target trials and 'non' is a set of N_{non} non-target trials. The two normalized summation terms respectively represent expectations of 'log costs' for target trials (left-hand term) and for non-target trials (right-hand term).

Let us look more closely at these log costs. For a target trial the cost is $C_{\text{tar}} = \log(1+e^{-\mathcal{L}'_t})$. If the detector correctly gives a high degree of support for the target hypothesis, $\mathcal{L}'_t \gg 1$, then the cost is low: $C_{\text{tar}} \approx 0$; but if it incorrectly gives a high degree of support for the non-target hypothesis, $\mathcal{L}'_t \ll -1$, then the cost is high[13]: $C_{\text{tar}} \approx |\mathcal{L}'_t|$. Conversely, the cost for non-target trials, $C_{\text{non}} = \log(1+e^{\mathcal{L}'_t})$, behaves the other way round.

[13] When degree of support is expressed as log-likelihood-ratio, then the behaviour of the log-cost is intuitively pleasing: if the detector output has the wrong sign, there is a cost which increases with the magnitude of the error. But if degree of support is instead expressed as a posterior probability, then a posterior of exactly 0 corresponds to $\mathcal{L}'_t = -\infty$ and then $C_{\text{tar}} = \infty$ (likewise, for a non-target trial, a posterior of 1 gives $C_{\text{non}} = \infty$). This is not a flaw of the C_{llr} metric. Rather it shows that a posterior of 0 or 1 is an unreasonable output to give in a pattern recognition problem where there can never be complete certainty about the answer. Working with system outputs (of moderate magnitude) in *log*-likelihood-ratio form, rather than likelihood-ratio form or posterior probability form naturally guards against this problem.

We have seen that extremely strong support for either hypothesis can have high cost, but what is the cost of a neutral log-likelihood-ratio? When $\mathcal{L}'_t = 0$, then $C_{\text{tar}} = C_{\text{non}} = \log 2$. This means that *the reference detector*, which does not process speech and which just outputs $\mathcal{L}'_t = 0$ for every trial, will earn itself a reference value of $C_{\text{llr}} = 1$. This is of course no coincidence, but is a consequence of the normalization factor in (9).

2.6 Discrimination/Calibration Decomposition: The PAV Algorithm

So far we have shown how the new cost measure C_{llr} generalizes C_{det}—but can we also find an analogy for $C_{\text{det}}^{\text{min}}$, the minimum achievable C_{det} if calibration were right? Again, the answer is affirmative. Just like a miscalibrated threshold can be fixed, post hoc, by choosing a different threshold that minimizes C_{det}, it is possible to find a *monotonic rising* warping function w, which, when applied applied to $\mathcal{L}'_t$ for every trial t, will minimize C_{llr} as measured on the warped log-likelihood-ratios $\mathcal{L}''_t = w\big(\mathcal{L}'_t\big)$. As before the minimization is performed given the truth reference for the evaluation, but note that it involves finding the whole warping function w rather than just a single threshold value. The warping function is constrained to be monotonic rising for several reasons:

- It is consistent with applying a single decision threshold to both $\mathcal{L}'_t$ and $\mathcal{L}''_t$.
- A monotonic rising function is invertible and therefore information-preserving. The warping function should correct only the *form* (calibration) of the output, but not the *content* (discriminative ability) of the score.
- The DET-curve (and therefore also the EER) is invariant under monotonic rising warping.
- If there were no constraint, C_{llr} would trivially be optimized to zero, which is a useless result.

How do we find w? Note first that since monotonicity is the only constraint, every value of w can be optimized independently for every trial, in a *non-parametric* way. There is a remarkable algorithm known as the *Pool Adjacent Violators* (PAV) algorithm[14] which can be employed to do this constrained non-parametric optimization. The input is the system-supplied log-likelihood-ratio scores for every trial as well as the truth reference. The output is a set of optimized log-likelihood-ratio values for these trials, where the sorted ordering of input and output scores remains the same, because of the monotonicity. With these optimally calibrated log-likelihood-ratios $w(\mathcal{L}'_t)$ we can apply (9) to find the *minimum* C_{llr}

$$C_{\text{llr}}^{\text{min}} = C_{\text{llr}}\big(\{w(\mathcal{L}'_t)\}\big). \tag{10}$$

It is beyond the scope of this chapter to go into the details of the PAV algorithm (details are available in [4] and references therein), but it may be instructive to see what the warping function $w(\mathcal{L})$ typically looks like. Let us take the system

[14] It is also known as *isotonic regression.*

that produced the score distributions in Fig. 1 and the DET-curve shown in Fig. 2. We plot the warping function $w(\mathcal{L})$ for this system, as found by the PAV algorithm, in Fig. 3. The PAV warping function has a stepped nature, which is a consequence of the 'pooling' of monotonicity violators. This system shows an average slope of 1 over a reasonable range of $\mathcal{L}$, but there is an offset. The log-likelihood-ratios given by this system are too optimistic towards target speakers. One can further observe a non-linear flattening of the curve at the extremes, indicating that the system-supplied log-likelihood-ratio tended to be over-optimistic in those regions.

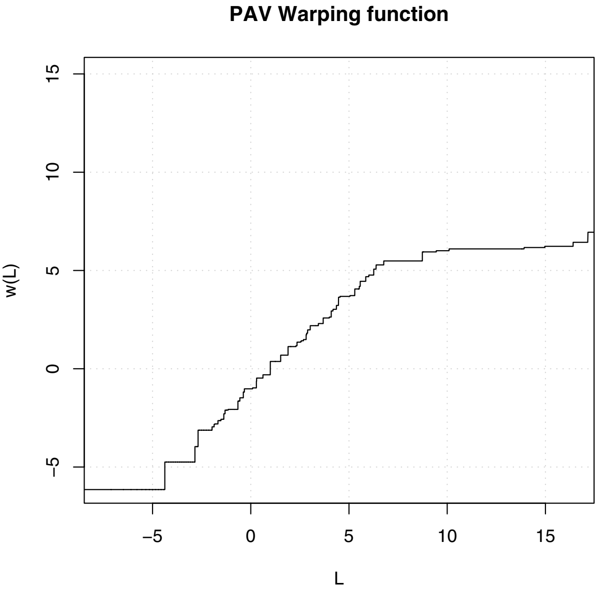


Fig. 3. The result of the PAV algorithm applied to the log-likelihood-ratio scores for which the score distributions were shown in Fig. 1

Note that the PAV algorithm can also be used as the basis for calibration. Just like a detector can be calibrated for a single application-type by choosing a threshold that minimizes C_{det} on some development test data, it is possible to calibrate log-likelihood-ratio scores by applying the PAV algorithm to development test data scores s, to minimize C_{llr} for that data. The warping function $w(s)$ can then be interpreted as a *score to log-likelihood-ratio* function $\mathcal{L}(s)$. Having said this, we leave the subject of calibration methods, since it is not a topic of this chapter. Rather, this is the story how to *measure* calibration.

Recall that C_{llr} is a measure of *both discrimination and calibration.* But since $C_{\mathrm{llr}}^{\mathrm{min}}$ has any calibration mismatch optimized away, it is a now pure measure of *discrimination.* This now allows us to decompose[15] C_{llr} to also obtain a pure measure of calibration. Because of the logarithmic nature of C_{llr}, it turns out that it is appropriate to form an additive decomposition: Our measure of calibration now becomes just $C_{\mathrm{llr}} - C_{\mathrm{llr}}^{\min}$. This difference is non-negative, is close to zero for well-calibrated systems, and grows without bounds as the system under calibration becomes increasingly miscalibrated. In summary, this PAV-based procedure forms the application-independent generalization of the traditional measures $C_{\mathrm{det}}^{\min}$ and $C_{\mathrm{det}} - C_{\mathrm{det}}^{\min}$.

As we shall further demonstrate with APE-curves below, the ability to do this discrimination/calibration decomposition is an important feature of the C_{llr} methodology. The ability to separate these aspects of detector performance empowers the designer of speaker detection systems to follow a divide-and-conquer strategy: First concentrate on building a detector with good discriminative ability, without having to worry about calibration issues. Then when you want to move on to practical applications, concentrate on also getting the calibration sorted out.

2.7 The APE-Curve: Graph of the C_{llr} Integral

The C_{llr}-integral, (7), is the integral of $C_{\mathrm{det}}(\theta)$ over the application parameter θ. We will now show that this integral can be visualized in a powerful graph. The essential part of the integrand of (7) is the error probability

$$P_e(\theta) = \tilde{P}_{\mathrm{tar}}(\theta) P_{\mathrm{miss}}(\theta) + (1 - \tilde{P}_{\mathrm{tar}}(\theta)) P_{\mathrm{FA}}(\theta). \tag{11}$$

Note that all of P_e, $\tilde{P}_{\mathrm{tar}}$, P_{miss} and P_{FA} are functions of θ. The graph of P_e against θ forms the basis of the *Applied Probability of Error* (APE)-plot.

In Fig. 4 we show the APE-plot for our example system. Along the horizontal axis we have θ, which as explained before can be called the 'prior log odds'. Note that the horizontal axis of the APE-plot is the whole real line, but that we plot[16] only the interesting interval close to $\theta = 0$. The vertical axis is the error-rate axis, which takes values between 0 and 1. On these axes, we plot three curves: solid, dashed and dotted, which are respectively error-rates of the actual, PAV-optimized and reference systems. From these plots we can read a wealth of information:

The solid curve. is $P_e(\theta)$ of (11). It shows the error-rate obtained (at each θ) when minimum-expected cost decisions are made with the log-likelihood-ratio scores $\mathcal{L}'_t$ as output by the system under evaluation. Note:

[15] In this chapter, we use the term *discrimination/calibration* decomposition. This is similar in spirit, but not in form, to the *refinement/calibration* decomposition which was introduced by De Groot two decades ago [16] and again recently examined for speaker detection in ref. [7]

[16] Recall that both of the axes in DET-curves are also infinite and that there too, we plot only a selected region.

- The area[17] under the solid curve is proportional to C_{llr}, which can be interpreted as the *total actual error* over the spectrum of applications.
- The vertical dashed line at $\theta = -\log 9.9$ represents the traditional NIST DCF parameters, so that the solid curve at this point gives[18] the traditional *actual* C_{det}.
- The error-rate goes to zero for large $|\theta|$, in such a way that the C_{llr} integral exists (has a finite value).[19]

The dashed curve. shows $P_e(\theta)$, but with scores $\mathcal{L}'_t$ replaced by $w(\mathcal{L}'_t)$ as found by the PAV algorithm.

- The area under the dashed line is proportional to $C_{\mathrm{llr}}^{\mathrm{min}}$, which can be interpreted as the *total discrimination error* over the whole spectrum of applications.
- The area between the solid and dashed curves represents the *total calibration error*.
- At the vertical line representing the NIST DCF parameter settings, $C_{\mathrm{det}}^{\mathrm{min}}$ can be read[20] from the dashed curve.
- The dashed curve has a unique global maximum, which is the equal-error-rate (EER). This maximum is typically located close to $\theta = 0$.

The dotted curve. represents the probability of error for the reference detector, which does not use the speech input, basing its decisions only on the prior $\tilde{P}_{\mathrm{tar}}$. As noted above, the reference detector outputs $\mathcal{L}'_t = 0$ for every trial. The error-rate of the reference detector is $P_e(\theta) = \min(\tilde{P}_{\mathrm{tar}}(\theta), 1 - \tilde{P}_{\mathrm{tar}}(\theta))$. Note here:

- The APE-plot scale does not show the maximum at $P_e = 0.5$.
- The area under the dotted curve is proportional to one (with the same scale factor as the areas under the other curves), and therefore represents the C_{llr}-value of the reference system.
- For $|\theta| \gg 1$, P_e goes to zero rapidly.
- For large negative θ we can observe that our example system performs *worse* than the reference detector!

The APE-curve is complementary to the traditional DET-curve. There is information, like the EER, that is duplicated in both curves, while some information displays better on the DET-curve, and other information better on the APE-curve. As a general rule, the DET-curve is a good tool for examining details of discriminative ability, while the APE-curve a a good tool for examining details of calibration. In addition, both curves have value as educational resources: As we know, the DET-curve demonstrates the error-tradeoff. The APE-curve demonstrates:

[17] The area is the analytically derived definite integral over the whole infinite θ-axis and not just the area under the visible part of the curve.

[18] The value of the solid curve is an *error-rate*, which is a scaled version of the *cost*, C_{det}, where the scaling factor is 1.09, as derived in footnote 12.

[19] This holds, provided that $|\mathcal{L}'_t| < \infty$, for every trial t. If however the system does output even a single log-likelihood-ratio of infinite magnitude having the wrong sign, then the C_{llr} integral will evaluate to infinity.

[20] Again subject to the scaling factor of 1.09.

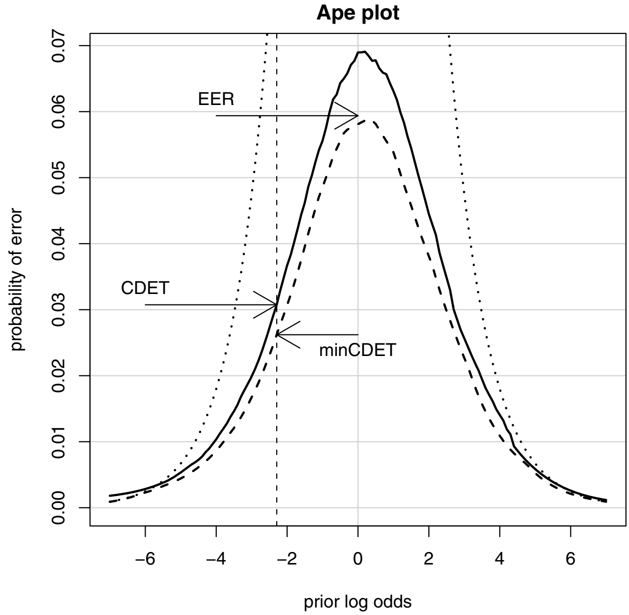


Fig. 4. APE-plot for our example system. Indicated are: $P_e(\theta)$ for observed $\mathcal{L}$ (solid curve), optimally calibrated $w(\mathcal{L})$ (dashed curve) and a reference detector (dotted curve).

- The derivation of C_{llr} as an integral of error-rate over the spectrum of applications.
- The importance of the EER as an application-independent indicator of discriminative ability.
- As discussed in more detail below, C_{llr} has the information-theoretic interpretation of being the amount of information that is lost between the input speech and the final decisions. The APE-curve is therefore a graphical demonstration of a relationship between information and error-rates—the more information you extract from the speech, the lower the error-rates will be.

Discussion. There is something interesting going on in the APE-curve around $\theta = 0$. On the one hand we see that P_e gives the biggest contribution to C_{llr} in this region. That would suggest that the task of the detector is hardest for $\theta \approx 0$, including the task of calibration. On the other hand, the benefit with respect to the reference detector is also the biggest in this region. Another way of phrasing this is that it seems that the information can be extracted from the speech signal most effectively when $\tilde{P}_{\mathrm{tar}} \approx 0.5$. For $|\theta| \gg 1$ there is already a lot of information in the prior, and it is difficult to add something useful by analyzing the speech signal, even though the probability of error is lower.

There is a further concern: it is also more difficult to accurately estimate error-rates when $|\theta| \gg 1$, because the absolute number of errors in these regions becomes small and eventually vanishes. So it seems the extreme regions of the APE-curve are regions where our detectors probably won't work so well, but also where we cannot estimate their performance accurately. In our APE-plots, we ignore these regions by not plotting them. This is just the same as is done with DET-curves. The horizontal and vertical axes of the DET-plot are infinite, but we always plot just a finite interesting region of this plot. Outside of this plot, the DET-curve becomes increasingly jagged, which is an indication of poor error-rate estimates.

The saving grace is that there are real-life effects that force reasonable applications to lie close to $\theta = 0$. There may certainly be applications where the prior P_{tar} becomes very small. But when things become scarce, their value generally increases. This means the cost of missing scarce events increases as the prior becomes smaller. Now recall (6) and note that a decrease in P_{tar} will be compensated for by an increase in C_{miss}, leaving θ approximately unchanged. Conversely, a similar argument shows that when $1 - P_{\text{tar}}$ becomes small, then C_{FA} would increase to compensate, again tending to keep θ roughly constant. It does therefore seem to make sense to concentrate our efforts to the benign central region of the APE-curve (or the corresponding region of the DET-curve).

2.8 Information-Theoretic Interpretation of C_{llr}

We have introduced C_{llr} as an integral of C_{det} over the spectrum of applications, but as hinted above, C_{llr} can be also be interpreted as a measure of *loss of information* [4].

Again, we will not do a rigorous information-theoretic derivation, but rather show informally how $1 - C_{\text{llr}}$ can be interpreted as the average information per trial (in bits of Shannon's entropy) that is gained by applying the detector. The information extracted by the detector from the speech is dependent on what is already known before considering the speech. This *prior knowledge* is encapsulated in the prior, P_{tar}. When $P_{\text{tar}} = 0$, or $P_{\text{tar}} = 1$, then there is already certainty about the speaker hypothesis and the detector cannot change this—the posterior will also be 0 or 1. However, values of P_{tar} between these extremes leaves a degree of prior uncertainty, up to a maximum of 1 bit where $P_{\text{tar}} = 0.5$. This maximum prior uncertainty is the reference level against which C_{llr} measures the information that the detector can extract from the speech. The information extracted from the speech by the detector, namely $1 - C_{\text{llr}}$ bits per trial, behaves in the following way:

- A (theoretically) perfect detector has $C_{\text{llr}} = 0$ and therefore $1 - C_{\text{llr}} = 1$, so it extracts *all* the information for every trial, transforming the prior uncertainty to posterior *certainty* in every case.
- A good, well-calibrated, real-life detector has $0 < C_{\text{llr}} < 1$, extracting an amount of information somewhere between 0 and 1 bit per trial.
- The reference detector which does not process the input speech has $C_{\text{llr}} = 1$ and therefore extracts 0 bits of information from every trial.

– A very badly calibrated[21] detector can do worse than this, having $C_{\text{llr}} > 1$, therefore extracting a *negative* amount of information. The negative sign indicates that on average over the APE-curve, the detector under evaluation has a higher error-rate than the reference detector. In this case it is therefore detrimental to use the detector and it is obviously better not to use (or at least to go and re-calibrate) the detector, because one could do better by just using the reference detector.

2.9 Comparison of Systems: DETs and APEs

Let us end this chapter with an example of the use of C_{llr} and APE-plots for comparing systems or conditions. This, in the end, is one of the key reasons to perform evaluations. To this purpose we use the data of two systems under evaluation of NIST SRE 2006 [5] which both may be called state of the art. The first system (which we have seen in earlier figures) consists of a single detector, the second system consists of the fusion of 10 separate detectors, of which the first system is one.

We further compare two evaluation conditions. The first condition includes trials with speech spoken in several languages, while the second condition has the subset of the trials where both speech segments are English.

We first look qualitatively at the DET-plot of three system/conditions in Fig. 5. Note how the DET warping of axes separates the three curves comfortably in the plot[22].

If we now inspect the curves more closely, we see that in terms of discrimination ability, the fused system performs favourably compared to the single system. Similarly we can conclude that, for the fused system, the English only trials were easier to discriminate than the whole collection of trials including several languages. (It does not really make sense to compare the upper and the lower curve, since both system and condition are different.) As for calibration, we can only conclude that for the NIST DCF the calibration was reasonable, and possibly better for the English only condition. We can finally observe that the lowest curve gets a bit noisy because a relatively low number of errors are made. For the English-only condition we have less than 30 target trial errors around $P_{\text{miss}} < 1.4\,\%$, so that if we apply George Doddington's 'rule of 30' [17] we find that for these low miss probabilities we are less than 90 % confident that the true P_{miss} is within 30 % of the observed P_{miss}.

We next look at the same systems evaluated on the same data, but depicted in APE-plots in Fig. 6. Here we have included a bar-graph of the C_{llr} and its

[21] It is only calibration problems that can cause $C_{\text{llr}} > 1$. If we remove calibration effects, considering only the discriminative ability of the detector, we find $0 \leq C_{\text{llr}}^{\text{min}} \leq 1$.

[22] With many different systems or conditions, the number of curves in a DET-plot is more often than not limited by the number of colours and/or line types. Also notice that the legend in the plot enumerates the curves in the same top-to-bottom order as the curves appear in the plot, i.e., according to the EER. (This practice is unfortunately not followed by all authors.)

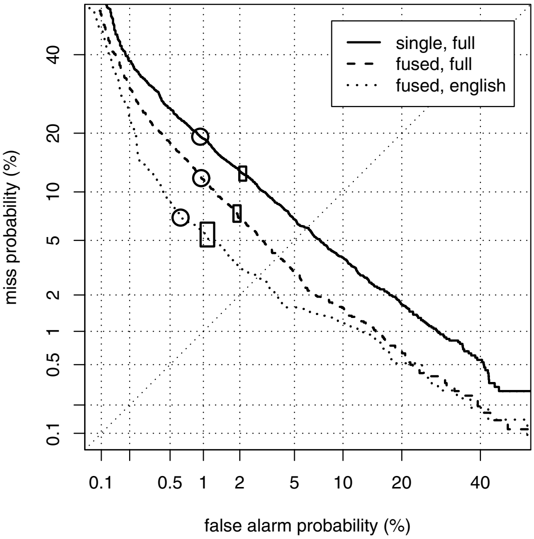


Fig. 5. A DET-plot for three system/conditions. From top to bottom: Single system, all trials; Fused system, all trials; and fused system, English trials. Notice that the upper and lower curve should not be compared with each other.

decomposition into discrimination and calibration loss, expressed in bits. The scales of the figures are the same, so that values can be compared visually. We can observe that although the fused system has much better discrimination power than the single system, the calibration error is roughly the same. Similarly, restricting trials to only English has a bigger effect on the discrimination than on the calibration. From the APE-curves we can learn that there is still quite some calibration performance to be gained for the fused system, especially at $\theta = 0$. All systems/conditions seem to suffer from being 'worse than the reference system' at very low θ.

One difference between DET and APE is the way that inaccuracies due to the limited number of trials show up. The curve in a DET-plot usually becomes ragged at the ends due to the low number of errors involved, showing that at each end, respectively P_{miss} or P_{FA} is poorly estimated. The fact that this effect is visible on the plot is a consequence of the magnification of small probabilities by the probit scale used in the DET-curve. In the APE-curve we do not see these effects, because when either P_{miss} or P_{FA} is poorly estimated, their value on the vertical axis is also small. Since C_{llr} is the area under the APE-curve, we see that fortunately these inaccuracies contribute relatively little to the total C_{llr} integral. Having said this, we must also remark that the proportions of the numbers of target and non-target trials in a NIST evaluation typically is 1:10, which leads to almost optimum accuracy at the operating point defined by C_{det}—this may be observed from the roughly equal 95 %-confidence intervals in

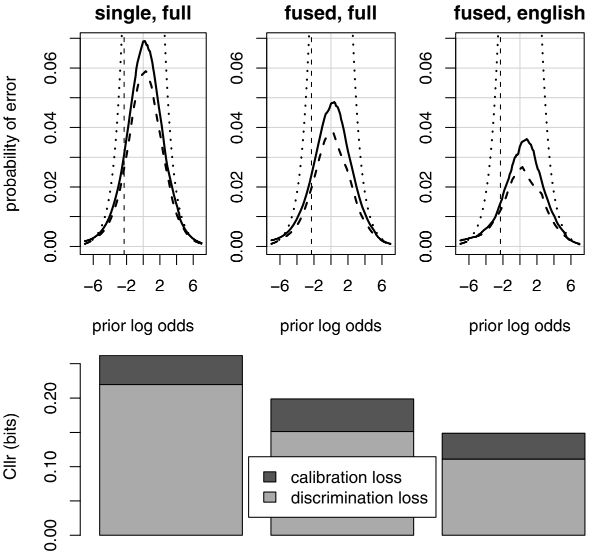


Fig. 6. APE-plots of the systems shown in Fig. 5. Note, that the graphs left and middle compare two systems, while the graphs middle and right compare two conditions.

the DET-plot around C_{det}. This 1:10 ratio has the effect that the left-hand side of the APE-plot is somewhat less noisy than the right-hand side.

3 Conclusion

We reviewed and appreciated the traditional measures that the speaker recognition community uses to assess the quality of automatic speaker recognition systems. The detection cost function C_{det} measures the application-readiness of a system for a particular application-type as defined by the parameters P_{tar}, C_{miss} and C_{FA}. NIST deserves credit for defining the task and evaluation measure and the progress that this has stimulated in the field. In particular, concentrating on detection rather than identification; and using expected cost, rather than error-rate for evaluation have had far-reaching effects. Moreover, the DET-curve, with its warped axes, show very well the trade-off between P_{FA} and P_{miss}, and allow for direct comparison of discrimination ability of many different systems or conditions in a single graph. Again, NIST deserves credit for introducing this type of analysis in the community—indeed, gradually DET-plots are being applied in other disciplines. Finally, when calibration is not an issue, the traditional EER remains a good single-valued summary of the discriminative capability of a detector. The utility of the EER as summary of discriminative ability can be appreciated in different ways in the DET and APE-plots.

We have further shown the limitations of C_{det} and $C_{\text{det}}^{\text{min}}$, in the sense that although they do measure calibration, they do so only in an application-dependent way. Of course, the DET-plot and the EER do not measure calibration.

Next, we reviewed the advantages of working with *log-likelihood-ratios* instead of merely with scores. Perhaps the most important advantage is that users can then set their own decision thresholds, where the thresholds are dependent only on properties of the application and not on the properties of the speaker detector. Despite these obvious and well-known advantages, the use of log-likelihood-ratio outputs in speaker recognition has not been common, presumably because such likelihood-ratio outputs are in practice subject to calibration problems, and without being able to measure these calibration problems, researchers had no good way to even start tackling this problem.

Our most important contribution in this chapter is therefore the introduction of a methodology to *measure the quality* of log-likelihood-ratios via C_{llr}. Moreover, we paid special attention to the issue of calibration, by forming a discrimination/calibration decomposition of C_{llr}. The practical calculation of C_{llr} via (9) is no more complex[23] than the traditional P_{miss} and P_{FA} calculations. The calculation of $C_{\text{llr}}^{\text{min}}$ is somewhat more complex, because it involves the PAV algorithm, but fortunately implementations are available to researchers, see e.g. [4].

Finally, we showed that the new metric C_{llr} has the interpretation not only as an integral of error-rates over the spectrum of applications, but also as the average information loss between speech input and decisions. This relationship is graphically demonstrated by the APE-plot, which indeed, for analysis of calibration, forms a useful complement to traditional DET-plots.

In conclusion, looking towards the future, it was announced at the June 2006 workshop of the NIST Speaker Recognition Evaluation that NIST intended to include the new measure C_{llr} as the primary evaluation measure in future evaluations. We hope this will stimulate more research on the subject of calibration, which is an important factor of the design of speaker recognition systems.

References

1. Martin, A.: Evaluations of Automatic Speaker Classification Systems. In: Müller, C. (ed.) Speaker Classification I. LNCS(LNAI), vol. 4343, Springer, Heidelberg (2007)
2. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of detection task performance. In: Proc. Eurospeech 1997, Rhodes, Greece, pp. 1895–1898 (1997)
3. Brümmer, N.: Application-independent evaluation of speaker detection. In: Proc. Odyssey, Speaker and Language recognition workshop, ISCA 2004, pp. 33–40 (2004)
4. Brümmer, N., du Preez, J.: Application-independent evaluation of speaker detection. Computer Speech and Language 20, 230–275 (2006)
5. NIST: The NIST year 2006 Speaker Recognition Evaluation Plan (2006), `http://www.nist.gov/speech/tests/spk/2006/index.htm`
6. Campbell, W.M., Reynolds, D.A., Campbell, L.P., Brady, K.J.: Estimating and evaluating confidence for forensic speaker recognition. In: Proc. ICASSP, pp. 717–720 (2005)

[23] With due respect for some numerical accuracy issues.

7. Campbell, W.M., Brady, K.J., Campbell, J.P., Granvile, R., Reynolds, D.A.: Understanding scores in forensic speaker recognition. In: Proc. Odyssey 2006 Speaker and Language Recognition Workshop (2006)
8. Ramos-Castro, D., González-Rodríguez, J., Ortega-Garcia, J.: Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. In: Proc. Odyssey 2006 Speaker and Language Recognition Workshop (2006)
9. Brümmer, N., van Leeuwen, D.A.: On calibration of language recognition scores. In: Proc. Odyssey 2006 Speaker and Language recognition workshop (2006)
10. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independetn speaker verification systems. Digital Signal Processing 10, 42–54 (2000)
11. Navrátil, J., Ramsawamy, G.N.: The awe and mistery of t-norm. In: Proc. Eurospeech, pp. 2009–2012 (2003)
12. Van Leeuwen, D.A., Martin, A.F., Przybocki, M.A., Bouten, J.S.: NIST and TNO-NFI evaluations of automatic speaker recognition. Computer Speech and Language 20, 128–158 (2006)
13. Swets, J.A.: Signal detection and recognition by human observers; contemporary readings. Wiley, New York (1964)
14. Green, D.M., Swets, J.A.: Signal Detection Theory and Psychophysics. Wiley, New York (1966)
15. Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, New York (1994)
16. DeGroot, M., Fienberg, S.: The comparison and evaluation of forecasters. The Statistician, 12–22 (1983)
17. Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A.: The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. Speech Communication 31, 225–254 (2000)

Author Index