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

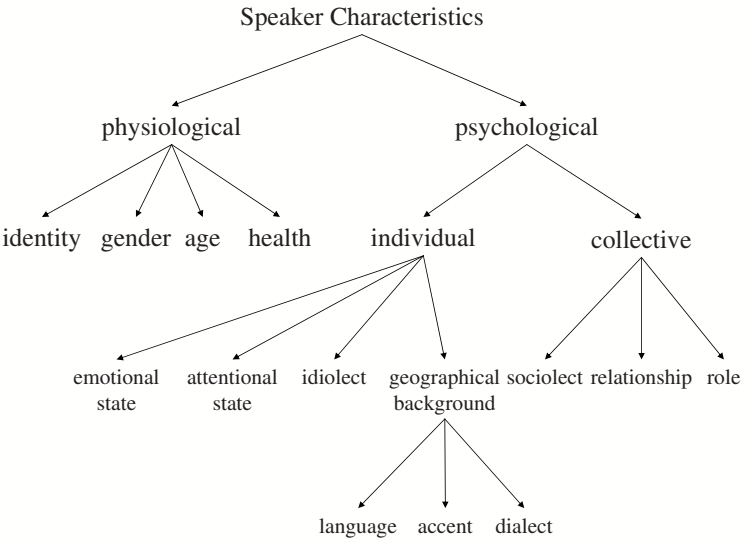

Fig. 1. Taxonomy of Speaker Characteristics

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.

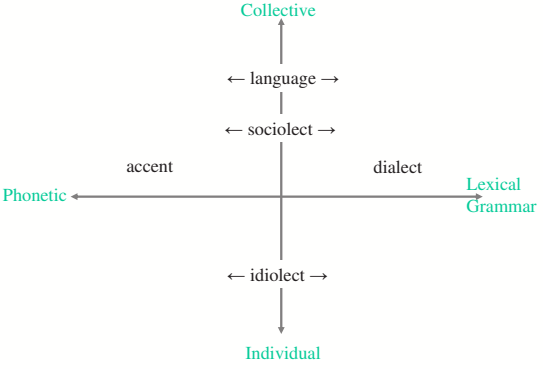

Fig. 2. Language-dependent Characteristics

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

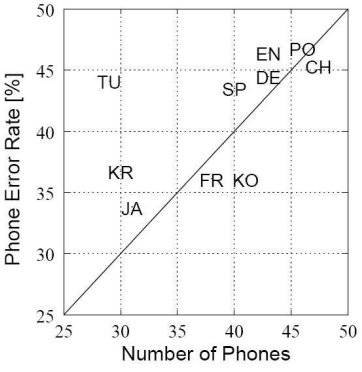

Fig. 3. Error rate vs number of phones for ten GlobalPhone languages

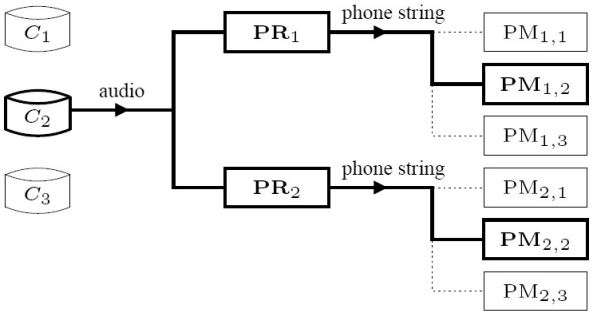

Fig. 4. Training of feature-specific phonetic models for 2 phone recognizers and a 3–class problem

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

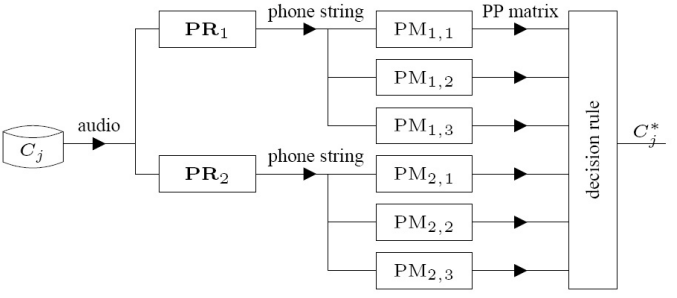

Fig. 5. MPM-pp classification block diagram

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 {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.

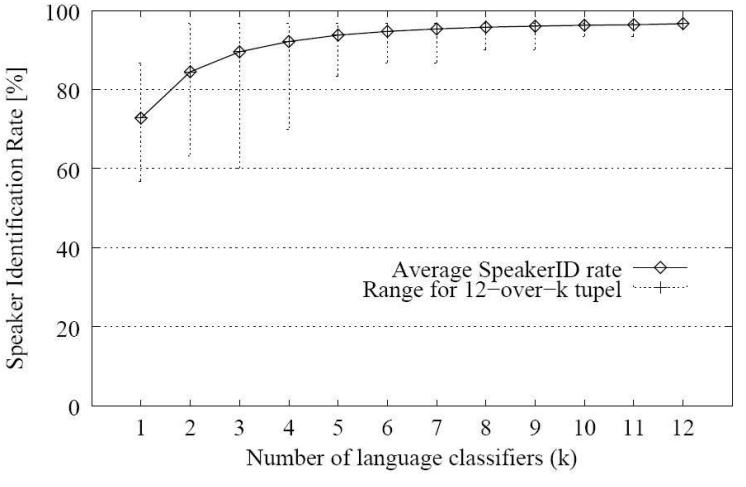

Fig. 6. Classification rate over number of phone recognizers

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 in-

dependent 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: Nonnative 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)