Doroteo Torre Toledano Alfonso Ortega Giménez
António Teixeira Joaquín González Rodríguez
Luis Hernández Gómez Rubén San Segundo Hernández
Daniel Ramos Castro (Eds.)

Communications in Computer and Information Science 328

Advances in Speech and Language Technologies for Iberian Languages

IberSPEECH 2012 Conference
Madrid, Spain, November 2012
Proceedings

Springer

Communications in Computer and Information Science 328

Doroteo Torre Toledano Alfonso Ortega Giménez
António Teixeira Joaquín González Rodríguez
Luis Hernández Gómez Rubén San Segundo Hernández
Daniel Ramos Castro (Eds.)

Advances in Speech and Language Technologies for Iberian Languages

IberSPEECH 2012 Conference
Madrid, Spain, November 21-23, 2012
Proceedings

Springer

Volume Editors

Doroteo Torre Toledano
Universidad Autonoma de Madrid, Spain
E-mail: doroteo.torre@uam.es

Alfonso Ortega Giménez
Universidad de Zaragoza, Spain
E-mail: ortega@unizar.es

António Teixeira
Universidade de Aveiro, Portugal
E-mail: ajst@ua.pt

Joaquín González Rodríguez
Universidad Autonoma de Madrid, Spain
E-mail: joaquin.gonzalez@uam.es

Luis Hernández Gómez
Universidad Politécnica de Madrid, Spain
E-mail: luis@gaps.ssr.upm.es

Rubén San Segundo Hernández
Universidad Politécnica de Madrid, Spain
E-mail: lapiz@die.upm.es

Daniel Ramos Castro
Universidad Autonoma de Madrid, Spain
E-mail: daniel.ramos@uam.es

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-35291-1 e-ISBN 978-3-642-35292-8
DOI 10.1007/978-3-642-35292-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012952611

CR Subject Classification (1998): I.2.7, I.2.6, H.5.1-2, H.5.5, I.5.2-4, I.7.5, I.2.1, J.5

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

It was a pleasure and an honor to organize the *IberSPEECH 2012: Joint VII "Jornadas en Tecnología del Habla" and III Iberian SLTech Workshop*, that took place during November 21–23 in Madrid, Spain, hosted by the ATVS Biometric Research Group, Universidad Autónoma de Madrid.

This conference is the result of the merging of two conferences: the *Jornadas en Tecnología del Habla* (Spanish Speech Technology Workshop) and the Iberian SLTech Workshop. The first has been organized by the *"Red Temática en Tecnología del Habla"* (Spanish Speech Technology Thematic Network, http://www.rthabla.es) since 2000. This network was created in 1999 and currently includes over 200 researchers and 30 research groups in speech technology in Spain. The first Iberian SLTech Workshop was organized in Porto Salvo, Portugal, in 2009, by the Special Interest Group on Iberian Languages (SIG-IL, http://www.il-sig.org/) of the International Speech Communication Association (ISCA, http://www.isca-speech.org) and has been organized in conjunction with the *"Jornadas en Tecnología del Habla"* since 2010.

As a result, the *IberSPEECH: Joint "Jornadas en Tecnología del Habla" and Iberian SLTech Workshop* is one of the most important research meetings in the field of speech and language processing focusing on Iberian languages, attracting many researchers (about 120 in last edition), mainly from Spain and Portugal, and is also a natural meeting for researchers from Latin America. However, although the main focus is on Iberian languages and the Iberian region, the conference is not restricted to them. Proof of this are the ALBAYZIN Technology Competitive Evaluations, organized in conjunction with the conference, which in this edition attracted the interest of several research groups from all around the world, including the USA, UK, France, Japan, China, and Switzerland, among others.

The ALBAYZIN Technology Competitive Evaluations have been organized alongside with the conference since 2006, promoting the fair and transparent comparison of technology in different fields related to speech and language technology. In this edition we had five different evaluations: Language Recognition, Audio Segmentation, Speech Synthesis, Search on Speech, and Handwriting Recognition. The organization of each of these evaluations requires preparing development and test data, providing data along with a clear set of rules to the participants, and gathering and comparing results from participants. This organization was carried out by different groups of researchers and was crucial for the success of the evaluations. Although results from the evaluations cannot be included in this volume owing to timing restrictions, we would like to express our gratitude to the organizers and also to the participants in the evaluations.

In this edition we had over 80 articles submitted to the conference, and only 29 were selected for this publication. This selection was based on the scores and comments provided by our Scientific Review Committee, which includes over 75 researchers from different institutions mainly from Spain, Portugal, and Latin America, to which we also would like to express our deepest gratitude. Each article was reviewed by three different reviewers and the authors had some time to address the comments before submitting the camera-ready paper. The articles have been organized, following the oral sessions of the conference, into six different topics:

- Speaker Characterization and Recognition
- Audio and Speech Segmentation
- Pathology Detection and Speech Characterization
- Dialogue and Multimodal Systems
- Robustness in Automatic Speech Recognition
- Applications of Speech and Language Technologies

Besides the excellent research articles included in this volume, the conference had the pleasure of having three extraordinary keynote speakers: Jan "Honza" Cernocky (Brno University of Technology, BUT, Czech Republic), Philip Rose (Australian National University, Australia), and Pedro Moreno (Google Research, NY, USA).

The conference was mainly organized and supported by the Spanish Thematic Network on Speech Technology *("Red Temática en Tecnología del Habla")* and the ISCA Special Interest Group on Iberian Languages (SIG-IL). Besides this, we also received support from the Universidad Autónoma de Madrid (UAM) and the Campus Internacional Excelencia UAM+CSIC, which not only provided a fantastic venue for organizing the conference (the Escuela Politécnica Superior), but also financial support. Also, several companies provided financial support for the conference, including Google, Microsoft, and Telefónica (through Cátedra UAM-Telefónica). Last but not least, we had financial support from the MA2VICMR consortium. Without the financial support of all of them this conference would simply have not been possible.

We would also like to thank Springer, and in particular Alfred Hoffmann and Leonie Kunz, for the possibility of publishing this volume and their help and great work in preparing it. This will help increase the international impact of this conference.

Finally, we would like to thank all the people that have been putting their efforts in organizing this conference, and in particular the Organizing Committee and the local Organizing Committee, as well as all the authors that have presented their articles at the conference.

Doroteo Torre Toledano
Alfonso Ortega Giménez
António Teixeira

Organization

General Chairs

Doroteo Torre Toledano	Universidad Autónoma de Madrid, Spain
Alfonso Ortega Giménez	Universidad de Zaragoza, Spain
António Teixeira	Universidade de Aveiro, Portugal

Program Chairs

Joaquín González Rodríguez	Universidad Autónoma de Madrid
Rubén San Segundo Hernández	Universidad Politécnica de Madrid
Luis Hernández Gómez	Universidad Politécnica de Madrid

Publication Chair

Daniel Ramos Castro	Universidad Autónoma de Madrid

Demo Chairs

Daniela Braga	Microsoft (MLDC)
Alberto Abad Gareta	INESC

Award Chairs

Miguel Sales Dias	Microsoft (MLDC)
Climent Nadeu Camprubi	Universitat Politècnica de Catalunya

Evaluation Chairs

Javier Tejedor Noguerales	Universidad Autónoma de Madrid
Javier González Domínguez	Universidad Autónoma de Madrid

Scientific Review Committee

Alberto Abad Gareta	INESC
Olatz Arregi Uriarte	Euskal Herriko Unibertsitatea
José Miguel Benedí Ruiz	Universitat Politécnica de Valencia
Mª Carmen Benítez Ortúzar	Universidad de Granada
Antonio Bonafonte Cávez	Universitat Politècnica de Catalunya
Daniela Braga	MLDC/Microsoft

Francisco Campillo	Universidad de Vigo
Antonio Cardenal	Universidade de Vigo
Valentín Cardeñoso Payo	Universidad de Valladolid
Paula Carvalho	Universidade de Lisboa
Francisco Casacuberta Nolla	Universitat Politécnica de Valencia
María José Castro Bleda	Universitat Politécnica de Valencia
José Colás Pasamontes	Universidad Autónoma de Madrid
Ricardo de Córdoba Herralde	Universidad Politécnica de Madrid
Carmen de la Mota Gorriz	Universitat Autònoma de Barcelona
Ángel de la Torre Vega	Universidad de Granada
Laura Docío Fernández	Universidade de Vigo
Daniel Erro Eslava	Euskal Herriko Unibertsitatea
David Escudero Mancebo	Universidad de Valladolid
Rubén Fernández Pozo	Universidad Politécnica de Madrid
Javier Ferreiros López	Universidad Politécnica de Madrid
Ascensión Gallardo Antolín	Universidad Carlos III de Madrid
Fernando García Granada	Universitat Politécnica de Valencia
Carmen García Mateo	Universidade de Vigo
Marta Gatius Vila	Universitat Politècnica de Catalunya
Juan Ignacio Godino Llorente	Universidad Politécnica de Madrid
Koldo Gojenola Galletebeitia	Euskal Herriko Unibertsitatea
Pedro Gómez Vilda	Universidad Politécnica de Madrid
Javier González Domínguez	Universidad Autónoma de Madrid
Joaquín González Rodríguez	Universidad Autónoma de Madrid
Inma Hernaez Rioja	Euskal Herriko Unibertsitatea
Luis A. Hernández Gómez	Universidad Politécnica de Madrid
Fco. Javier Hernando Pericas	Universitat Politècnica de Catalunya
Lluís Felip Hurtado Oliver	Universitat Politécnica de Valencia
Eduardo Lleida Solano	Universidad de Zaragoza
Joaquím Llisterri Boix	Universitat Autònoma de Barcelona
Fernando Jesús López Colino	Universidad Autónoma de Madrid
Eduardo López Gonzalo	Universidad Politécnica de Madrid
María Teresa López Soto	Universidad de Sevilla
Ramón López-Cózar Delgado	Universidad de Granada
Nuno Mamede	INESC
José B. Mariño Acebal	Universitat Politècnica de Catalunya
Carlos David Martínez Hinarejos	Universitat Politécnica de Valencia
Helena Moniz	INESC
Juan Manuel Montero Martínez	Universidad Politécnica de Madrid
Nicolás Morales Mombiela	Nuance
Asunción Moreno Bilbao	Universitat Politècnica de Catalunya
Climent Nadeu Camprubi	Universitat Politècnica de Catalunya
Juan Luis Navarro Mesa	Universidad de Las Palmas de Gran Canaria
Eva Navas Cordón	Euskal Herriko Unibertsitatea
Juan Nolazco Flores	Tecnológico de Monterrey

Alfonso Ortega Giménez	Universidad de Zaragoza
Antonio Miguel Peinado Herreros	Universidad de Granada
Carmen Peláez Moreno	Universidad Carlos III de Madrid
José Luis Pérez Córdoba	Universidad de Granada
Paulo Quaresma	Universidade de Évora
Daniel Ramos Castro	Universidad Autónoma de Madrid
Andreia Rauber	Universidade Católica de Pelotas
José Adrián Rodríguez Fonollosa	Universitat Politècnica de Catalunya
Eduardo Rodríquez Banga	Universidade de Vigo
Luis Javier Rodríguez Fuentes	Euskal Herriko Unibertsitatea
Antonio José Rubio Ayuso	Universidad de Granada
Victoria Eugenia Sánchez Calle	Universidad de Granada
Joan Andreu Sánchez Peiró	Universitat Politécnica de Valencia
Emilio Sanchís Arnal	Universitat Politécnica de Valencia
Rubén San-Segundo Hernández	Universidad Politécnica de Madrid
Kepa Sarasola Gabiola	Euskal Herriko Unibertsitatea
Ibon Saratxaga Couceiro	Euskal Herriko Unibertsitatea
Encarnación Segarra Soriano	Universitat Politécnica de Valencia
Mário Silva	Universidade de Lisboa
Alberto Simões	ESEIG/IPP
Daniel Tapias Merino	SIGMA Technologies
António Teixeira	Universidade de Aveiro
Javier Tejedor Noguerales	Universidad Autónoma de Madrid
Doroteo Torre Toledano	Universidad Autónoma de Madrid
María Amparo Varona Fernández	Euskal Herriko Unibertsitatea
José Luis Vicedo González	Universitat d'Alacant

Local Organizing Committee

Doroteo Torre Toledano	Universidad Autónoma de Madrid
Joaquín González Rodríguez	Universidad Autónoma de Madrid
Daniel Ramos Castro	Universidad Autónoma de Madrid
Javier González Domínguez	Universidad Autónoma de Madrid
Javier Franco Pedroso	Universidad Autónoma de Madrid
Javier Tejedor Noguerales	Universidad Autónoma de Madrid
Fernando Jesús López Colino	Universidad Autónoma de Madrid
Javier Ortega García	Universidad Autónoma de Madrid
Julián Fiérrez Aguilar	Universidad Autónoma de Madrid
Javier Galbally Herrero	Universidad Autónoma de Madrid
Rubén Vera Rodríguez	Universidad Autónoma de Madrid
Pedro Tomé González	Universidad Autónoma de Madrid
Leonardo Campillo Llanos	Universidad Autónoma de Madrid
Olga León Zurdo	Universidad Autónoma de Madrid
Antonio Moreno Sandoval	Universidad Autónoma de Madrid
Sara Antequera	Universidad Autónoma de Madrid

Table of Contents

Speaker Characterization and Recognition

Audio and Speech Segmentation

Pathology Detection and Speech Characterization

Dialogue and Multimodal Systems

Robustness in Automatic Speech Recognition

Applications of Speech and Language Technologies

Reliability Estimation of the Speaker Verification Decisions Using Bayesian Networks to Combine Information from Multiple Speech Quality Measures

Jesús Villalba, Eduardo Lleida, Alfonso Ortega,
and Antonio Miguel

Communications Technology Group (GTC),
Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba,lleida,ortega,amiguel}@unizar.es

Abstract. In some situations the quality of the signals involved in a speaker verification trial is not as good as needed to take a reliable decision. In this work, we use Bayesian networks to model the relations between the speaker verification score, a set of speech quality measures and the trial reliability. We use this model to detect and discard unreliable trials. We present results on the NIST SRE2010 dataset artificially degraded with different types and levels of additive noise and reverberation. We show that a speaker verification system, that is well calibrated for clean speech, produces an unacceptable actual DCF on the degraded dataset. We show how this method can be used to reduce the actual DCF to values lower than 1. We compare results using different quality measures and Bayesian network configurations.

Keywords: speaker recognition, Bayesian networks, reliability, quality.

1 Introduction

In some situations, the quality of the signals involved in the speaker verification process is not as good as needed to take a reliable decision. This causes a dramatic drop of the system performance. The purpose of this work is finding a method to estimate the reliability of the speaker verification (SV) decision for each trial. We intend to discard the unreliable trials in order to be able to assure that the decisions taken with the trials that we keep produce low error rates. We infer the trial reliability from a set of measures, that we call quality measures, extracted from the training and testing segments of the trial.

In the last years, several approaches have been proposed to combine different sources of information into a global confidence measure [1], [2], [3]. We work on the approach introduced by Richiardi in [4] and [5]. In these works, a Bayesian network is used to obtain a probabilistic measure of the reliability of the trial given the speaker verification score and some quality measures. In case of low reliability, the system asks the user to utter a new sentence and chooses the

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 1–10, 2012.

one with higher reliability. In [6], this approach is compared with confidence measures presented in previous works [7] [8] [9] showing that the probabilistic approach outperforms previous approaches. In [10], the same technique is used to detect the reliability of several biometric modalities in a multimodal identity verification system. Then, the reliability estimations are used to improve the fusion of the scores of the different systems.

In this paper, we extend Richiardi's work introducing other quality measures and comparing the performance of these measures detecting the trial reliability. Besides, we introduce and compare other configurations of the Bayesian network. We use this model to discard unreliable trials on an artificially degraded version of the NIST SRE10 dataset achieving an important improvement of the actual DCF.

The rest of the paper is organized as follows: Section 2 describes the quality measures. Section 3 describes the Bayesian networks used for reliability estimation. Section 4 describes the experimental setup and shows the results. Finally, Section 5 shows the conclusions.

2 Speech Quality Measures

2.1 Signal/Noise Ratio

Additive noise is known to have a negative impact on speaker verification performance. We measure the SNR using a method that takes advantage of the periodicity properties of voiced speech intervals. The most part of the energy of voiced speech is concentrated in multiples of its pitch frequency while additive noise has a more uniform frequency distribution. This allows to get an estimation of the clean and noise signals separately using the adapted comb filters H_s and H_n respectively:

$$H_s(z,t) = \frac{0.5z^{T_p(t)} + 1 + 0.5z^{-T_p(t)}}{1 - \alpha_s z^{-T_p(t)}} \quad H_n(z,t) = \frac{-0.5z^{T_p(t)} + 1 - 0.5z^{-T_p(t)}}{1 + \alpha_n z^{-T_p(t)}} \tag{1}$$

where $T_p(t)$ is the pitch period at time t and, α_s and α_n are coefficients that modify the bandwidth of the filter. We set $\alpha_s = 0.25$ and $\alpha_n = 0.7$. As the pitch period changes along the speech segment these are time varying filters. We had used a pitch estimator based on the RAPT algorithm [11].

The SNR for a frame t is calculated from the power of the clean and noisy signal estimations. Finally we, average the SNR over all voiced frames.

2.2 Modulation Index

The modulation index at time t is calculated as

$$Indx(t) = \frac{v_{max}(t) - v_{min}(t)}{v_{max}(t) + v_{min}(t)} \, . \tag{2}$$

where $v(t)$ is the envelope of the signal and $v_{max}(t)$ and $v_{min}(t)$ are the local maximum and minimum of the envelope in the region close to time t. The envelope is approximated by the absolute value of the signal $s(t)$ down sampled to 60 Hz as in [12]. The envelope of a recording with noise or reverberation should have higher local minimums and, therefore, a lower modulation index. We average the index over all speech frames.

2.3 Spectral Entropy

Entropy is a measure related to the peakiness or flatness of a probability distribution. The spectral entropy is computed over the values of the short-term power spectrum, where the spectral values are normalized to sum 1 and thus, forming a pdf. Thus, the entropy for a frame t is calculated as follows:

$$H(t) = -\sum_{\omega} \frac{|X(\omega,t)|^2}{\sum_{\omega'} |X(\omega',t)|^2} \log \frac{|X(\omega,t)|^2}{\sum_{\omega'} |X(\omega',t)|^2} \tag{3}$$

where $|X(\omega,t)|^2$ is the power spectrum of the signal. The use of the entropy relies in the assumption that a clean signal should have a more organized spectrum, while a noisy signal should have a flatter spectrum. We average the entropy over all speech frames.

2.4 UBM Likelihood

The Universal Background Model (UBM) is a GMM that represents the probability distribution of the speech features of the development database, that is usually a good quality database. Speaker models are adapted from this UBM. Degraded signals are more likely to differ from the UBM than non-degraded ones. Therefore, they will produce a worse estimation of the speaker models. Thus, the likelihood of the utterance given the UBM can be used as a measure of speech degradation. This measure was first used in [13] with some good results.

3 Bayesian Networks for Reliability Estimation

In order to estimate a global measure of the trial reliability from the quality measures we have adopted an approach similar to that used in [4] and [5]. These works use a Bayesian network (BN) to model the relationships between the random variables involved in the verification process. A Bayesian network is a directed graphical model [14] that describes the dependencies of a set of random variables. Figure 1 shows the Bayesian Network that describes our problem. Empty nodes denote *hidden variables*, shaded nodes denote *observed variables* and small solid nodes denote *deterministic parameters*. A node or group of nodes surrounded by a box, called a *plate*, labelled with N indicates that there are N nodes of that kind (for example N trials). The arcs between the nodes point from

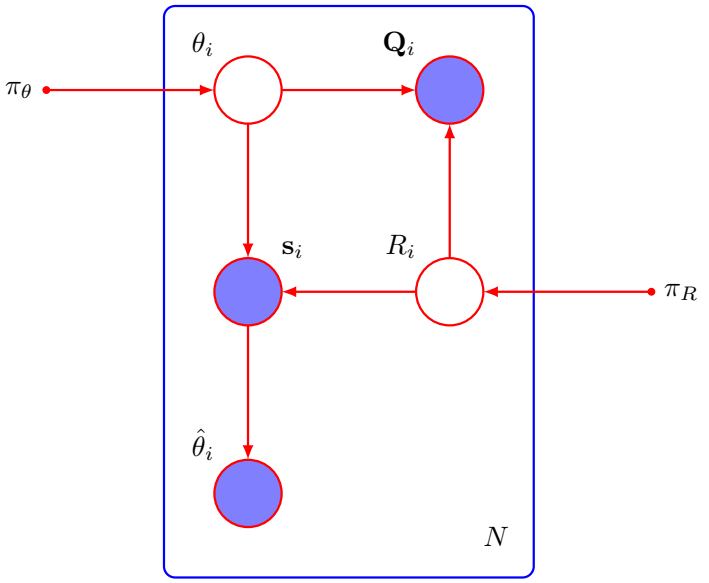


Fig. 1. BN for reliability estimation based on score and quality measures

the parent variables to their children variables. They represent the conditional dependencies between parents and children.

Following, we introduce the variables included in the graph. s is the SV score. $\mathbf{Q}$ are the quality measures related to one trial. $\theta \in \{\mathcal{T}, \mathcal{N}\}$ is the label of the trial, where $\mathcal{T}$ is the hypothesis that the training and test segments belong to the same speaker and $\mathcal{N}$ to different speakers. $\hat{\theta}$ is the SV decision after applying a threshold ξ_θ to s. $R \in \{\mathcal{R}, \mathcal{U}\}$ is the reliability of the trial, where $\mathcal{R}$ is the hypothesis that the decision is reliable and $\mathcal{U}$ unreliable. $\pi_\theta = (P_\mathcal{T}, P_\mathcal{N})$ is the hypothesis prior where $P_\mathcal{T}$ is the target prior and $P_\mathcal{N} = 1 - P_\mathcal{T}$ the non-target prior. Finally, $\pi_R = (P_\mathcal{R}, P_\mathcal{U})$ is the reliability prior.

The BN allows to write the joint probability distribution of the variables as a product of conditional distributions:

$$P\left(s, \mathbf{Q}, R, \theta, \hat{\theta} | \pi_\theta, \pi_R\right) = P\left(s|R, \theta\right) P\left(\mathbf{Q}|R, \theta\right) P\left(\hat{\theta}|\theta, R\right) P\left(\theta|\pi_\theta\right) P\left(R|\pi_R\right). \tag{4}$$

Using (4) we can write the posterior distribution of R given the observable variables as

$$P\left(R|s, \mathbf{Q}, \hat{\theta}, \pi_\theta, \pi_R\right) = \frac{\sum_{\theta \in \{\mathcal{T}, \mathcal{N}\}} P\left(s, \mathbf{Q}, R, \theta, \hat{\theta} | \pi_\theta, \pi_R\right)}{\sum_R \sum_{\theta \in \{\mathcal{T}, \mathcal{N}\}} P\left(s, \mathbf{Q}, R, \theta, \hat{\theta} | \pi_\theta, \pi_R\right)} \tag{5}$$

The distributions $P\left(s|R,\theta,\hat{\theta}\right)$ are modelled by Gaussians and $P\left(\mathbf{Q}|R,\theta\right)$ by GMM. $P\left(\hat{\theta}|\theta,R\right)$ is a discrete distribution that is 1 if $\hat{\theta}=\theta$ and $R=\mathcal{R}$ or if $\hat{\theta}\neq\theta$ and $R=\mathcal{U}$ and it is 0 otherwise.

Our Bayesian network differs from that in [5] in which ours adds a link from θ to $\mathbf{Q}$. In this manner, we suppose that the speech degradation affects differently to targets and non-target trials. In Section 4, we show results comparing both BN configurations.

The previous model estimates the reliability based on the score and quality measures. However, We would like to assess the ability of the quality measures to estimate the reliability on their own without using the score. For that, we remove s from the network. The joint distribution of this other network is the same as in equation (4) without the term $P\left(s|R,\theta\right)$.

4 Experiments

4.1 Database

We take the telephone part of NIST SRE08 and SRE10 databases assuming that they are quite clean. Then, we have created a synthetic database degrading NIST with different noise levels and reverberation times.

Dataset with Additive Noise. The dataset with additive noise has been created with a similar protocol than the Aurora2 dataset [15]. We have added different Aurora2 noises to enrollment and test:

- Enrollment: suburban train, babble, car and exhibition hall.
- Test: restaurant, street, airport and train station.

The noises are previously filtered by the ITU MIR telephone frequency response to simulate that they have pass through a telephone channel. The noise for each file is selected randomly.

We have used the open source FaNT Tool [16] for adding noise to the signals. We have signal-to-noise ratios of 20dB, 15dB, 10dB, 5dB and 0dB.

Dataset with Reverberation. In order to create the reverberant dataset we have used a free Matlab package based on [17]. This package includes two tools:

- RIR: calculates the impulse response of a rectangular room given the room dimensions, the reflection coefficients of the walls and the speaker and microphone locations.
- FCONV: used to convolve the room impulse response (RIR) with the clean signal.

We created random room impulse responses with the following criteria:

- 8 sizes of room, from small room to basketball court.
- Add a random number to the room size to change it ±50%.
- 8 different materials for the walls: rubber, granite, clay, concrete, steel, aluminium, brick and glass.
- Random speaker position inside the room.
- Random mic position in a square with 4m. of side around the speaker.

For each RIR we compute the reverberation time (RT) as the time that the filter energy takes to fall 60dB. We assign each RIR to one of 8 groups by the nearest reverberation time among 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75 and 1 second. Each RIR is used only to degrade one file. The RIR for each file is selected randomly.

4.2 Experimental Setup

We trained the Bayesian network with trials from the SRE08 dataset and tested on SRE10. The training set includes all trials that can be done scoring all SRE08 telephone training segments versus all SRE08 telephone test segments for all SNR and RT pairs. This dataset has 1269 target trials (424 male, 845 female) and 766605 non-target trials (176090 male, 590515 female) per noisy condition.

The test set is the core det5 (phn-phn) condition of SRE10. This dataset has 708 target trials (353 male, 355 female) and 29665 non-target trials (13707 male, 15958 female) per noisy condition.

The SV baseline system is based on i-vectors with two-covariance model. We have used 400 dimensional i-vectors. They are extracted using 20 short-time Gaussianized MFCC plus deltas and double deltas and a 2048 component diagonal covariance UBM. The UBM, the i-vector extractor and the two-covariance model are gender independent and they were trained using telephone data from SRE04, SRE05 and SRE06. The i-vectors preprocessing includes centering, whitening and length normalization.

The SV verification scores are calibrated using the *Bosaris Toolkit* to optimize the old NIST operating point ($C_{Miss} = 10$, $C_{FA} = 1$, $P_{\mathcal{T}} = 0.01$). The calibration is trained with the clean part of the SRE08 dataset. Then, we use this calibration function on all the conditions of SRE08 and SRE10. We use the Bayes decision threshold (2.29). Thus, on the clean part of SRE10, we achieve an EER=2.2%, minDCF=0.14 and actDCF=0.17. When we pool all the noisy conditions we get minDCF=0.99 and actDCF=4.05. That means that our SV system is not useful any more. Our goal is to discard the unreliable trials in order to make the actDCF lower than 1.0 keeping the threshold that we set using only clean trials. Consequently, we show results that compare actDCF versus the number of trials that we keep. In this context the DCF is defined as

$$C_{DCF} = C_{Miss} P_{\mathcal{T}} P_{Miss\hat{\mathcal{R}}} + C_{FA}(1 - P_{\mathcal{T}}) P_{FA\hat{\mathcal{R}}} \tag{6}$$

where $P_{Miss\hat{\mathcal{R}}}$ and $P_{FA\hat{\mathcal{R}}}$ are computed on the trials that are classified as reliable.

$$P_{Miss\hat{\mathcal{R}}} = \frac{N_{Miss}}{N_{\hat{\mathcal{R}}\mathcal{T}}} \; P_{FA\hat{\mathcal{R}}} = \frac{N_{FA}}{N_{\hat{\mathcal{R}}\mathcal{N}}} \tag{7}$$

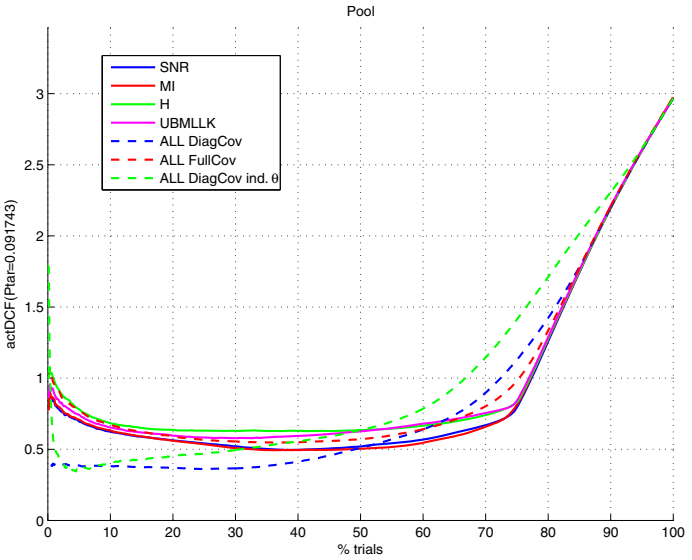


Fig. 2. % trials vs. actDCF in dataset with additive noise for BN with score and quality measures

4.3 Results with Additive Noise

In this section, we show results training and testing the Bayesian networks with the datasets with additive noise only. Figure 2 shows the actDCF versus the number of trials that we keep for the BN that estimates the reliability from the SV score and the quality measures. To plot this curves we put a varying threshold on the posterior probability of reliable shown in equation (5). The lower and steepy the curves are, the better. That means that we remove the worst trials first.

The distributions of the quality measures are mixtures of diagonal or full covariance Gaussians. We tried different number of mixture components and found that 4 components is enough.

The curves show that the system removes easily the first 30% of bad trials. These trials are the ones with the lowest SNR pairs. We get a dramatic improvement of actDCF from 2.96 to 0.75. All quality measures perform similarly in this range. From this point, we go on removing trials that are not so noisy but that are still causing verification errors. The best quality measures working alone are SNR and modulation index. With these measures we can get an actDCF=0.5 removing 50% of the trials. After that, the actDCF grows again. In a perfect reliability estimation system, the actDCF should always decrease, when we remove trials. However, in practice, there is a certain amount of reliable trials with low reliability estimations and vice verse. Then, if we remove a reliable trial and compute the error rates with equations (7), the denominator decreases and the numerator remains constant. Therefore, the error rate and the actDCF grow.

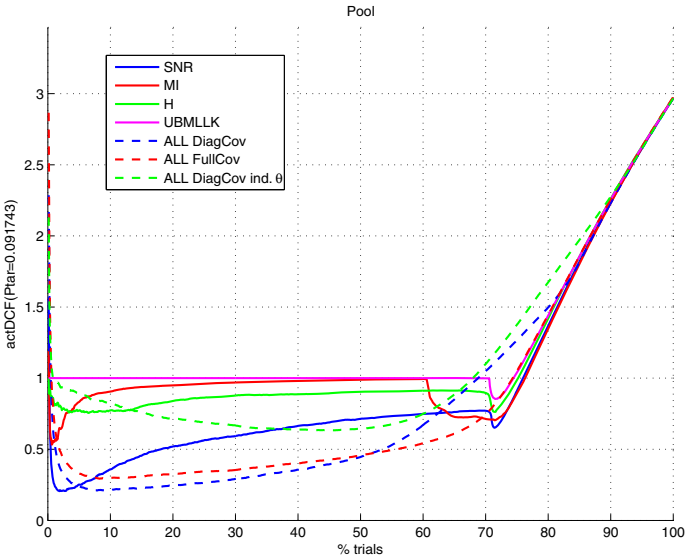


Fig. 3. % trials vs. actDCF in dataset with additive noise for BN with quality measures

In a good reliability estimation system the actDCF should decrease fast as we prune trials and it should not grow again unless most of the bad trials have been removed. We get the lowest actDCF=0.36, using all the quality measures together modeled by diagonal Gaussians and keeping only the 30% of the trials.

All curves but the last one (green dotted line), use a BN that assumes dependence between the quality measures and the trials labeling. We get better results assuming dependence. That confirms our belief that degradations affect differently to target and non-target trials.

Figure 3 shows the actDCF versus the number of trials for the BN that estimates the reliability from the quality measures only. Here, we have more difference between curves than in the previous figure. For the first 30% of trials removed, it performs similarly to the previous case. After, that results are worse. That indicates that the SV score is an important help for the reliability estimation. SNR is the best measure followed by Entropy, modulation index and, finally, UBM likelihood. When we combine all the quality measures we can achieve actDCF=0.21 removing 90% of the trials. Here, as in the previous figure, the results assuming that the quality measures depend on the trial labelling are better than assuming independence.

4.4 Results with Additive and Convolutional Noise

In this section, we show results training and testing the Bayesian network with the datasets with additive noise and reverberation. Figure 4 shows the actDCF versus the number of trials for the BN with SV score and quality measures. We pool all combinations of additive noise and reverberation.

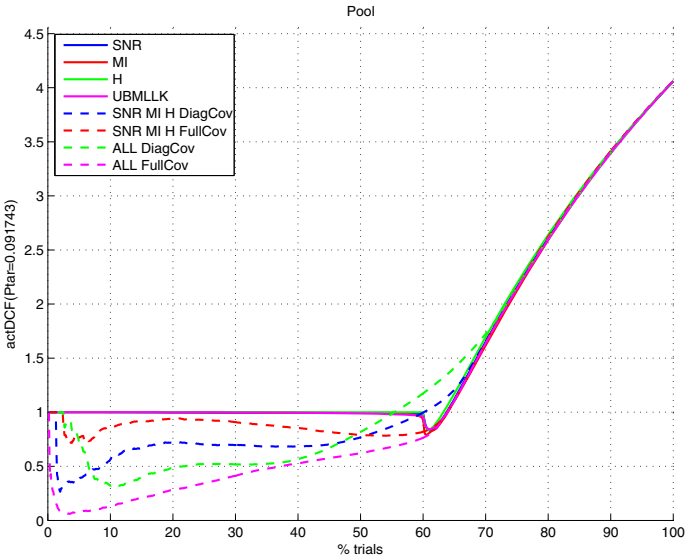


Fig. 4. % trials vs. actDCF in dataset with additive and convolutional noise for BN with score and quality measures

We have worse results than for the dataset with additive noise only. If we use only one quality measure the system only allows to remove 40% of the trials and get actDCF=0.79. That means that our measures are not as good detecting reverberation as detecting noise. However, combining measures improves the results, especially if we add the UBM likelihood to the other three measures. In the best case, we can get actDCF=0.5 pruning 50% of the trials or actDCF=0.13 pruning 90% of the trials.

5 Conclusions

In this paper, we presented a method to detect bad classified trials in a speaker verification system. The method is based on modeling speech quality measures with Bayesian networks. We revisited previous works on Bayesian networks and compare different quality measures and network configurations. We have shown experiments on an artificially degraded database including noise and reverberation. In this experiments, we take a speaker verification system trained and calibrated with clean speech. Using this system on the degraded dataset we get an unacceptable performance. Then, we used our system to remove unreliable trials achieving a dramatic improvement of the actual detection cost function.

Acknowledgment. This work has been supported by the Spanish Government through national projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

References

1. Huggins, M.C., Grieco, J.J.: Confidence metrics for speaker identification. In: 7th ICSLP, Denver, Colorado (2002)
2. Campbell, W.M., Reynolds, D.A., Campbell, J.P., Brady, K.J.: Estimating and evaluating confidence for forensic speaker recognition. In: ICASSP 2005, vol. 1, pp. 717–720 (2005)
3. Solewicz, Y., Koppel, M.: Considering Speech Quality in Speaker Verification Fusion. In: Interspeech 2005 (2005)
4. Richiardi, J., Drygajlo, A., Prodanov, P.: A probabilistic measure of modality reliability in speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005 (2005)
5. Richiardi, J., Drygajlo, A., Prodanov, P.: Confidence and reliability measures in speaker verification. Journal of the Franklin Institute 343(6), 574–595 (2006)
6. Richiardi, J., Drygajlo, A., Prodanov, P.: Speaker Verification with Confidence and Reliability Measures. In: Proc. of ICASSP, vol. 1(6), pp. 641–644 (2006)
7. Nakasone, H., Beck, S.D.: Forensic automatic speaker recognition. In: Odyssey Speaker and Language Recognition Workshop (2001)
8. Bengio, S., Marcel, C., Marcel, S., Mariethoz, J.: Confidence Measures for Multimodal Identity Verification. Information Fusion 3(4), 267–276 (2002)
9. Poh, N., Bengio, S.: Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 474–483. Springer, Heidelberg (2005)
10. Kryszczuk, K., Richiardi, J., Prodanov, P., Drygajlo, A.: Reliability-Based Decision Fusion in Multimodal Biometric Verification Systems. EURASIP Journal on Advances in Signal Processing, 1–10 (2007)
11. Talkin, D.: A robust algorithm for pitch tracking (RAPT). In: Speech Coding and Synthesis, pp. 495–518 (1995)
12. Villalba, J., Lleida, E.: Detecting Replay Attacks from Far-Field Recordings on Speaker Verification Systems. In: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N.C., Fairhurst, M.C. (eds.) BioID 2011. LNCS, vol. 6583, pp. 274–285. Springer, Heidelberg (2011)
13. Harriero, A., Ramos, D., Gonzalez-Rodriguez, J., Fierrez, J.: Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 434–442. Springer, Heidelberg (2009)
14. Bishop, C.: Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC (2006)
15. Hirsch, H.-G., Pearce, D.: The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In: 6th International Conference on Spoken Language Processing, ICSLP 2000, pp. 16–19. Citeseer, Beijing (2000)
16. Hirsch, H.-G.: FaNT - Filtering and Noise Adding Tool (2005)
17. McGovern, S.: A Model for Room Acoustics (2004)

On the use of Total Variability and Probabilistic Linear Discriminant Analysis for Speaker Verification on Short Utterances

Javier González Domínguez, Rubén Zazo, and Joaquin González-Rodríguez

Biometric Recognition Group (ATVS),
Escuela Politecnica Superior, Universidad Autonoma de Madrid
javier.gonzalez@uam.es
http://atvs.ii.uam.es

Abstract. This paper explores the use of state-of-the-art acoustic systems, namely Total Variability and Probabilistic Linear Discriminant Analysis for speaker verification on short utterances. While the recent advances in the field dealing with the session variability problem have proved to greatly outperform speaker verification systems on typical scenarios where a reasonable amount of speech is available, this performance rapidly degrades at the presence of limited data in both enrolment and verification stages. This paper studies the behaviour of TV and PLDA on those scenarios where a scarce amount of speech (∼10s) is available to train and testing a speaker identity. The analysis has been carried out on the well defined and standard 10s-10s task belonging to the NIST Speaker Recognition Evaluation 2010 (NIST SRE10) and it explores the multiple parameters, which define TV and PLDA in order to give some insight about their relevance in this specific scenario.

Keywords: i-vectors, Total variability, PLDA, short utterances.

1 Introduction

The remarkable advances dealing with the session variability problem accomplished during last years, have led to highly reliable speaker verification systems at the presence of a reasonable amount of speech.

In this context, techniques based on Factor Analysis such as Joint Factor Analysis (JFA) [1] [2], Total Variability (TV) [3] or more recently Probabilistic Linear Discriminant Analysis (PLDA) [4] have demonstrated an outstanding behavior even when facing vast and challenging evaluation scenarios such as the NIST Speaker Recognition Evaluation, NIST SRE10 [5].

Unfortunately those excellent results rapidly degrade as long as the available amount of enrolment and verification speech decreases [6] [7]. This fact made critical the design and use of the speaker verification systems in real applications such as access control or forensics while penalizing its application in other everyday applications.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 11–19, 2012.

The purpose of this paper is to evaluate and analyse the state-of-the-art acoustic systems TV and PLDA on those scenarios where just a very limited amount of speech (∼10s) is available for both, enrolment and verification. This analysis is driven through the different design parameters of TV and PLDA, with the aim of discovering which of them have a greater impact dealing with scarce amount of data. This last point is of particular interest, as often, systems are presented adjusted to typical scenarios, overshadowing the actual relevance of the different design parameters in specific tasks.

The rest of this work is organized as follows. A description of the Total Variability and Probabilistic Linear Discriminant Analysis based system is given in Section 2. Section 3 is devoted to present the experimental set-up and obtained results. Finally, main conclusions and future work lines are summarized in Section 4 and Section 5 respectively.

2 Systems Description

2.1 Total Variability

Total Variability [3] represents a step further on the use of Joint Factor Analysis [1] [2] where a single subspace is trained to jointly model both session and speaker variability. This subspace, the so-called *total variability* subspace, T, aims to constraint in a low dimensional space both the session and the speaker variability. Mathematically, this generative latent variable model can be formulated as

$$\overline{\mu}'_s = \overline{\mu}'_{UBM} + Tw \ . \tag{1}$$

where μ_s and μ_{UBM} are the speaker and the Universal Background Model (UBM) model supervector respectively, T is the total variability matrix and w are the the latent factors of the mode, also called *total vectors* or *i-vectors*.

Since T constrains all the variability, speaker and session, and it is shared for all the speakers models/excerpts, the i-vectors, w, can be considered enough to represent the set of differences between one excerpt to each other. Now, the *disentangling* phase between the speaker information and non-desired information can be accomplished at the i-vectors domain. This phase is typically carried out via classical Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) [8]. The use of those techniques is now guaranteed as the dimensional reduction performed allows obtaining a non-singular within-class covariance matrix. Hereafter, we refer the Total Variability system followed by the classical LDA and WCCN as simply Total Variability or TV.

Finally, in order to obtain an score, a straightforward cosine distance between the i-vector coming from the speaker modeling w_1 and a test excerpt i-vector w_2 is computed as

$$S_{w_1,w_2} = \frac{(A^t w_1)W^{-1}(A^t w_2)}{\sqrt{(A^t w_1)W^{-1}(A^t w_1)}\sqrt{(A^t w_2)W^{-1}(A^t w_2)}} \ . \tag{2}$$

where A is the LDA matrix and W is the within class covariance matrix corresponding to WCCN.

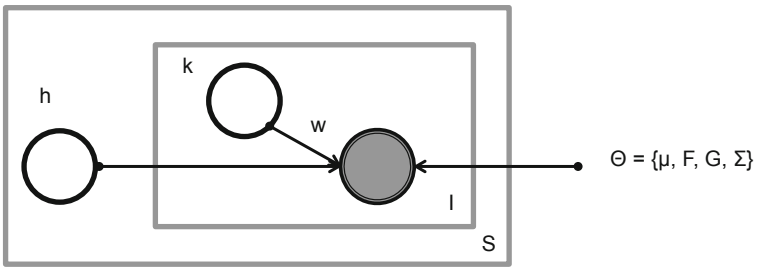


Fig. 1. Probabilistic Linear Discriminant Analysis graphical model representation for S speakers and I utterances. The observed variable w (i-vector), is explained through the identity latent factor h, the session variability hidden variable k and the set of hyperparameters Θ.

2.2 Probabilistic Linear Discriminant Analysis

As stated in the above section, the total variability framework has the main advantage of reducing a given speech utterance to a low-dimensional fixed length representation: the i-vector. From this point, i-vectors can be directly used for classification opening the door to classical methods such as Linear Discriminant Analysis (LDA) to accomplish the disentangling phase between speaker and session variability.

Probabilistic Linear Discriminant analysis (PLDA) is a generative latent variable model that has been recently used to successfully modelling i-vectors [4]. PLDA can be seen as a probabilistic version of classical LDA [9], where a specific i-vector i of a given speaker s is assumed to be decomposed as

$$w_{si} = \mu + Fh_s + Gk_i + \epsilon_i \ . \tag{3}$$

where F and G represents the new speaker and session variability subspaces respectively, h_s and k_i their respective latent variables associated and ϵ_i is a residual noisy term assumed to be normal distributed with zero mean and diagonal covariance matrix Σ. Figure 1 shows the PLDA probabilistic graphical model.

From above equation 3 the analogy between classical stated JFA and PLDA modelling approaches turns out evident. Nonetheless, two mayor important differences, in the context of speaker verification, must be taken into account

- JFA acts over speaker supervectors (high-dimensionality) while PLDA acts over i-vectors (low-dimensionality).
- JFA assumes speaker supervectors as generated by a mixture of multivariate Gaussians, while PLDA assumes i-vectors generated by a single multivariate Gaussian.

Following the PLDA model the similarity measure or score S_{w_1,w_2} between two given i-vectors, w_1 and w_2, can be computed as the ratio of the two alternative hyphotesis: H_0, both w_1 and w_2 belongs to a same identity (same h_s) and H_1, w_1 and w_2 belongs to different identities (different h_s). This ratio can be expressed as

$$S_{w_1,w_2} = \frac{p(w_1, w_2|H_0)}{p(w_1|H_1)p(w_2|H_1)} = \frac{\int p(w_1, w_2|h)p(h)dh}{\int p(w_1|h_1)p(h_1)dh_1 \int p(w_2|h_2)p(h_2)dh_2} \,. \quad (4)$$

Assuming Gaussian priors for the latent variables in the model, it can be seen that integrals involved in above equation 4 turn out tractable and therefore the score, S_{w_1,w_2}, can be easily derived in a closed-form solution. Further details can be found in [9] [10].

3 Experiments

3.1 Experimental Setup

Experiments has been carried out on the telephone male part of the *10s-10s* NIST SRE10 task, where just ∼10s over a telephone channel are provided for both enrolment and verification stages. Specifically, a total number of 10858 trials has been evaluated belonging from 264 and 290 different models and tests segments respectively. The performance was assessed following the NIST SRE10 protocol [11] and results are presented in function of the Equal Error Rate (EER) and the minimum decision cost function (DCF).

Development data for training different Universal Background Models (UBMs) and system hyperparameters belonging to SWBI, SWBII and past NIST SRE evaluations (SRE04, SRE06 and SRE08). Utterances used with this purpose belongs to 1conv/short2 SRE tasks and therefore contains around 2.5m of speech. Specifically a total number of 5638 files from 823 speakers were used to train T, F and Σ [1]; LDA matrix was trained via 5214 files from 611 speakers while 4705 files belonging to 466 speakers were used to estimate the corresponding within class covariance matrix of WCCN method.

Symmetric score normalization (SNorm) [12] was used to finally normalize raw scores generated from the systems. A cohort of a 1000 files from the same development dataset was used for this purpose. Also, the length normalization method proposed in [10] was applied before PLDA modelling.

Regarding the feature extraction configuration, it consists of 38 MFCC coefficients ($19 + \Delta$) extracted by using a sliding Hamming window of 20ms and a 50% of overlapping. MEL filters were scaled between 300 and 3000Hz to focus as much as possible to speech voice.

[1] Given the intrinsic low-dimensionality of the i-vectors and the amount of speech for training the PLDA model available, we opted by grouping the noisy terms in equation 3 into a full-covariance matrix Σ.

Table 1. EER/DCF for Total variability and Probabilistic Linear Discriminant Analysis systems depending on the number of Gaussians used (male SRE10 10s-10s)

	# Gaussians				
System	64	128	256	512	1024
TV-LDA	21.08/0.8202	21.40/0.0785	18.41/0.0710	17.53/0.0698	16.22/0.0687
PLDA	18.79/0.0766	17.27/0.0699	16.00/0.0674	16.22/0.0647	**15.36/0.0614**

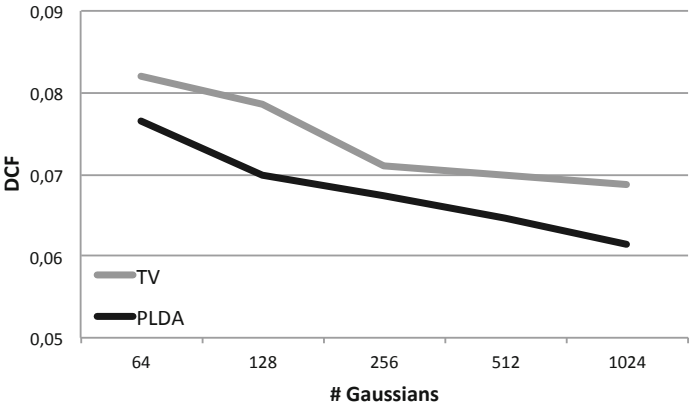


Fig. 2. DCF of TV and PLDA systems as a function of the number of Gaussians

3.2 Results

As the starting point of this analysis, the performance obtained for both TV and PLDA systems was evaluated. Table 1 shows the results of both systems in function of the number of Gaussians used to build the Universal Background Model, and therefore of the i-vector extractor [2]. At a first glance, two observations can be done. First, PLDA method outperforms the LDA followed by WCCN method proposed originally to separate speaker and session variability on i-vectors. This result reinforces, on short utterances, the conclusions extracted in [4] [7], and highlights the mayor ability of the probabilistic framework followed in PLDA to manage uncertainty versus non-probabilistic frameworks. Second, as it can be better observed in Figure 2, increasing the number of Gaussians used in the i-vector extractor turns out in performance gains. The fact of obtaining better performance by doing the system *heavier* (much more free parameters to be trained) beside the nature of the faced problem where just an small amount of speech is available could seem contradictory. However, note that the inherent advantage of using the i-vector framework is that finally, regardless the *size* of the i-vector extractor, classification is done in a low-dimensional space. This

[2] As a reference performance on longer utterances, the same 1024 Gaussians PLDA system achieves a 2.64/0.0149 of EER and DFC on SRE10 task condition 5 (male part only).

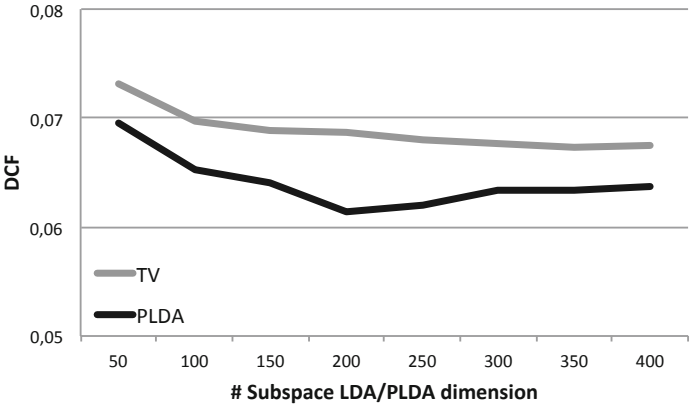


Fig. 3. DCF of TV and PLDA systems as a function of the number of kept dimensions in the LDA supspace and sepaker variability subspace F respectively

last point allows to work with *heavier* and more robust systems at the development time to finally performing classification in a much lower-dimensional space without suffering a performance degradation.

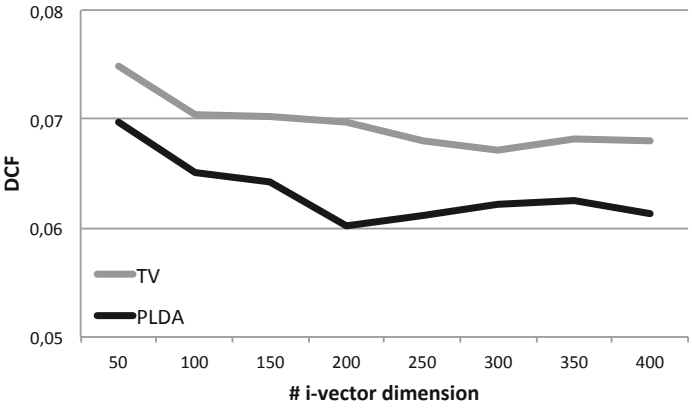


Fig. 4. DCF of TV and PLDA systems as a function of the i-vector dimension

Another aspect explored in this study was the relevance of the LDA and the speaker variability subspace F dimensions in TV and PLDA systems respectively. Figure 3 shows a comparison of both systems by moving those dimensions from 50 to the maximum i-vector dimension used, 400. Here, it can be seen that while the DCF in LDA kept mostly constant from 150 dimensions, PLDA find the minimum DCF at 200 dimension to slightly degrade when using higher dimensions. This result confirms, on short utterances, the studies performed for

Table 2. EER/DCF for Total Variability system in function of the number of Gaussians and LDA dimensions (male SRE10 10s-10s)

	# **Gaussians**				
LDA dim.	64	128	256	512	1024
50	23.37/0.0837	22.23/0.0797	19.17/0.0753	18.96/0.0756	17.61/0.0731
100	22.61/0.0835	21.03/0.0774	18.89/0.0711	17.75/0.0726	16.12/0.0697
150	21.56/0.0825	21.08/0.0769	18.79/0.0698	17.52/0.0698	16.12/0.0688
200	21.08/0.820	21.40/0.0785	18.41/0.0710	17.53/0.0698	16.22/0.0687
250	21.47/0.0834	20.04/0.0773	18.26/0.0712	16.72/0.0720	16.31/0.0680
300	21.56/0.0810	20.05/0.0768	18.55/0.0726	17.27/0.0692	**16.12**/0.0677
350	21.47/0.0819	20.42/0.0768	18.70/0.0723	17.53/0.0697	16.22/**0.0673**
400	21.47/0.0827	20.27/0.0773	17.64/0.0716	17.23/0.0701	16.82/0.0675

Table 3. EER/DCF for Probabilistic Linear Discriminant Analysis system in function of the number of Gaussians and F subspace dimension (male SRE10 10s-10s)

	# **Gaussians**				
F **dim.**	64	128	256	512	1024
50	20,04/0.0824	17,27/0,0813	17,00/0,0751	16,50/0,0747	16,87/0.0695
100	19,17/0,0797	17,27/0,0729	16,89/0,0670	16,12/0,0664	15,36/0.0653
150	19,10/0,0781	16,89/0,0711	16,30/0,0664	16,12/0,0658	16,12/0.0640
200	18,79/0.0766	17.27/0.0699	16.00/0.0674	16.22/0.0647	15.36/**0.0614**
250	19,28/0,0781	16,50/0,0710	16,11/0,0656	16,50/0,0655	15,74/0.0620
300	19,28/0,0774	16,72/0,0715	**14,97**/0,0657	15,84/0,0657	15,74/0.0633
350	19,01/0,0763	16,50/0,0720	15,54/0,0655	16,12/0,0669	15,36/0.0634
400	18,41/0,0776	16,50/0,0717	15,84/0,0667	15,97/0,0665	15,26/0,0638

longer durations, where the optimum size of the PLDA speaker variability subspace use to be lower than the i-vector space [13]. Tables 2 and 3 complete in detail the above described results for both systems, exploring different number of Gaussians and LDA, F subspaces dimensions.

Finally the i-vector dimension used in both systems was also analysed. Figure 4 summarizes the results obtained in terms of DCF by increasing the i-vector dimension from 50 to the standard 400 dimensions. As it can be observed, a minimum at the 300 and 200 dimensions is found for the TV and PLDA systems respectively. These results suggest again that for the short utterances problem i-vector size under 400 dimensions might fit better the problem. Moreover, it encourages the use of PLDA rather than TV followed by LDA and WCCN when using lower dimensional i-vectors; note that a relative improvement of 14% in DCF is achieved by using PLDA instead of TV.

4 Conclusions

A wide analysis on the use of state-of-the art acoustic approaches for speaker verification on short utterances has been carried out in this work. While Total

Variability and Probabilistic Linear Discriminant Analysis methods have demonstrated to achieve outstanding results at the presence of a *reasonable* amount of data, this performance rapidly decrease when just short utterances are available for both enrolment and verification stages. This work has explored the limits of those systems when dealing with short durations. To this aim a leave-one-out analysis of the main configuration parameters of TV and PLDA system has been performed. On one hand, results show that due to the final low-dimensionality dimension of the i-vector, systems designed with complex or *heavy* i-vector extractors (high number of Gaussians, i-vector dimension) are able to obtain gains over lighter ones. On the other hand, the probabilistic framework followed by PLDA has demonstrated to better manage the implicit uncertainty of the task than the classical LDA and WCCN methods.

5 Future Work

Although the use of the i-vector framework achieves acceptable results on the challenging short utterances problem, specially by using PLDA as a modelling technique, some aspects should be explored. On one hand, a deep analysis of the differences among i-vectors extracted from different utterances durations has to be carried out. This study could give some insight, as well as turn out into a better treatment, of the duration utterance effect in speaker verification. Using development short utterances similar to the evaluation conditions, as performed in Joint Factor Analysis [6], could be a possible line in this context.

On the other hand, note that into the Total Variability framework presented, i-vectors are computed as MAP point estimates of the latent factor w. However, it is well known that the use of limited amount of data in order to get point estimates could derive in non reliable i-vectors. In this sense, alternatives as fully Bayesian frameworks as used in [14] for Joint Factor Analysis could be a more appropriate way of facing the short durations problem.

Acknowledgments. This research has been supported by the Ministerio de Ciencia e Innovacion under the proyect TEC2009-14719-C02-01 and Catedra UAM-Telefonica.

References

1. Kenny, P., Boulianne, G., Oullet, P., Dumouchel, P.: Speaker and Session Variability in GMM-Based Speaker Verification. IEEE Trans. on Audio, Speech and Language Processing 15(4), 1448–1460 (2007)
2. Vogt, R., Sridharan, S.: Explicit Modeling of Session Variability for Speaker Verification. Computer Speech & Language 22(1), 17–38 (2008)
3. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing 19(4), 788–798 (2011)

4. Kenny, P.: Bayesian Speaker Verification with Heavy-Tailed Priors. In: Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28-July 1 (2010)
5. Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S.S., Shriberg, E., Stolcke, A.: The SRI NIST 2010 Speaker Recognition Evaluation System. In: ICASSP, pp. 5292–5295 (2011)
6. Vogt, R., Baker, B., Sridharan, S.: Factor analysis subspace estimation for speaker verification with short utterances. In: INTERSPEECH, pp. 853–856 (2008)
7. Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S., Mason, M.W.: I-Vector Based Speaker Recognition on Short Utterances. In: Interspeech 2011, pp. 2341–2344. International Speech Communication Association (ISCA), Firenze Fiera (2011), http://eprints.qut.edu.au/46313/
8. Hatch, A.O., Kajarekar, S.S., Stolcke, A.: Within-class covariance normalization for svm-based speaker recognition. In: INTERSPEECH (2006)
9. Prince, S., Li, P., Fu, Y., Mohammed, U., Elder, J.H.: Probabilistic models for inference about identity. IEEE Trans. Pattern Anal. Mach. Intell. 34(1), 144–157 (2012), http://dblp.uni-trier.de/db/journals/pami/pami34.html#PrinceLFME12
10. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of I-Vector Length Normalization in Speaker Recognition Systems. In: INTERSPEECH, pp. 249–252 (2011)
11. National Institute of Standards and a. o. Technology, The NIST Year 2010 Speaker Recognition Evaluation Plan (2010), http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplanr6.pdf
12. Shum, S., Dehak, N., Dehak, R., Glass, J.R.: Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In: Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic (2010)
13. Matejka, P., Glembek, O., Castaldo, F., Alam, M.J., Plchot, O., Kenny, P., Burget, L., Cernocký, J.: Full-Covariance UBM and Heavy-Tailed PLDA in I-Vector Speaker Verification. In: ICASSP, pp. 4828–4831. IEEE (2011), http://dblp.uni-trier.de/db/conf/icassp/icassp2011.html#MatejkaGCAPKBC11
14. Zhao, X., Dong, Y.: Variational bayesian joint factor analysis models for speaker verification. IEEE Transactions on Audio, Speech & Language Processing 20(3), 1032–1042 (2012)

Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition*

Javier Franco-Pedroso, Fernando Espinoza-Cuadros, and Joaquin Gonzalez-Rodriguez

ATVS – Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
javier.franco@uam.es

Abstract. In this paper, the contributions of different linguistic units to the speaker recognition task are explored by means of temporal trajectories of their MFCC features. Inspired by successful work in forensic speaker identification, we extend the approach based on temporal contours of formant frequencies in linguistic units to design a fully automatic system that puts together both forensic and automatic speaker recognition worlds. The combination of MFCC features and unit-dependent trajectories provides a powerful tool to extract individualizing information. At a fine-grained level, we provide a calibrated likelihood ratio per linguistic unit under analysis (extremely useful in applications such as forensics), and at a coarse-grained level, we combine the individual contributions of the different units to obtain a highly discriminative single system. This approach has been tested with NIST SRE 2006 datasets and protocols, consisting of 9,720 trials from 219 male speakers for the 1side-1side English-only task, and development data being extracted from 367 male speakers from 1,808 conversations from NIST SRE 2004 and 2005 datasets.

Keywords: automatic speaker recognition, forensic speaker identification, temporal contours, linguistic units, cepstral trajectories.

1 Introduction

Automatic speaker recognition has focused in the last decade on two concurrent problems: the compensation of session variability effects, mainly through high-dimensional supervectors and latent variable analysis [2] [7] [8], and the production of an application-independent calibrated likelihood ratio per speaker recognition trial [1], able to elicit useful speaker identity information to the final user with any given application prior. The results are highly efficient text-independent systems in controlled conditions, as NIST SRE evaluations, where lots of data from hundreds of speakers in similar conditions are available. Thus, all the speech available in every trial is used to produce detection performances difficult to imagine a decade ago.

* Supported by MEC grant PR-2010-123, MICINN project TEC09-14179, ForBayes project CCG10-UAM/TIC-5792 and Catedra UAM-Telefonica. Thanks to ICSI (Berkeley, CA) for hosting the preliminary part of this work. Thanks to SRI for providing Decipher labels for SRE datasets.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 20–29, 2012.

However, in the presence of strong mismatch (as e.g. in forensic conditions, where acoustic and noise mismatch, apart from highly different emotional contexts, speaker roles or health/intoxication states can be present between the control and questioned speech), those acoustic/spectral systems could be unusable as all our knowledge about the two speech samples is deposited into a single likelihood ratio, obtained from all the available speech in the utterance, that could be strongly miscalibrated (being then highly misleading) as the system has been developed under severe database mismatch between training and testing data. Moreover, it is difficult (or even impossible) to collect enough data to develop a system robust to every combination of mismatch factors present in actual case data, an important problem in real applications.

A usual procedure in forensic laboratories is that a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable speech between both samples, segments being from seconds long to just some short phonetic events in given articulatory contexts. The number and types of comparable units for analysis is always a case-dependent subject, and therefore flexible strategies for analysis and combination are needed.

The proposed approach gives an answer to this application framework, providing informative calibrated likelihood ratios for every linguistic unit under analysis. Moreover, the combination of the different units yields good discrimination capabilities allowing to obtain speaker detection performance levels similar to equivalent acoustic/spectral systems when enough usable units are available.

The remainder of the paper is organized as follows. In Sections 2 and 3 we present, respectively, our proposed front-end for feature extraction over linguistic units and the system in use. Section 4 describes the databases and the experimental protocol used for testing the system. Section 5 shows results for the different linguistic units individually and for several combination methods, to finally conclude in Section 6 summarizing the main contributions and future extensions of this work.

2 Cepstral Trajectories Extraction from Linguistic Units

Many attempts have been made to incorporate the temporal dynamics of speech into features, from the simplest use of the velocity (delta) and acceleration (delta-delta) derivative coefficients to modulation spectrograms, frequency modulation features or even TDCT (temporal DCT) features (see [9] for a review). However, to the best of our knowledge none of the previous approaches, with the exception of SNERFs [4] and [12] for prosodic information, take advantage of the linguistic knowledge provided by an automatic speech recognizer (ASR) to extract non-uniform-length sequences of spectral vectors to be converted into constant-size feature vectors characterizing the spectro-temporal information in a given linguistic unit. In our proposed front-end, we obtain a constant-size feature vector from non-uniform-length MFCC features sequence within a phone unit.

2.1 ASR Region Conditioning

In this work, both phone and diphone units have been used for defining time intervals in order to extract the temporal contours over the MFCC features. For this purpose, the phonetic transcription labels produced by SRI's Decipher conversational telephone speech recognition system [6] were used first. For this system, trained on English data, the Word Error Rate (WER) of native and nonnative speakers on transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and 36.1% respectively. These labels define both phonetic content and time interval of speech regions containing the phone units to be segmented. For this work, 41 phone units from an English lexicon were used, represented by the Arpabet phonetic transcription code [13]. Diphone units are defined by the combination of any two consecutive phone units, although only a subset of 98 diphones of the possible combinations was used (those presenting higher frequency of occurrence).

2.2 Cepstral Trajectories Parameterization

By means of SRI's Decipher phone labels, trajectories (i.e., the temporal evolution of each MFCC vector dimension) of 19 static MFCC are extracted from phone and diphone units, yielding a MFCC matrix of 19 coefficients x #frames/unit for each linguistic unit. This variable-length segment is duration equalized to a number of frames equivalent to 250 ms. Finally, those trajectories are coded by means of a fifth order discrete cosine transform (DCT), yielding our final 19 x 5 fixed-dimension feature vector for each linguistic unit.

3 System Description

3.1 Unit-Dependent Acoustic Systems

Proposed systems are based on the well known GMM-UBM framework [11], using duration-equalized DCT-coded MFCC trajectories per linguistic unit as feature vectors. The GMM-UBM systems have been the state-of-the-art in the text-independent speaker recognition field for many years until the emergence of JFA [7] and total variability [2] techniques, which have outperformed the former ones through accurately modeling the existing variability in the supervector feature space. For this work, GMM-UBM systems have been chosen for two main reasons: i) as we are using a new type of features, we need first to find the optimal configuration for this GMM-UBM new-framework, which is the basis of supervector-based systems; and ii) because we aim to model speakers in a unit-dependent way, a much smaller amount of data is available for training purposes, so probably not enough data would be available to capture the existing variability in each unit domain (also having into account that we only have ASR labels from the SRE04, SRE05 and SRE06 datasets).

Three different unit-dependent GMM-UBM configurations were tested previously to perform experiments reported in this paper:

1. UBM and speaker models trained on unit-independent data; evaluation trials performed on unit-dependent test data (as we did in our first approach [5]).
2. UBM trained on unit-independent data; speaker models adapted from unit-dependent training data; evaluation trials performed on unit-dependent test data.
3. UBM and speaker models trained on unit-dependent data; evaluation trials performed on unit-dependent test data (fully unit-dependent).

For each configuration, different numbers of mixtures were tested, ranging from 2 up to 1024 mixtures increasing in powers of 2. It was found out that best results were obtained for the fully unit-dependent configuration, using 8 mixtures in the case of phone units and 4 mixtures in the case of diphone units. These configurations are those used to obtain the individual linguistic unit results reported in this paper.

3.2 Fusion Schemes and Linguistic Units Combinations

Both individual unit performance and different unit combinations have been analyzed in this paper. On the one hand, individual linguistic-unit systems allow us to report useful speaker verification LR's for very short speech samples where usual state-of-the-art systems are not directly applicable (as it is the case of forensic applications). On the other hand, when more data is available, individual units can be combined to achieve better discriminative capabilities.

In addiction to obtaining test results for each linguistic unit, these individual systems were combined in both intra- and inter-unit manners, i.e. fusing phone/diphone units between them and fusing phone and diphone units together. Two different fusion techniques were used: sum fusion and logistic regression fusion. The former one was performed after linear logistic regression calibration, while the latter one was performed in a single calibration/fusion step.

Another issue is what should be the selected units to be fused. Two strategies have been used in this work. The first of them is to select the n-best performing units by setting a threshold for the EER of the units to be fused, leaving out those performing worse. However, this procedure do not guaranty that the best fused system will be achieved because some units with lower performance by itself could contribute to the fused system if its LR's are sufficiently low correlated with those produced by the other units to be fused. On the other hand, testing all of the possible combinations would be a very complex task, so we used a unit selection algorithm (similar to that used in [3]) based on the following steps:

1. Take the best performing unit in terms of EER as the initial units set.
2. Take the next best performing unit and fuse with the previous set. If the fusion improves the performance of the previous set, this unit is added to the units set, otherwise rejected.
3. The previous step is repeated for all the units in increasing EER order.

This procedure allows us to find complementarities between units that otherwise would not have been revealed, but avoiding the complex task of testing each possible combination.

4 Datasets and Experimental Setup

NIST SRE datasets and protocols have been used to develop and test our proposed system, in particular those of years 2004, 2005 and 2006. As region conditioning for linguistic units definition and extraction rely on SRI's Decipher ASR system (trained on English data), English-only subsets of the NIST SRE datasets have been used. SRE 2004 and 2005 datasets were used as the background dataset for UBM training, consisting of 367 male speakers from 1,808 conversations (only male speakers were used for this work). English-only male 1side-1side task from SRE 2006 was used for testing purposes. This dataset and evaluation protocol comprises both native and nonnative speakers across 9,720 same-sex different-telephone-number trials from 298 male speakers. SRE 2005 evaluation set was also used to obtain scores in order to train the calibration rule (linear logistic regression).

Performance evaluation metrics used are the Equal Error Rate (EER) and the Detection Cost Function (DCF) as defined in the NIST SRE 2006 evaluation plan [10]. Cllr and minCllr [1] (and its difference, calibration loss) are also used to evaluate the goodness of the different detectors after the calibration process.

5 Results

5.1 Reference System Performance

As we are using the GMM-UBM framework to model unit-dependent systems, our baseline reference system is also a GMM-UBM system based on MFCC features. A classical configuration with 1024 mixtures and diagonal covariance matrices was used, and MFCC features include 19 static coefficients plus first order derivatives, cepstral mean normalization, RASTA filtering and feature warping. The performance of this system in the English-only male 1side-1side task from SRE 2006 is EER=10.26% and minDCF=0.0457. This system does not include any type of score normalization.

5.2 Phone Units: Individual and Combined Systems Performances

Table 1 shows individual performance of phone units for the NIST SRE 2006 English-only male 1side-1side task. It can be seen that, although most of the phones have high EER and minDCF values, almost all of them are well calibrated (low difference between Cllr and minCllr). This allows us to obtain informative calibrated likelihood ratios from very short speech samples (as low as some phone units), as we can see in the tippet plot in Figure 1 for the best performing phone unit ('N'). Moreover, there are lots of units that can be combined, and despite their lower individual performance (around 60% worse than the reference system for the best performing phone), combined system can outperform reference system by means of sum or logistic regression fusion, as it can be seen in Figure 2. This is due to the highly complementarity of acoustic systems coming from different linguistic content.

Table 1. EER (%), minDCF, Cllr and minCllr for phone units in the NIST SRE 2006 English-only male 1side-1side task

Phone unit	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AA	32.20	0.0983	0.8633	0.8452
AE	18.98	0.0813	0.6087	0.5832
AH	29.39	0.0969	0.8235	0.7967
AO	34.36	0.0992	0.9065	0.8838
AW	36.99	0.0991	0.9241	0.9111
AX	27.08	0.0947	0.7882	0.7512
AY	21.68	0.0869	0.6822	0.6428
B	34.50	0.0986	0.8922	0.8778
CH	42.59	0.1000	0.9686	0.9538
D	32.07	0.0965	0.8661	0.8500
DH	28.43	0.0934	0.8403	0.7857
DX	40.44	0.0998	0.9670	0.9484
EH	31.69	0.0975	0.8574	0.8283
ER	35.18	0.0987	0.9107	0.8901
EY	26.40	0.0925	0.7713	0.7515
F	39.63	0.0993	0.9561	0.9397
G	35.71	0.1000	0.9291	0.9040
HH	39.80	0.0992	0.9527	0.9414
IH	26.95	0.0948	0.7964	0.7495
IY	23.32	0.0923	0.7453	0.7002
JH	39.69	0.0997	0.9487	0.9339
K	27.76	0.0961	0.8219	0.7832
L	26.51	0.0935	0.7789	0.7451
M	22.28	0.0857	0.6824	0.6583
N	15.92	0.0713	0.5520	0.5082
NG	29.37	0.0934	0.9977	0.7958
OW	24.65	0.0987	0.7917	0.7396
P	39.50	0.0988	0.9466	0.9335
PUH	24.18	0.0908	0.7359	0.7149
PUM	34.15	0.0953	0.8644	0.8419
R	24.65	0.0887	0.7295	0.7116
S	30.04	0.0973	0.8451	0.8059
SH	39.36	0.0996	1.0546	0.9294
T	27.89	0.0921	0.8256	0.7647
TH	38.37	0.1000	1.1207	0.9298
UH	41.53	0.1000	0.9717	0.9593
UW	24.79	0.0898	0.7391	0.7198
V	35.86	0.0990	0.9093	0.8932
W	35.82	0.0993	0.9167	0.8966
Y	24.00	0.0906	0.7313	0.7062
Z	32.07	0.0968	0.8487	0.8312

It should be noted that results equivalent to that of the reference system can be achieved by combining only 4 phone units (‘AE’, ‘AY’, ‘M’, ‘N’). Also, it can be seen that the unit selection algorithm used can achieve better fusion results than simply setting a threshold for the EER of the units to be fused, both for sum and logistic regression fusions. Furthermore, it is worth noting that some of the phone units selected to be fused have very low performance (‘CH’ in the sum fusion, ‘AO’ in both sum and logistic regression fusions).

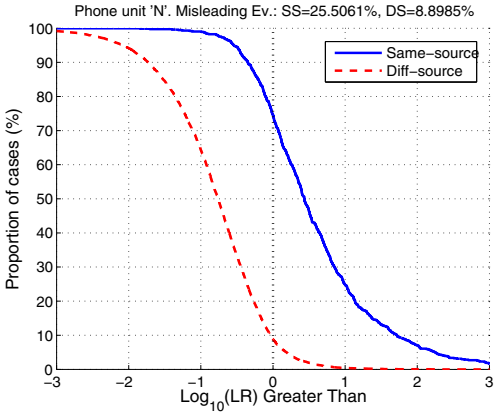


Fig. 1. Tippet plot for the best performing phone unit (‘N’) in the NIST SRE 2006 English-only male 1side-1side task

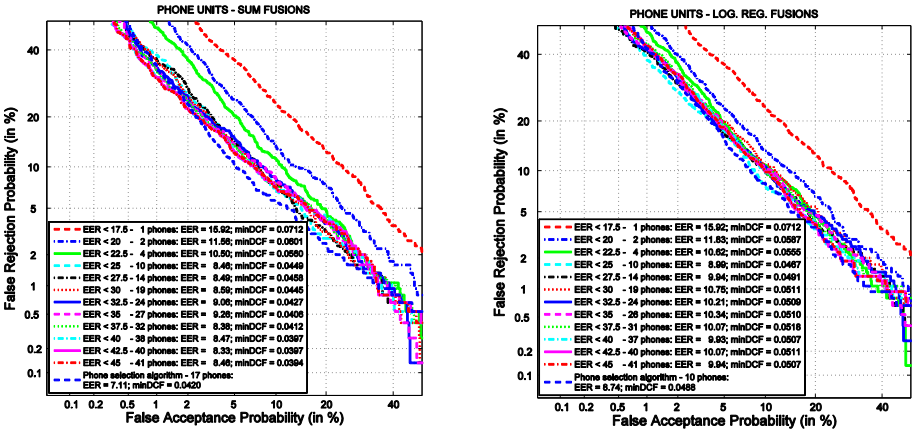


Fig. 2. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different phone selection schemes

5.3 Diphone Units: Individual and Combined Systems Performances

Table 2 shows individual performance for the ten best performing diphone units for the NIST SRE 2006 English-only male 1side-1side task. As it can be seen, diphone units have much lower performance than phone units. This may be due to the fact that, while diphones cover a longer time span that can present more complex trajectories, we are still using a 5 order DCT to code these trajectories. However, as it can be seen in Figures 3, diphone fusions can achieve as good performance as the phones unit fusions, although more units are needed to be fused.

Table 2. EER (%), minDCF, Cllr and minCllr for the 10 best performing diphone units in the NIST SRE 2006 English-only male 1side-1side task

Diphone unit	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AEN	30.72	0.0993	0.8479	0.823
AET	31.89	0.0969	0.872	0.8526
AXN	23.84	0.0899	0.7583	0.7097
AYK	32.45	0.0970	0.8494	0.8356
LAY	29.11	0.0972	0.8156	0.7955
ND	24.92	0.0876	0.7563	0.7037
NOW	30.86	0.0995	0.8455	0.8185
UWN	32.20	0.0953	0.8417	0.8188
YAE	29.78	0.0976	0.8383	0.8094
YUW	27.18	0.0960	0.8223	0.7812

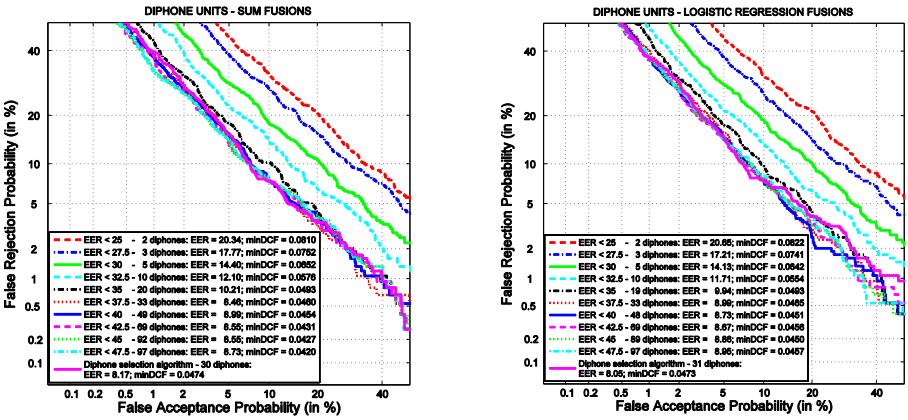


Fig. 3. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different diphone selection schemes

5.4 Inter-unit Combined System Performance

In the previous paragraphs we have seen how well combine different units from each type (i.e., different phones between them and different diphones between them), but it is also interesting to see how can be combined units from different types between

them. For this purpose, same fusion techniques and combination schemes have been used putting together both phones and diphones, yielding results show in Figure 4.

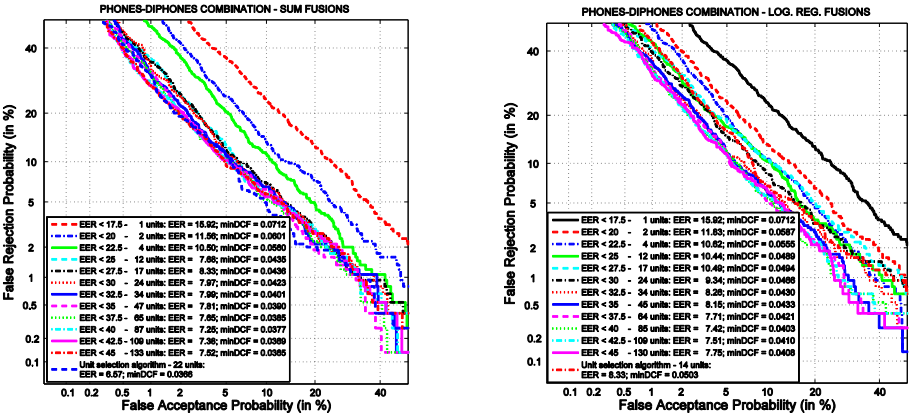


Fig. 4. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different phone-diphone selection schemes.

It can be seen that better results can be achieve by combining phones and diphones units than working in a intra-unit manner, taking advantage of different linguistic levels. This way, it is possible to achieve improvements around 35% in terms of EER over the reference system, as it can be seen in Table 3.

6 Summary and Conclusions

In this paper we have presented an analysis of the contributions of individual linguistic units to automatic speaker recognition by means of their cepstral trajectories, showing that some of them can be used to obtain informative likelihood ratios very useful in forensic applications, with the advantage of being a completely automatic system and using parameters similar to those used by linguists or phoneticians. This way it is possible to deal with uncontrolled scenarios where only some short segments are available to be compared, making it possible to infer a conclusion about the speaker identity in the speech sample. This procedure cannot be done by the usual automatic speaker recognition systems because they use all available speech data as a

Table 3. Performance comparison between the reference system and unit-based fused systems in the NIST SRE 2006 English-only male 1side-1side task

System	# fused units	EER (%)	minDCF
Reference	-	10.26	0.0457
Phones – best fused system (sum)	17	7.11	0.0420
Diphones – best fused system (log. reg.)	31	8.05	0.0473
Phones+diphones – best fused system (sum)	22	6.57	0.0366

whole, and usually they are tuned to work with fixed-length training and testing segments. Furthermore, when more testing data is available, individual units can be combined to improve the discrimination capabilities of the resulting system, having shown that these combinations, both at intra- and inter-unit levels, can outperform the results obtained with the same system framework based on MFCC features.

References

1. Brummer, N., et al.: Application-independent evaluation of speaker detection. Comp. Speech Lang. (20), 230–275 (2006)
2. Dehak, N., et al.: Front-End Factor Analysis for Speaker Verification. IEEE Trans. on Audio, Speech and Lang. Proc. 19(4), 788–798 (2011)
3. de Castro, A., Ramos, D., Gonzalez-Rodriguez, J.: Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking. In: Proceedings of Interspeech 2009, pp. 2343–2346 (September 2009)
4. Ferrer, L.: Statistical modeling of heterogeneous features for speech processing tasks. Ph.D. dissertation, Stanford Univ. (2009), http://www.speech.sri.com/people/lferrer/thesis.html
5. Franco-Pedroso, J., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., Ramos, D.: Fine-grained automatic speaker recognition using cepstral trajectories in pone units. In: Proceedings of IAFPA 2012, Santander, Spain (2012)
6. Kajarekar, S., et al.: The SRI NIST 2008 Speaker Recognition Evaluation System. In: Proc. IEEE ICASSP 2009, Taipei, pp. 4205–4209 (2009)
7. Kenny, P., et al.: A Study of Inter-speaker Variability in Speaker Verification. IEEE Trans. on Audio, Speech and Lang. Proc. 16(5), 980–988 (2008)
8. Kenny, P.: Bayesian speaker verification with heavy tailed priors. Keynote Presentation at Odyssey 2010, Brno (2010)
9. Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: from features to supervectors. Speech Communication 52, 12–40 (2010)
10. NIST SRE 2006 Evaluation Plan (2006), http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf
11. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)
12. Shriberg, E.: Modeling prosodic feature sequences for speaker recognition. Speech Communication 46(3-4), 455–472 (2005)
13. Wikipedia contributors. Arpabet. Wikipedia, The Free Encyclopedia (July 19, 2012), http://en.wikipedia.org/wiki/Arpabet

Improving the Quality of Standard GMM-Based Voice Conversion Systems by Considering Physically Motivated Linear Transformations

Tudor-Cătălin Zorilă[1,2], Daniel Erro[2], and Inma Hernaez[2]

[1] POLITEHNICA University of Bucharest (UPB), Bucharest, Romania
ztudorc@gmail.com
[2] AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
{derro,inma}@aholab.ehu.es

Abstract. This paper presents a new method to train traditional voice conversion functions based on Gaussian mixture models, linear transforms and cepstral parameterization. Instead of using statistical criteria, this method calculates a set of linear transforms that represent physically meaningful spectral modifications such as frequency warping and amplitude scaling. Our experiments indicate that the proposed training method leads to significant improvements in the average quality of the converted speech with respect to traditional statistical methods. This is achieved without modifying the input/output parameters or the shape of the conversion function.

Keywords: voice conversion, Gaussian mixture models, dynamic frequency warping, amplitude scaling, linear transformation.

1 Introduction

Voice conversion (VC) has acquired a lot of attention from speech technologies researchers during the last two decades [1–13], being a subject still far from conclusion. VC can be understood as the process by which the voice characteristics of a speaker (source speaker) are replaced by those of another speaker (target speaker) so that the modified speech signal will sound as if it had been produced by the target speaker. VC can be applied to a full range of applications. It can provide an almost costless source of voice variability in text-to-speech (TTS) synthesis, where re-recording new voices is an expensive process and not always possible. This technique can also be applied for voice modifications in movie, music and computer game industries or can be used to repair pathological voices.

VC systems operate in two different modes: training and conversion. During the training phase, given speech recordings from the two involved speakers, the VC systems learn a function to transform the source speaker's acoustic space into that of the target speaker. During the conversion phase, this function is applied to transform new input utterances from the source speaker. Various types of VC techniques have been studied in the literature: vector quantization and mapping codebooks [1], more

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 30–39, 2012.

sophisticated solutions based on fuzzy vector quantization [2], frequency warping transformations [3, 4], artificial neural networks [5], hidden Markov models [6], classification and regression trees [6], etc. However, another technique, namely statistical parametric VC based on Gaussian mixture models (GMM), has prevailed over them.

GMM-based VC systems [7, 8] use statistical principles to partition the acoustic space into a finite number of overlapping classes. Then, a linear transformation is learnt for each class. The function applied during the conversion stage is a statistically weighted combination of these linear transforms. The main problem associated with this well known technique is referred to as oversmoothing. This phenomenon is a consequence of the limited capability of this specific statistical conversion function to capture the correspondence between source and target features in all its variability. As a result of it, the converted speech will sound excesively smoothed and not very natural in terms of subjective quality. Existing methods to alleviate oversmoothing either oversimplify the conversion function [9] or apply sophisticated transformations involving utterance-level features such as the global variance of the converted parameters [10], thus losing the capability of performing frame-by-frame VC in real-time applications.

This paper follows the line of previous works in which frequency warping (FW) based transformations were combined with traditional GMM-based systems [11–13]. FW functions map the frequency axis of the source speaker's spectrum into that of the target speaker. Since they do not remove any detail of the source spectrum, they yield high-quality converted speech judged as quite natural by listeners. However, the conversion accuracy they achieve is moderate because the FW procedure does not modify the relative amplitude of meaningful parts of the spectrum. For this reason, FW was combined with traditional GMM-based systems in several ways [11–13]. In all of these systems, the shape of the VC function had to be modified and more sophisticated signal models and vocoders had to be used to make this combination possible.

In this paper we propose an alternative way of training the set of linear transformations to be applied by a traditional GMM-based VC system. In this new training method, the matrices and vectors of the transformation are calculated according to physical criteria: the matrices are forced to correspond to a FW operation, and the vectors play the role of corrective filters. During conversion, the system operates in the same way as a traditional one and uses the same input/output parameters, i.e. Mel-cepstral coefficients. Despite this, its performance is significantly enhanced in terms of subjective quality, because the degree of oversmoothing is effectively reduced and the converted voice sounds more natural.

The remainder of the paper is structured as follows. Section 2 contains a brief description of the fundamentals of GMM-based voice conversion, including a mathematical interpretation of the oversmoothing effect. In section 3 we show the details of one of the most popular FW training methods. In section 4 we explain the novel training method in which FW-based transformations are integrated into the traditional statistical framework. The effectiveness of this method is experimentally shown in section 5. Finally, the conclusions of this work are summarized in section 6.

2 Traditional GMM-Based VC

The conversion function applied by traditional GMM-based VC systems [7, 8] is a probabilistic combination of m linear transforms:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i^{(\theta)}(\mathbf{x}) \left[\mathbf{v}_i + \mathbf{\Gamma}_i \mathbf{\Sigma}_i^{(xx)^{-1}} \left(\mathbf{x} - \boldsymbol{\mu}_i^{(x)} \right) \right] \tag{1}$$

where m is the number of Gaussian mixtures of the model θ, $\boldsymbol{\mu}_i^{(x)}$ and $\mathbf{\Sigma}_i^{(xx)}$ are the mean vector and covariance matrix that characterize the i^{th} Gaussian mixture of θ, and $p_i^{(\theta)}(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to that specific mixture. Alternatively, the VC function can be expressed as

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i^{(\theta)}(\mathbf{x}) \left[\mathbf{A}_i \mathbf{x} + \mathbf{b}_i \right] \tag{2}$$

where

$$\mathbf{A}_i = \mathbf{\Gamma}_i \mathbf{\Sigma}_i^{(xx)^{-1}}, \ \mathbf{b}_i = \mathbf{v}_i - \mathbf{A}_i \boldsymbol{\mu}_i^{(x)} \tag{3}$$

Given a training set of paired vectors contained acoustic parameters (Mel-cepstral coefficients in this case), the unknown vectors and matrices of this VC function, $\{\mathbf{v}_i\}$ and $\{\mathbf{\Gamma}_i\}$, can be obtained either by least squares based minimization of the conversion error [7] or by joint density modeling of the concatenated pairs of vectors [8]. In both cases, the resulting converted speech will be perceived by listeners as oversmoothed. Previous investigations on the reasons why oversmoothing appears [9] showed that most of the elements of the matrices $\{\mathbf{A}_i\}$ yielded by traditional training methods were very close to zero due to the limited capability of the GMMs to model the source-target correspondence. In these conditions, the transformation given by expression (1) can be approximated by a simple weighted combination of m vectors $\{\mathbf{v}_i\}$, which explains the observed oversmoothing phenomenon.

In the next section we will show that alternative training methods based on physical principles can provide the traditional linear VC function with matrices and vectors that make it less prone to oversmoothing.

3 Fundamentals of Dynamic Frequency Warping

Dynamic FW (DFW) [3] is a procedure that calculates the FW function that should be applied to a set of (N+1)-point log-amplitude semispectra, $\{X_t\}$, to make them maximally close to their paired counterparts, $\{Y_t\}$. It is based on a cost function $D(i, j)$ which indicates the accumulated log-spectral distortion that would be obtained if the i^{th} bin of the source spectra were mapped into the j^{th} bin of the target spectra following the "best" path from (0, 0) to (i, j). $D(i, j)$ can be expressed mathematically as follows:

$$D(i,j)=\min\begin{Bmatrix}D(i-1,j)+d(i,j)\\ D(i-1,j-1)+w\cdot d(i,j)\\ D(i,j-1)+d(i,j)\end{Bmatrix},\quad i,j=0\ldots N \tag{4}$$

where w, $1 \le w < 2$, is an adjustable weighting coefficient that controls the relative penalty of vertical and horizontal paths ($w \approx 2$ means no penalty for them, while $w \approx 1$ means strong penalty), and $d(i, j)$ is a local distortion measure involving exclusively the i^{th} source bin and the j^{th} target bin. In our implementation, $d(i, j)$ is calculated simultaneously from all the available training vectors to globally optimize the warping procedure:

$$d(i,j)=\sum_{t=1}^{T}\left(X_t[i]-Y_t[j]\right)^2 \tag{5}$$

The frequency warping path P is given by a sequence of points,

$$P=\left\{(0,\ 0),(i_1,j_1),(i_2,j_2),\ldots,(i_K,j_K)\right\}, \tag{6}$$

such that the presence of (i, j) in P indicates that the i^{th} bin of the source spectrum should be mapped into the j^{th} bin of the target spectrum for an optimal warping in terms of log-spectral distortion. In this work, i_K and j_K are forced to be equal to N, so the remaining points of P are backtracked from (N, N) following the minimal-distortion path in inverse order. Note that this path is determined by the recursion in expression (4).

4 Physically Motivated Linear Transforms

DFW is not trainable directly in the parametric domain. Therefore, the first step in the training procedure is translating p^{th}-order cepstral vectors into $(N+1)$-point discrete log-amplitude semispectra. By definition, this can by done by multiplying the cepstral vectors by the following matrix:

$$\mathbf{S}=\begin{bmatrix}1 & 2\cos(\omega_0) & 2\cos(2\omega_0) & \cdots & 2\cos(p\omega_0)\\ 1 & 2\cos(\omega_1) & 2\cos(2\omega_1) & \cdots & 2\cos(p\omega_1)\\ 1 & 2\cos(\omega_2) & 2\cos(2\omega_2) & \cdots & 2\cos(p\omega_2)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & 2\cos(\omega_N) & 2\cos(2\omega_N) & \cdots & 2\cos(p\omega_N)\end{bmatrix},\quad \omega_k=g\left(k\frac{\pi}{N}\right) \tag{7}$$

where $g(\ )$ is an optional perceptual frequency scale. Note that $g(\ )$ is directly related to the frequency scale assumed during the cepstral analysis.

Similarly, the p^{th}-order cepstral representation of a discrete log-amplitude spectrum can be recovered through the technique known as regularized discrete cepstrum [14],

which is equivalent to multiplying the (N+1)-point discrete log-amplitude semispectrum in vector form by

$$\mathbf{C} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{R}\right)^{-1}\mathbf{S}^{\mathrm{T}} \tag{8}$$

where $\mathbf{S}$ is given by (7), $\mathbf{R}$ is a regularization matrix that imposes smoothing constraints to the cepstral envelope,

$$R = 8\pi^2 \cdot diag\left\{0, 1^2, 2^2, \ldots p^2\right\}, \tag{9}$$

and λ is an empirical constant typically equal to $2{\cdot}10^{-4}$ [14]. In practice, since the 0^{th} cepstral coefficient (the one carrying the energy) is not considered in voice transformation tasks, we use modified versions of these matrices, $\hat{\mathbf{S}}$ and $\hat{\mathbf{C}}$, where $\hat{\mathbf{S}}$ results from removing the first column of $\mathbf{S}$ and $\hat{\mathbf{C}}$ results from removing the first row of $\mathbf{C}$.

After the training vectors are converted into spectra using matrix $\hat{\mathbf{S}}$, an optimal warping path P is obtained via the DTW training procedure in section 3. Then, we can define the following matrix containing the source-target correspondence:

$$\mathbf{M}[i, j] = \begin{cases} 1, & (i, j) \in P \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

The multiplication of a source semispectrum by $\mathbf{M}^{\mathrm{T}}$ would yield a warped version of the same semispectrum if there were no one-to-many mappings between target and source bins. However, one-to-many mappings are unavoidable according to the structure of P, which is conditioned by the recursion in (4). Therefore, we define the following warping matrix $\mathbf{W}$ in which multiple source bins paired with the same target bin are just averaged:

$$\mathbf{W}[i, j] = \frac{\mathbf{M}[j, i]}{\sum_{k=1}^{N}\mathbf{M}[k, i]} \tag{11}$$

Once $\mathbf{W}$ has been determined, the matrix that converts a p^{th}-order cepstral vector into another cepstral vector representing the warped version of the original spectrum can be easily obtained as

$$\tilde{\mathbf{A}} = \hat{\mathbf{C}} \cdot \mathbf{W} \cdot \hat{\mathbf{S}} \tag{12}$$

Since the frequency response of a corrective filter can be seen as an additive term in the cepstral domain, the cepstral correction vector that is necessary to compensate for the differences between frequency-warped source vectors and target vectors is

$$\tilde{\mathbf{b}} = \mathbf{y}_{avg} - \tilde{\mathbf{A}}\mathbf{x}_{avg} \tag{13}$$

where $\mathbf{x}_{avg}$ and $\mathbf{y}_{avg}$ are computed simply by averaging the source and target cepstral vectors over the training dataset. As a result of this training procedure, we get the following physically motivated linear transformation:

$$\mathbf{y} = \tilde{\mathbf{A}}\mathbf{x} + \tilde{\mathbf{b}} \tag{14}$$

We suggest applying this linear transformation in a traditional statistical framework:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i^{(\theta)}(\mathbf{x})\left[\tilde{\mathbf{A}}_i\mathbf{x} + \tilde{\mathbf{b}}_i\right] \tag{15}$$

The matrices and vectors of the transformation can be trained independently for each class of the GMM using exclusively the vectors in that class. For a hard classification, we can assume that $\mathbf{x}$ belongs to the i^{th} class of model θ when $p_i^{(\theta)}(\mathbf{x}) > p_j^{(\theta)}(\mathbf{x})$ for $j = 1\ldots m, j \neq i$. Although such a hard partition of the acoustic space during training is inconsistent with the soft partition used during conversion (15), this does not have any remarkable perceptual consequence according to our listening tests.

5 Experiments and Discussion

The speech data used in the evaluation experiments were taken from the CMU ARCTIC database [15]. Four speakers were selected from this database: two female speakers, *slt* and *clb*, and two male speakers, *bdl* and *rms*. From now on, for the sake of simplicity, they will be referred to as *f*1, *f*2, *m*1 and *m*2, respectively. 50 parallel training sentences per speaker were randomly selected for training and a different set of 50 sentences was separated for testing purposes. The remaining sentences of the database were simply discarded. The sampling frequency of the signals is 16 kHz. We used the vocoder described in [16] to translate the speech signals into Mel-cepstral coefficients and to reconstruct the waveforms from the converted vectors. The order of the cepstral analysis was 24 (plus the 0^{th} coefficient containing the energy, which does not take part in the conversion). The frame shift was set to 8ms. During conversion, the mean and variance of the source speaker's $\log f_0$ distribution were replaced by those of the target speaker by means of a linear transformation. In order to find the correspondence between the source and target cepstral vectors extracted from the parallel training utterances, we calculated a piecewise linear time warping function from the phoneme boundaries given by the available segmentation. The GMMs used in all the experiments had 32 mixtures with full-covariance matrices. Such a number of mixtures was chosen according to phonetic criteria, objective scores measured on separate validation sets, and informal listening tests. During DTW-related computations, N was set to 512.

In the first experiment, different configurations of the proposed method are compared in terms of average Mel-cepstral distortion (MCD) between converted and target vectors. Three specific aspects of the method are studied:

- The influence of the perceptual frequency scale applied when resampling the cepstral envelopes in expression (7). We consider Mel and linear frequency scale. These two configurations will be labeled as "mel" and "lin" respectively.
- The effect of removing the glottal source spectrum from $\{X_t\}$ and $\{Y_t\}$ before training the DFW paths, as suggested in earlier works [3]. In our implementation, we

assume that the glottal spectrum is mainly related with the 1st cepstral coefficient. According to this, we remove the glottal spectrum by settin $c_1 = 0$. This configuration with will be labelled as "c1=0".

— The effect of considering just one representative vector for each class in expression (5), i.e. the average vector, instead of considering all the vectors simultaneously during DFW training. We use labels "avg" and "all" for these configurations.

The MCD scores in Fig. 1, which have been obtained by calculating global scores over all possible combinations of voices, reveal that: (i) considering all the training vectors instead of their average is significantly advantageous; (ii) removing the glottal spectrum is mandatory when only average representative vectors are considered, but it is not crucial when all the vectors are considered during DFW training; (iii) no significant differences can be seen between Mel- and linear-frequency resampling of cepstral envelopes. These observations hold for individual conversion directions.

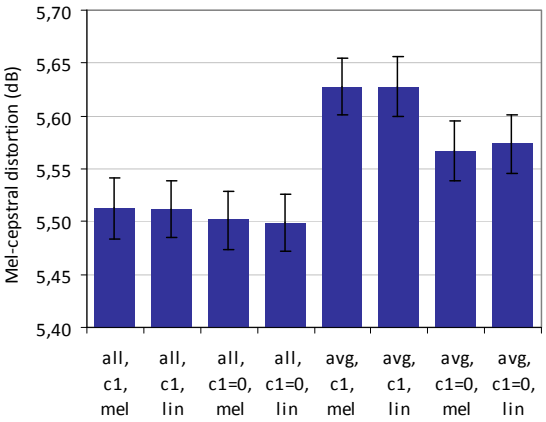


Fig. 1. Average MCD scores and 95% confidence intervals for different configurations of the system and for all combinations of voices

Table 1. Objective comparison between traditional and proposed GMM-based VC systems

	No conversion	Traditional GMM	Proposed GMM
MCD (dB)	7.05 ± 0.03	4.78 ± 0.03	5.50 ± 0.03

Table 1 indicates that a traditional GMM-based system based on joint-density modeling [8] gives significantly better MCD scores than the proposed system regardless of its configuration. Similar observations were made in previous related works [12], where it was also shown that objective distortion measures do not necessary correlate well with subjective measures when the nature of the methods under comparison is heterogeneous. Therefore, we conducted a perceptual mean opinion score (MOS) test to compare the best configuration of the proposed system in terms of

MCD (the one labeled as "all, c1=0, lin") with a traditional GMM-based VC system. In this test, 18 volunteer evaluators listened to reference utterances from the target speakers (previously parameterized and reconstructed with the same vocoder as the converted speech) followed by converted utterances. The listeners were asked to rate the similarity between converted and target voices and the quality of the converted voices in a 5-point scale. As usual, 5 points was the best score and 1 point was the worst. Comparisons were made for 4 different conversion directions: *m*1-*f*1, *f*1-*f*2, *f*2-*m*2, and *m*2-*m*1. The results of the test are shown in Fig. 2. On average, the proposed method significantly outperforms the traditional system in terms of quality while achieving comparable scores in terms of similarity. A more detailed case-by-case analysis reveals that the proposed system is relatively less successful in cross-gender cases. In fact, there is one conversion direction, namely "*f*2-*m*2", in which no quality improvements are achieved. Further analyses indicated that this can be due to the particularities of this specific pair of voices and to some possibly inaccurate decisions regarding the manually adjustable weights and permitted paths in expression (4). These issues will be tackled in future works.

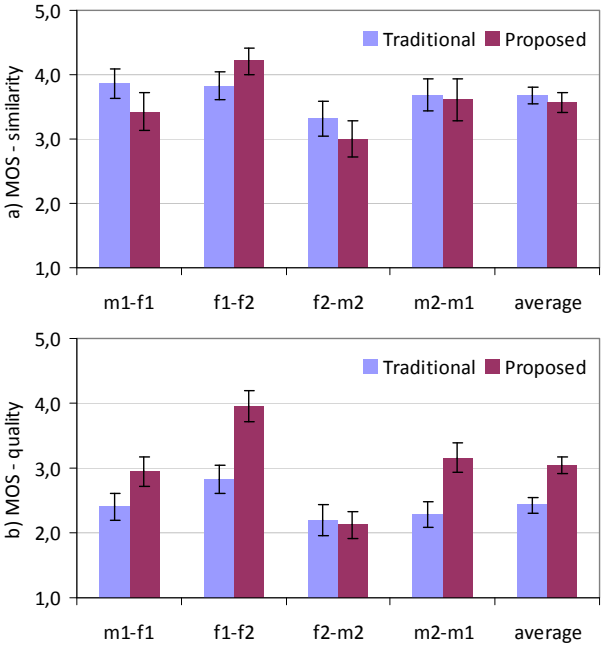


Fig. 2. Mean opinion scores and 95% confidence intervals: a) similarity; b) quality

6 Conclusions

This paper has shown that the performance of traditional voice conversion systems based on Gaussian mixture models and linear transforms can be improved by

imposing some physically meaningful constraints to the matrices and vectors of the transformation. The resulting system is applicable in the same circumstances as the traditional one. Subjective listening tests indicate that on average the proposed method produces evident and statistically significant improvements in quality. Future works will aim at finding the optimal configuration of the system for it to be more robust against the particularities of some specific voice pairs.

Acknowledgements. This work has been partially supported by the Romanian Ministry of Labour, Family and Social Protection (financial agreement POSDRU/88/1.5/S/61178), the Spanish Ministry of Science and Innovation (Buceador, TEC2009-14094-C04-02) and the Basque Government (Berbatek, IE09-262; ZURE_TTS, SPE11UN081).

References

1. Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 655–658 (1988)
2. Arslan, L.M.: Speaker transformation algorithm using segmental codebooks (STASC). Speech Commun. 28, 211–226 (1999)
3. Valbret, H., Moulines, E., Tubach, J.P.: Voice transformation using PSOLA technique. Speech Commun. 1, 145–148 (1992)
4. Sündermann, D., Ney, H.: VTLN-based voice conversion. In: Proc. IEEE Symp. Signal Process. Inf. Technol., pp. 556–559 (2003)
5. Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B.: Transformation of formants for voice conversion using artificial neural networks. Speech Commun. 16(2), 207–216 (1995)
6. Duxans, H., Bonafonte, A., Kain, A., van Santen, J.: Including dynamic and phonetic information in voice conversion systems. In: Proc. Int. Conf. Spoken Lang. Process., pp. 1193–1196 (2004)
7. Stylianou, Y., Cappé, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech and Audio Process. 6, 131–142 (1998)
8. Kain, A.: High resolution voice transformation. Ph.D. thesis, Oregon Health & Science University (2001)
9. Chen, Y., Chu, M., Chang, E., Liu, J.: Voice conversion with smoothed GMM and MAP adaptation. In: Proc. Eurospeech, pp. 2413–2416 (2003)
10. Toda, T., Black, A.W., Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio, Speech, Lang. Process. 15(8), 2222–2235 (2007)
11. Toda, T., Saruwatari, H., Shikano, K.: Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 841–844 (2001)
12. Erro, D., Moreno, A., Bonafonte, A.: Voice conversion based on weighted frequency warping. IEEE Trans. Audio, Speech, Lang. Process. 18(5), 922–931 (2010)
13. Tamura, M., Morita, M., Kagoshima, T., Akamine, M.: One sentence voice adaptation using GMM-based frequency-warping and shift with a sub-band basis spectrum model. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 5124–5127 (2011)

14. Cappé, O., Laroche, J., Moulines, E.: Regularized estimation of cepstrum envelope from discrete frequency points. In: IEEE Workshop on Apps. Signal Process. to Audio & Acoustics, pp. 213–216 (1995)
15. CMU ARCTIC speech synthesis databases, http://festvox.org/cmu_arctic/
16. Erro, D., Sainz, I., Navas, E., Hernaez, I.: HNM-based MFCC+F0 extractor applied to statistical speech synthesis. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 4728–4731 (2011), http://aholab.ehu.es/ahocoder

Evaluation of a New Beam-Search Formant Tracking Algorithm in Noisy Environments

Dayana Ribas González[1], José Enrique García Laínez[2], Antonio Miguel[2], Alfonso Ortega Gimenez[2], Eduardo Lleida[2], and José Ramón Calvo de Lara[1]

[1] Advanced Technologies Application Center (CENATAV), 7a 21812 e/ 218 y 222, Rpto. Siboney, Playa, C.P. 12200, La Habana, Cuba
[2] Communications Technology Group (GTC), Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{dribas,jcalvo}@cenatav.co.cu, {jegarlai,amiguel,ortega,lleida}@unizar.es

Abstract. In this work we present the experimental evaluation of a new beam-search formant tracking algorithm under noisy conditions and compare its performance with three formant tracking methods. The proposed formant tracking algorithm makes use of the roots of the polynomial of a Linear Predictive Coding (LPC) as formant candidates. The best combination of formant candidates respect to a defined cost function are selected applying a beam-search algorithm. The cost function makes use of information about local and neighbor frames using trajectory functions in order to preserve the dynamics of the frequency of formants. Experiments were carried out with a subset of the TIMIT database, contaminated with various types and levels of noises. The results show that the beam-search formant tracker have a robust behavior in noisy environments and it is clearly more precise than the rest of compared methods.

Keywords: formant tracking, beam-search algorithm, noisy environments.

1 Introduction

The resonance frequencies of the vocal tract, known as formants, carry useful information to identify the phonetic content and articulatory information of speech as well as speaker and emotion discriminative information That is why formant tracking methods are widely used in automatic speech processing applications like speech synthesis, speaker identification, speech and emotions recognition. Those methods have to deal with the problem of the variability of the amount of formants depending on phoneme and the merging and demerging of neighboring formants over time, very common with F2 and F3. This is why, formant tracking is a hard task to face [1].

For decades, a number of works have been dedicated to designing formant tracking methods. Formant trackers usually consists of two stages: firstly the

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 40–48, 2012.

speech is represented and analyzed for obtaining some formant frequency candidates and secondly the selection of those candidates is done, taking into account some constraints. Those constraints are related with the acoustical features of the formant frequencies, the continuity of formant trajectory, etc.

One of the most extended methods of spectral analysis for formant tracking consists of extracting the roots of the polynomial of LPC, that has been shown to be effective in detecting the peaks of the spectrum [2]. In [3], a Gammatone filterbank followed by a difference of gaussians spectral filtering shown to enhance the formant structure. In [4], a method to segment the spectrum as a tuple of order-2 resonators was proposed. The method produces smooth formant frequencies in a frame by frame basis without any temporal information. However, it has the drawback of not representing well frames with more than 4 formants.

There has been considerable effort in the speech community to propose methods in the stage of formant selection. Probabilistic methods for estimating formant trajectories have been used successfully in recent years. Within this group are methods based on the Bayesian filtering like Kalman Filters [5] and particle filters [3] or Hidden Markov Models (HMM) [6]. Previous algorithms based on continuity constraints made use of dynamic programming and the Viterbi algorithm [7] [8] [9]. However, Viterbi based algorithms have the limitation that the cost function of a hypothesis only depends on the current observation, and the last state. In [10] we proposed a beam-search algorithm for formant tracking, that is able to incorporate trajectory information to the cost function, overcoming the limitation of the Viterbi search. In this paper we evaluate this algorithm in several noisy environments and we compare its performance with three formant tracking methods.

2 The Proposal: Beam-Searching Algorithm

The proposed formant detector can be decomposed in two main stages: The first is the formant frequency candidate extractor, where a set of frequencies and their bandwidths are chosen as possible formants. The roots of the polynomial of the LPC coding were used as formant candidates [7] [9].

The second stage is a beam-search algorithm for finding the best sequence of formants, given the frequency candidates. A mapping as proposed in [7] [9] of frequency candidates to all possible combinations of formants is chosen. For this purpose, $h_t = \{F1; B1; F2; B2; F3; B3; F4; B4\}$ is a possible formant tuple at frame t, obtained by means of a mapping from frequency candidates and formed by frequency (F) and bandwidth (B) information. The algorithm tries to find the best sequence of mappings, by applying a cost function that makes use of both local and global information. Its main advantage is to make no Markovian assumptions about the problem, i.e the evaluation of hypothesis in a frame takes into account the hypothesis defined in all previous frames unlike the Viterbi search [7, 9] which only uses previous state information. This feature allows to incorporate efficiently trajectory functions in the algorithm for representing the formant frequency dynamics.

The set of M active hypotheses in a frame t is represented by the group $P_t = \{p_{t,1}, p_{t,2}, ...p_{t,x}....p_{t,M}\}$, where a hypotheses $p_{t,x}$ is composed of an accumulated cost $acc_{x,t}$ and a history of mappings $z_t = h_1, h_2...h_t$. For obtaining the hypotheses P_t set in frame t, the set is propagated through all possible combinations of formant candidates $o_t = \{h_{t,1}...., h_{t,w}, ...h_{tU}\}$ where U is the total number of possible frequency mappings at frame t. This gives the extended group $PE_t = \{p_{t,1,1}, p_{t,x,w}...p_{t,M,U}\}$, and the accumulated cost of each new hypotheses $p_{t,x,w}$ is:

$$acc_{x,w,t} = acc_{x,t-1} + c(p_{t,x,w}, h_w) \tag{1}$$

The set PE_t is sorted according to the accumulated cost $acc_{x,w,t}$, and it produces the new group of M active hypotheses P_{t+1}, where the hypotheses with higher accumulated cost are maintained. This process is repeated for each frame until the end of the stream is detected, and the history of formants of the best hypothesis is selected as the final result. This search algorithm is illustrated in Fig. 1, where the M value represents a compromise between accuracy and execution speed.

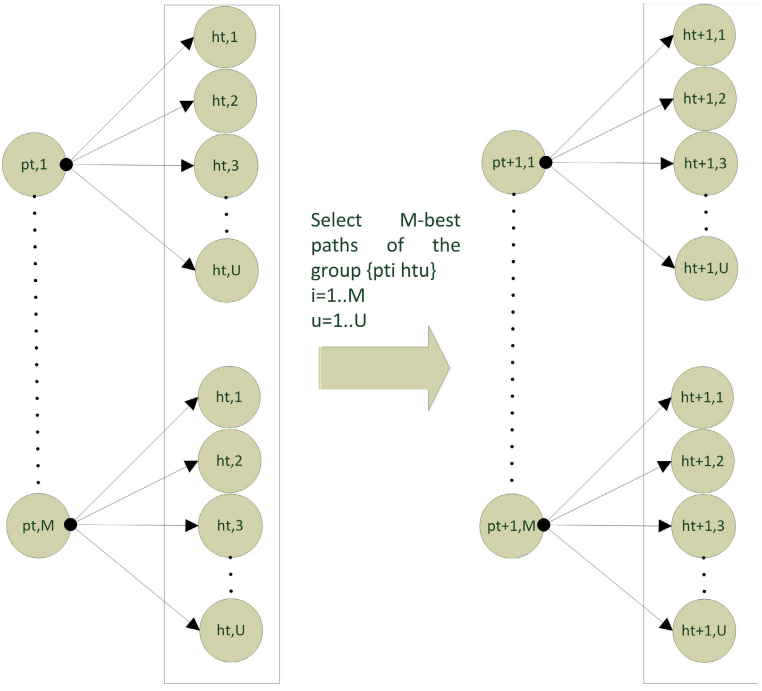


Fig. 1. Diagram of beam-search algorithm

2.1 The Cost Function

The cost function is defined as:

$$c(p_{t,x,w}, h_w) = cfrequency + cbandwidth + ctrajectory + cmapping \quad (2)$$

It uses both local and global observations for choosing the best sequence of formants. The part of the cost function that makes use of local information (that is, the current frame) contains the terms cfrequency, cbandwidth (defined as in [7]) and cmapping:

$$cfrequency = \alpha \sum_i |(F_i - norm_i)/norm_i| \quad (3)$$

$$cbandwidth = \beta \sum_i (B_i) \quad (4)$$

where $norm_i = 500, 1500, 2500, 3500$ and $i = \{1, ..., 4\}$ is the formant number.

$$cmapping_i = \begin{cases} 0 & if\ BWmin_i > THR \\ \frac{THR - BWmin_i}{\gamma_i} & if\ BWmin_i < THR \end{cases} \quad (5)$$

$$cmapping = \sum_i cmapping_i \quad (6)$$

where $BWmin_i$ is the minimum bandwidth of the frequency candidates that are discarded and that would be valid for the formant i in this mapping; γ_i and THR are constants. The part of the cost function that employs global information assumes that the frequency of each formant follows a smooth trajectory. This term is intended to take into account when a mapping is discarding some frequency peak with a low bandwidth.

$$ctrajectory = \theta \sqrt{\sum_{i,w} \frac{F_{i,w} - F_{i,\widehat{w}}}{B_i}} \quad (7)$$

Where $w = \{0, ..., W-1\}$ and W is the order of the trajectory function and $\widehat{F}_{i,t-w}$ is the estimated value of formant i, at frame $t-w$, assuming that $F_{i,t}, ..., F_{i,t-(W-1)}$ is approximated by a known function; $1/B_i$ is the weighted term of the trajectory, in order to give more importance to frames that have lower bandwidth; α, β and θ are constant for representing the weight of the terms. In the experiments, linear and quadratic functions were used, approximated with the least squares method. However, we assume that there is room for improvement in the modeling of such trajectory.

The trajectory term that makes use of several past frames justifies the use of the tree beam-search algorithm in place of the Viterbi decoding algorithm. One of the main benefits of this trajectory model is that it allows to recover observation errors in frames between obstruent and vowel, thanks to contiguous frame evidences.

An advantage of this continuity constraint compared with previous works is that this function does not increment costs when a change in the value of two consecutive frequencies occurs, as considered in [7] [9]. In addition, this global function will help the algorithm to correct errors in difficult frames where the frequency candidates do not give clear evidences. Within this group are frames between obstruent and vowel and frames corrupted by noise.

3 Experiments and Results

For comparison purpose three formant tracking methods were selected: Mustafa's proposal [11], Welling and Ney's algorithm [4] and Wavesurfer's method from Snack toolkit [12]. The performance of the formant tracking methods evaluated were measured carrying a quantitative evaluation using the VTR-Formant database [13]. This database contains the formant labels of a representative subset of the TIMIT corpus with respect to speaker, gender, dialect and phonetic context. In these experiments, 420 signals from VTR database were processed and the mean absolute error (MAE) between formants estimated for all formant tracking methods and VTR database were computed. All speech material used was digitized at 16 bits, at 10000 Hz sampling rate. The pitch ESPS algorithm from Snack toolkit was used, for obtaining the MAE only taking into account voiced frames.

Figure 2 shows the formant estimation achieved in a selected speech signal of TIMIT database, with the method proposed and the three methods used for comparison, besides the reference computed with VTR database. This qualitative view of the formant trackers obtained with each method allows to see the benefits of our tracking algorithm. In the figure it can be seen how Welling and Ney's algorithm achieve formant tracking lines quite accurate, however sometimes it has a poor performance, mainly in the tracking of F1. Wavesurfer's obtained tracking lines very similar to the reference, however the method proposed sometimes outperforms it, for example in the tracking of F3 and F4 between 0,5 and 1 second. Mustafa's algorithm achieved the worst performance of all the methods used.

Table 1. MAE (Hz) for formant estimations obtained with LPC beam-search algorithm, Wavesurfer, Welling and Ney and Mustafa's algorithms

Methods	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
LPC-beam-search	18.39	27.96	35.26	69.01
Wavesurfer	29.95	57.66	76.53	76.44
Welling-Ney	37.53	47.33	52.53	67.32
Mustafa	28.11	80.22	82.54	75.63

The Table 1 shows the performance of the four methods evaluated in clean speech. It can be observed how the proposed tracking algorithm outperforms consistently all the formant extractor in most cases. Notice that the order of accuracy

in the methods evaluated is: LPC-beam-search, Welling-Ney, Wavesurfer and finally Mustafa, however in F1, Mustafa outperforms Welling-Ney. Wavesurfer is better than Welling-Ney in the tracking of F1, however for F2 and F3 its performance decrease, taking into account that these are the harder resonances to follow. The F4 performance has less importance because this formant in VTR-database is not manually labeled.

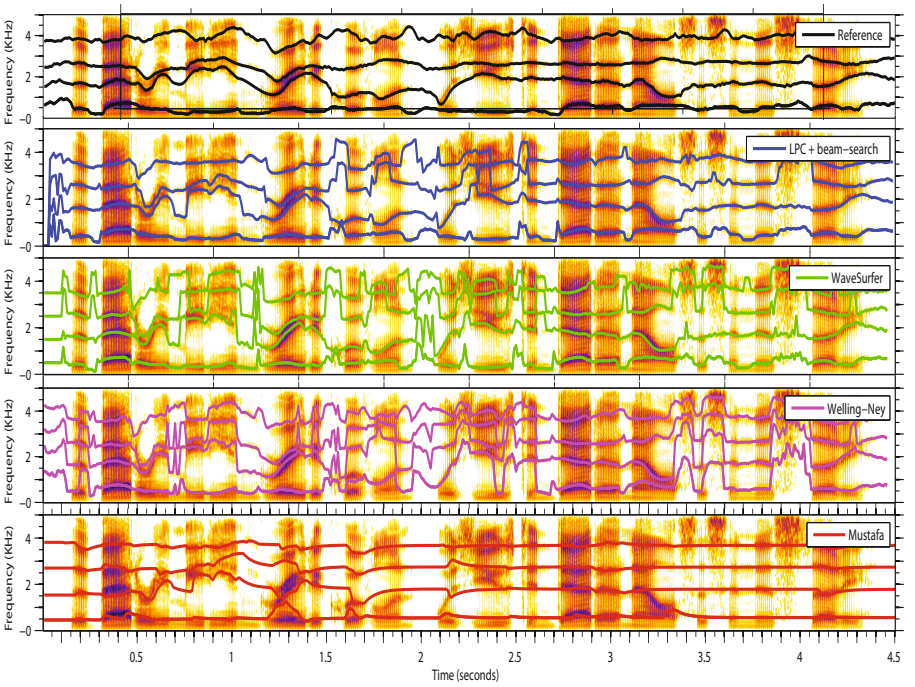


Fig. 2. Example results of a signal of TIMIT database with all the methods evaluated

Additional tests were carried out in noisy environments. The corrupted speech signals come from four different noise environments:

- stationary white noise
- pseudostationary street noise, which is a mixture of different noises
- music from Guns and Roses band, highly harmonic and non-stationary noise
- babble noise, special case of non-stationary noise, is the voice of other speakers

All those types of noise were added electronically to test speech signals at different SNR levels, from 0 to 20 dB in 5 dB steps.

Figure 3 shows the MAE in the noisy environments evaluated. For each type of noise the behavior of the methods is quite different. Notice that stationary white noise is the most challenge type of noise, given by the worst MAE of formant trackers shown in the corresponding plot. On the other hand, for all methods

in street, music and babble noise, from SNR = 10 dB, F1 has a behavior quite stable, besides F2 and F3 have a slight decrease of the slope of the MAE curves. This fact gives an idea of the robustness of formant trackers over SNR = 10 dB.

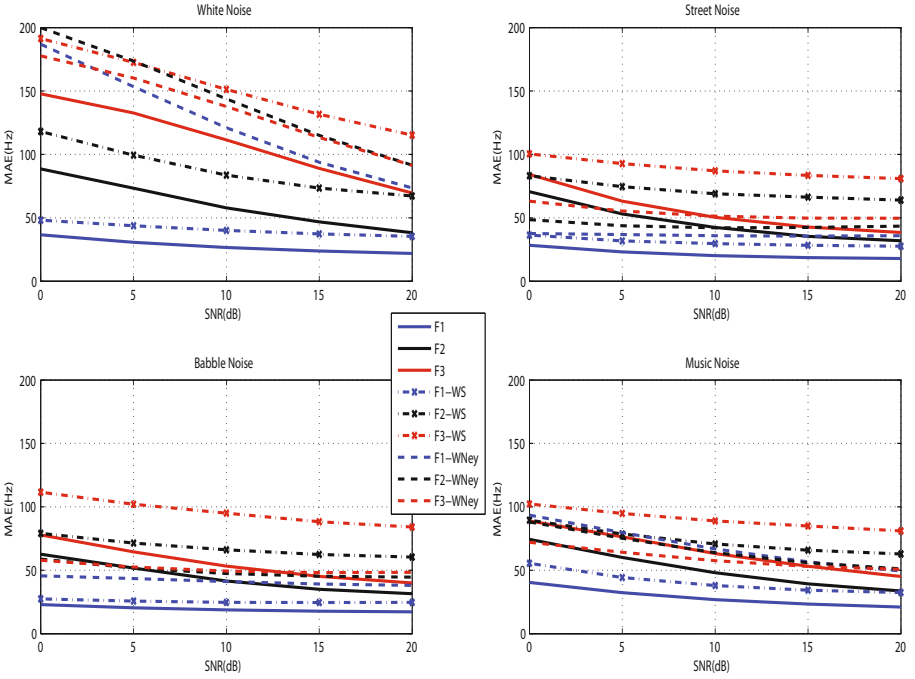


Fig. 3. MAE of formant estimation with LPC-beam search algorithm, Welling-Ney (WNey) algorithm and Wavesurfer(WS) algorithm vs VTR database in noisy environments

Figure 3 shows that the proposed method in noisy environments outperforms the other methods in most conditions evaluated. Nevertheless, Welling-Ney algorithm obtains the most precise F3 in music and street noise for SNR below 10 dB, and also is the best method in F2 for street noise in SNR below 10 dB. Concluding that the Welling-Ney method is more robust to narrow band noise (music and street noise) than the methods based on LPC (Wavesurfer and LPC beam-search). The spectral segmentation performed in the Welling-Ney method based on the searching of the 4 best spectral regions with dynamic programming, makes this method robust against this kind of noise, unlike LPC based methods that use as formant candidates 5 or 6 peaks. A narrow band noise is a good candidate to be confused with a formant and to be selected, because frequently it has lower bandwidth than a speech formant.

In white noise Welling-Ney and Wavesurfer's algorithms performs very inaccurate, with MAE near 200 Hz. However for babble noise Welling-Ney achieved very low errors, even in F2 and F3, for low values of SNR, it outperforms LPC beam-search method.

Figure 4 shows the formant tracking obtained with three of the methods evaluated and the reference over a spectrogram of the same speech signal used in Fig. 2 corrupted by babble noise with SNR = 10dB. Notice that the proposed method achieves soft formant curves, thanks to the trajectory functions combined with the beam-search algorithm. The other methods generate curves with a lot of spikes, which are due to the uncertainty introduced by the noise, that could mask the spectral features for detecting the formant candidates. So, if poor continuity constraints are incorporated, the formant trackers become very unstable and tend to have fast changes in the detected formants, in noisy environments. This is the case of the Wavesurfer formant tracker. On the other side the Welling-Ney formant tracker does not include any continuity constraint, and this is why it has this behavior.

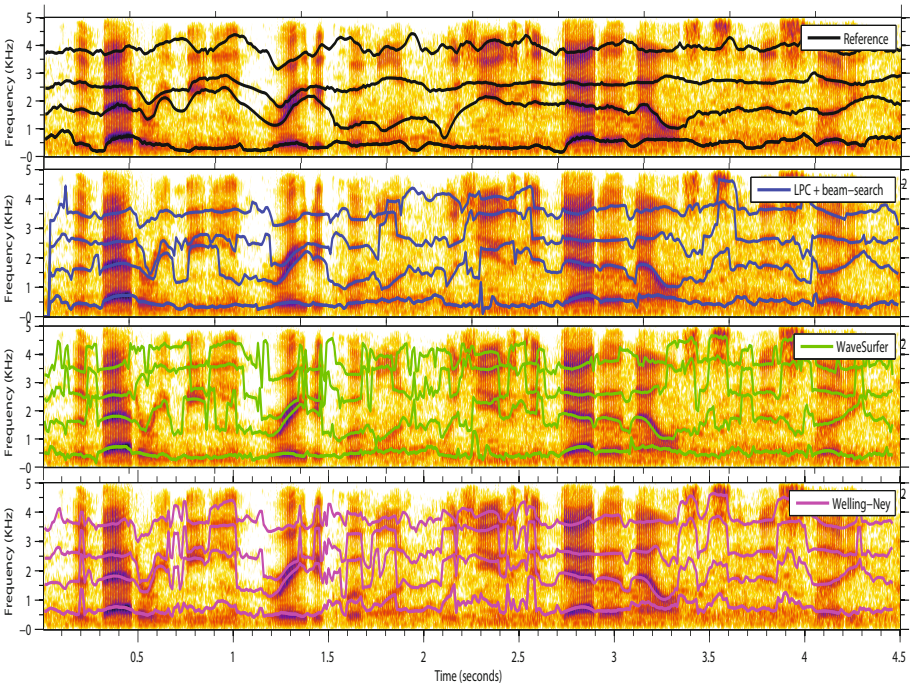


Fig. 4. Example results of a signal of TIMIT database corrupted by babble noise with SNR = 10dB with all the methods evaluated

4 Conclusions

In this paper we present an evaluation of the LPC-beam searching method in noisy environments and a comparison with three formant tracking algorithms. In spite of the proposed method not being designed with specific techniques noise compensation, it presents a very robust performance for all the types of noises

evaluated. In fact, results show that in most cases LPC beam-search method proposed performs better than Wavesurfer's, Mustafa's and Welling-Ney formant tracking algorithm. Furthermore, a feature that makes the beam-search algorithm attractive is that it produces smooth formant trajectories even in corrupted signals, while the other methods are very spiky in presence of noise.

Acknowledgements. This work has been partially funded by Spanish national program INNPACTO IPT-2011-1696-390000.

References

[1] Rose, P.: Forensic Speaker Identification. Taylor and Francis Forensic Science Series (Robertson, J. (ed.)). Taylor and Francis, London (2002)
[2] McCandless, S.: An algorithm for automatic formant extraction using linear prediction spectra. IEEE TASSP ASSP-22, 135–141 (1974)
[3] Gläser, C., Heckmann, M., Joublin, F., Goerick, C.: Combining auditory preprocessing and Bayesian Estimation for Robust Formant Tracking. IEEE Trans. Audio Speech Lang. Process. (2010)
[4] Welling, L., Ney, H.: Formant Estimation for Speech Recognition. IEEE Transactions on Speech and Audio Processing 6(1) (1998)
[5] Mehta, D.D., Rudoy, D., Wolfe, P.J.: KARMA: Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. Stat. AP (2011)
[6] Messaoud, Z.B., Gargouri, D., Zribi, S., Hamida, A.B.: Formant Tracking Linear Prediction Model using HMMs for Noisy Speech Processing. Int. Journal of Inf. and Comm. Eng. 5(4) (2009)
[7] Talkin, D.: Speech formant trajectory estimation using dynamic programming with modulated transition costs. JASA 82(S1), 55 (1987)
[8] Deng, L., Bazzi, I., Acero, A.: Tracking Vocal Tract Resonances Using an Analytical Nonlinear Predictor and a Target-Guided Temporal Constraint (2003)
[9] Xia, K., Espy-Wilson, C.: A new strategy of formant tracking based on dynamic programming. In: Proc. ICSLP (2000)
[10] García Laínez, J.E., Gonzalez, D.R., Artiaga, A.M., Solano, E.L., De Lara, J.R.C.: Beam-Search Formant Tracking Algorithm based on Trajectory Functions for Continuous Speech. To be Plublished in Proceedings of CIARP 2012 (2012)
[11] Mustafa, K., Bruce, I.C.: Robust formant tracking for continuous speech with speaker variability. IEEE Transactions on Speech and Audio Processing (2006)
[12] Snack toolkit, http://www.speech.kth.se/wavesurfer
[13] Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y., Alwan, A.: A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. In: ICASSP (2006)

On the Influence of Automatic Segmentation and Clustering in Automatic Speech Recognition

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo, and Antonio Cardenal-Lopez

Multimedia Technologies Group (GTM),
AtlantTIC Research Center, Universidade de Vigo
E.E. Telecomunicación, 36310 Vigo Spain
{plopez,ldocio,carmen,cardenal}@gts.uvigo.es

Abstract. An automatic speech recognition (ASR) system needs a previous segmentation stage that differentiates between speech and non-speech. Other information such as "who spoke when" can be proportioned to the ASR system, allowing it to perform speaker adaptation. This paper studies the influence of automatic speech segmentation and speaker clustering on ASR performance, in order to detect the weak points of the diarization system by analyzing what causes the different types of recognition errors: insertions, suppressions and substitutions. Experiments are run on the Galician broadcast news database Transcrigal, and results show that the speaker diarization system presented in this work is suitable as a previous step to ASR, as the performance is almost the same as the obtained when using manual segmentation and clustering.

Keywords: automatic segmentation, automatic speech recognition.

1 Introduction

Automatic speech recognition (ASR) is a task in which a computer has to identify the words that are spoken by a human in order to generate a transcription of their speech. Nowadays it has a huge range of applications in different fields: online or offline transcription of TV programs [9], with the aim of adapting television to disabled people by generating subtitles automatically or just transcribing and storing the programs in databases for searching in multimedia contents; automatic translation of spoken documents, obtaining a transcription or speech spoken in the target language by means of a text-to-speech system; recognition of speech for natural language question answering [15]; communication with devices such as mobile phones or GPS navigators while driving [12]; and so forth.

The data input of an ASR system should be speech only, because other types of audio information such as music or noise will cause the recognizer to unsuccessfully try to recognize these data. Thus, everything that is not speech has to

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 49–58, 2012.

be removed before recognition. ASR systems commonly use a voice activity detector (VAD) to discriminate between speech and non-speech, but no additional information is proportioned, such as speaker turns. Therefore, another approach should be applied in order to enhance the performance of the ASR system, such as a speaker diarization system.

Speaker diarization is the task consisting in, given an audio stream, deciding "who spoke when", which embraces different tasks: discrimination between speech and non-speech, detection of speaker change-points in the audio stream, and labeling of the speech segments by speaker. In other words, speaker diarization includes two tasks: audio segmentation and speaker clustering. ASR can obtain huge benefits from speaker diarization; for example, only speech segments will be addressed to the ASR system for their recognition. Moreover, splitting the speech parts obtaining segments where there is only speech from one speaker makes it easier to detect the beginning or the end of the sentences, which facilitates the task of the language models and, therefore, of the ASR system. Also, when performing speaker clustering, it is known which segments include speech from the same speaker; thus, this data can be used to train a specific model for each speaker, which helps to improve the performance of ASR [5]. On the other hand, when the accuracy of the speaker diarization task is poor, the contrary effect might be produced on the ASR stage; for example, non-speech segments may be addressed to the ASR, or sentences may be split creating two sentences that make no sense for the language model. Thus, diarization is important in speech recognition when a good diarization is obtained. Two types of errors can be found when performing speech/non-speech detection: missed speech (speech segments are labeled as non-speech) and false alarm speech (non-speech segments are labeled as speech). In the same way, two types of errors can be found in speaker segmentation: insertions (detecting change-points that are not actual speaker change-points) and deletions (actual speaker change-points that are not detected). Also, in speaker clustering, speech from a speaker can be assigned to an incorrect speaker. All these types of errors might affect ASR performance.

This paper studies the influence of speaker diarization on ASR by comparing the recognition results obtained with automatic segmentation and clustering of the audio stream and with manual segmentation and clustering. A segmentation algorithm based on the Bayesian information criterion (BIC) approach is used, and it features a probabilistic approach that models the occurrence of false alarms by means of a Poisson process [7]. For speaker clustering an agglomerative hierarchical clustering (AHC) strategy is used. A database in Galician language is used to assess the performance of the whole system, the Transcrigal broadcast news database [4]. This database features spontaneous and planned speech; performance on both types of speech is also assessed.

The rest of the paper is organized as follows: Sects. 2 and 3 describe the speaker diarization and the ASR systems, respectively; Sect. 4 describes the experimental framework; Sect. 5 presents the results obtained; a discussion of the results is presented in Sect. 6; and Sect. 7 describes some future work.

2 Automatic Speaker Diarization System

2.1 Speaker Segmentation and Speech/Non-Speech Classification

The segmentation strategy used in this paper is fully described in [7], but a brief description is given here. As shown in Fig. 1, it is a four stage segmentation strategy:

- Change detection: a coarse segmentation is done by means of the distance changing trend segmentation (DCTS) algorithm [14].
- Change refinement: anytime the DCTS algorithm detects a change-point, it is refined by using the BIC algorithm. The value of ΔBIC is observed [11]; there are three possibilities at this stage:
  - $\Delta BIC < 0 \Rightarrow$ the change-point is discarded and the system returns to the change detection stage.
  - $\Delta BIC > \Theta$ the change-point is accepted and the system goes to the next stage.
  - $0 < \Delta BIC < \Theta$ the change-point is accepted with probability p.

  Θ is a threshold for ΔBIC, because if ΔBIC is high the change-point is more likely to be a real change-point, while if ΔBIC is too low it is possible that the change-point is not a real change-point. The probability p is a discard probability, and it increases following a Poisson cumulative density function. The mean of this Poisson distribution is the expected number of change-points μ. This approach is fully explained in [7].
- Segment classification: the accepted change-point and the previous one form a segment of data. The likelihood of this data with Gaussian Mixture Models (GMM) trained with speech, non-speech and music is computed, assigning the type corresponding to the GMM that achieves the highest likelihood.
- Adjacent segments merger: when there are two speech segments in a row and they are both labeled as "male" or "female", it is possible that the change-point in the middle of them is a false alarm. Thus, the cross likelihood ratio (CLR) of the two segments is computed: if this value exceeds a threshold, the segments are too similar to each other and the change-point between them is discarded.

The output of the speech segmentation algorithm is a set of segments labeled as speech or non-speech. The non-speech segments are discarded and the speech ones are addressed to the ASR system in order to transcribe them.

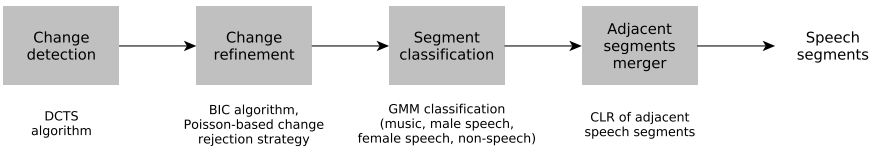


Fig. 1. Speaker segmentation system

2.2 Speaker Clustering

In the speaker clustering strategy developed for this system, each speech segment is modeled as in a common approach for speaker recognition [10]: a universal background model (UBM) is trained, and this UBM is adapted to each speech segment by using the Maximum a Posteriori (MAP) technique. Given the set of speech segments $S = (S_1, \ldots, S_n)$ (where n is the number of segments), obtained as described in Sect. 2.1, the UBM is adapted to each of them obtaining a set $\Theta = (\Theta_1, \ldots, \Theta_n)$, where $\Theta_i \in \Re^{M \times N}$ (M is the number of Gaussian mixtures and N is the size of the feature vectors) are the normalized means of the adapted Gaussian components. These means (rows) are concatenated in order to obtain a set of supervectors $V = (V_1, \ldots, V_n)$ [1]. A matrix $M \in \Re^{n \times MN}$ is constructed, where rows i are the supervectors V_i. Thus, the whole set of segments is represented by means of matrix M.

Once M is obtained, AHC is performed by using the general purpose clustering toolkit CLUTO [3]. In this implementation, the most similar pair of clusters is merged according to the group average-link algorithm, using the cosine distance as similarity measure. The desired number of clusters, which is not calculated by CLUTO, is computed after applying AHC: the number of clusters that obtains a trade-off between the intra-cluster and extra-cluster similarities is chosen. The aim is to minimize the intra-cluster similarity and maximize the extra-cluster similarity, obtaining a set of clusters with similar elements in each cluster and non-similar elements in different clusters.

Figure 2 summarizes the whole clustering procedure.

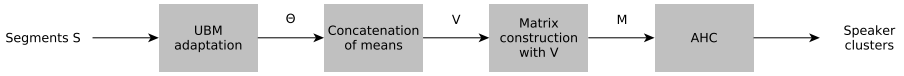


Fig. 2. Speaker clustering system

3 Automatic Speech Recognition System

A large vocabulary continuous speech recognition module is used in this work for generating the automatic transcription of the audio files [2]. The basic decoder has two stages, a Viterbi algorithm working in a synchronous way with a beam search and an A^* algorithm to obtain the N-best hypothesis. The transcription of each audio document is obtained in two passes. The first pass is as follows:

– First, acoustic model selection is performed by applying a phonetic recognizer to the 10 first seconds of a speech segment. The models that achieve the best acoustic score are selected.

– A VAD based on energy thresholds and a simple state machine is then applied to further divide the segment in smaller chunks. This procedure provides an important reduction of the computation effort, with a minor degradation of the recognition results.
– Lastly, the decoder is applied to the VAD segments using a 3-gram based language model (LM) and the selected models, obtaining a first transcription.

The second pass is basically an acoustic-model adaptation stage plus a new recognition pass:

– Using the results provided by the first pass, a phone-level transcription is obtained.
– This transcription is used to perform a maximum likelihood linear regression+maximum a posteriori (MLLR+MAP) adaptation of the acoustic models.
– A second recognition pass using the new acoustic models is then performed.

Usually the acoustic model adaptation is applied sequentially to each segment provided by the speaker segmentation module. However, when the speaker clustering information is available, the first pass is applied to the whole audio document, and then the acoustic model adaptation is performed grouping all segments belonging to the same speaker.

This whole procedure is represented in Fig. 3.

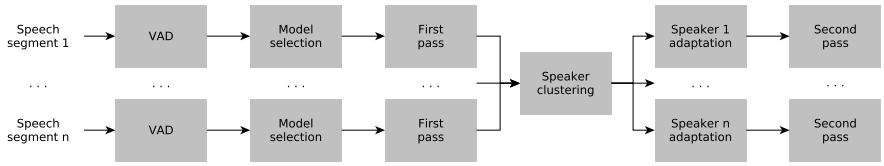


Fig. 3. Speech recognition system

4 Experimental Framework

4.1 Description of the Database

Performance of the ASR system when using manual and automatic segmentations is assessed by running some experiments on Transcrigal-DB [4], a database featuring broadcast news shows in Galician language. As it is usual in this kind of programs, there are several habitual speakers, who usually speak planned speech, and there are also other speakers speaking spontaneous speech. Different accents and dialectal varieties of the Galician language appear, specially in the case of the non-habitual speakers.

The database has been manually segmented, labeled and transcribed, so there is a reference for the diarization and for the transcription. Three datasets were defined: a train set (25 hours) for training the GMMs of speech, music and non-speech; a development set (2 hours) for tuning the parameters of the automatic segmentation system; and a test set (13 hours) to assess the performance of the segmentation and the transcription.

4.2 Metrics

A popular metric used to measure the performance in the ASR stage is the word error rate (WER):

$$WER = \frac{S + D + I}{W} \tag{1}$$

The WER is the average number of word errors taking into account three different types of error: word substitutions (S), word deletions (D) and word insertions (I). WER is defined as the addition of these three types of error divided by the number of reference words W. When the automatic transcription is identical to the reference transcription, WER is equal to zero.

The amount of speech that is lost in the segmentation stage (missed speech, MS) and the amount of non-speech that is labeled as speech (false alarm speech, FA) are measured in order to analyze the performance of the segmentation system and to try to find out if there is any relationship between the WER and these segmentation and classification errors. The accuracy of the speaker diarization stage is measured by means of the speaker error (SPKE), which is the percentage of speech assigned to an incorrect speaker. The combination of these three metrics, which measures the performance of the whole diarization procedure, is known as diarization error rate (DER) [13].

These performance measures are computed by using the speech recognition scoring toolkit [8] developed by NIST for the Rich Transcription Evaluation campaigns [13]. Concretely, the tools md-eval and SCLITE are used to assess the diarization and the transcription, respectively.

4.3 Description of the Experiments

The aim of this paper is to analyze the influence of automatic speaker segmentation and clustering on the ASR task. To do so, two experiments are performed: in the fist one, the test set of the database described in Sect. 4.1 is automatically segmented and transcribed; and in the second one, the test set is automatically segmented and clustered, and then it is transcribed. It has to be noticed that in the first experiment, only the first pass of ASR is performed, while in the second experiment two passes plus speaker model adaptation are performed, as described in Sect. 3.

Some reference results are necessary in order to study the impact of the automatic diarization on the transcription; hence, the manual segmentation of the test dataset is also transcribed by the ASR system; this will be considered as the

baseline of the whole system, because the best diarization that can be obtained by the automatic diarization system is ideally equal to the manual one. Both diarization and recognition results are presented.

As commented in Sect. 4.1, there are habitual speakers speaking planned speech, and other speakers speaking spontaneous or quasi-spontaneous speech. In order to assess the performance of the recognizer in these two cases, ASR results for all the speakers, habitual speakers only and other speakers are presented.

5 Experimental Results

5.1 Features

In this work, the acoustic features extracted from the speech utterances are 12 Mel-frequency Cepstral Coefficients (MFCC), extracted using a 25ms Hamming window at a rate of 10ms per frame, and augmented with the normalized log-energy and their delta and acceleration coefficients. In the segmentation algorithm only 13 features are used (12MFCC and log-energy), while in the classification task with GMMs, the clustering stage and the recognition stage the 39 features are used.

5.2 Results

The test dataset of Transcrigal-DB is automatically segmented and the segments are classified as speech/non-speech as described in Sect. 2. The free parameters of the system were previously tuned on the development dataset: the parameter λ of the BIC algorithm [11], the threshold Θ and the number of expected change-points μ. The tuned values are $\lambda = 2.7$, $\Theta = 500.0$ and $\mu = 20$. The GMMs

Table 1. ASR results on Transcrigal database with speaker segmentation

	Speakers	Substitutions	Deletions	Insertions	WER
Manual segmentation	All	14.9%	4.9%	4.4%	(24.2 ± 3.82)%
	Habitual	12.1%	3.4%	4.6%	(18.3 ± 4.55)%
	Others	18.2%	5.9%	4.3%	(28.4 ± 2.1)%
Automatic segmentation	All	14.9%	6.5%	3.6%	(25.0 ± 4.61)%
	Habitual	10.6%	4.5%	3.6%	(18.7 ± 5.53)%
	Others	17.9%	7.8%	3.7%	(29.4 ± 2.3)%
Automatic segmentation and clustering	All	13.7%	6.2%	3.3%	(23.1 ± 4.5)%
	Habitual	9.9%	4.3%	3.3%	(17.6 ± 5.2)%
	Others	16.4%	7.5%	3.3%	(27.2 ± 2.3)%

Table 2. Diarization results on Transcrigal database

MS	FA	SPKE	DER
5.6%	5.3%	19.5%	30.41%

used to classify the segments in speech, music or non-speech have 64 Gaussian mixtures.

Table 1 shows the ASR results obtained both with manual and automatic segmentations. The WER is represented with the 95% confidence interval.

6 Discussion

Results in Table 1 show the performance of the ASR system when using manual segmentation, automatic segmentation and automatic segmentation followed by clustering (with adaptation of the speaker models in the ASR stage). It can be seen that manual and automatic segmentation achieve the same number of substitutions, but automatic segmentation has more deletions and manual segmentation has more insertions. Thus, what makes the difference between the two segmentations are the segment boundaries: the automatic segmentation has more tight boundaries, which causes the initial phoneme of the sentences to be cut in some occasions, making the ASR system to get lost and generating suppressions; on the other hand, the manual segmentation usually has non-speech frames at the beginning and the end of speech segments, leading to insertions when the recognizer tries to recognize audio parts that are not speech. This last type of error may be due to the fact that the ASR system has models for speech, silence and music, but it does not have a model for noise; thus, noise may be confused with speech and, therefore, recognized as vocabulary words. Table 1 also shows that when performing a second pass in the recognition stage adapting the speech segments to the speaker models as indicated by the automatic clustering, a general improvement is obtained. The WER in this case is even lower than with the manual segmentation.

As expected, the habitual speakers obtain lower error rates than the other speakers. This is due to the fact that in general the habitual speakers speak planned speech and the others speak in a more spontaneous way, and also because the habitual speakers have their own speaker models. In both cases WER improves when performing speaker adaptation, obtaining better results than with the manual segmentation.

There are other details that have been extracted from a thorough examination of the recognition results. For example, some segments of the habitual speakers showed a WER much higher than expected. Sometimes the automatic segmentation includes non-speech at the beginning of the sentences (false alarm speech error) and, as the acoustic model selection is performed on the ten first seconds of the segment, if the segment starts with non-speech the selection of the model is incorrect, causing the WER to rise. A similar problem occurs when speaker

turns are missing: when a speaker segment includes speech from different speakers, the acoustic model that is selected might be good for one of the speakers but not for the other one, raising the WER.

Table 2 shows the diarization results, which have some influence on the WER. 5.6% of the speech is missing in the automatic segmentation, causing suppressions on the transcription, and 5.3% of non-speech has been labeled as speech, causing insertions, as the recognizer transcribes data that is not speech.

After commenting the experimental results, it has to be said that the WER obtained with automatic segmentation and clustering is the lowest one, but the p-values show that this difference is not meaningful. Thus, the presented diarization system is perfectly suitable for its integration with an ASR system, because performance is as good as when using the best possible segmentation.

7 Future Work

As commented in Sect. 6, the automatic diarization system described in Sect. 2 achieves a good performance when compared to a manual diarization. Nevertheless, the influence of the miss-detected speech is noticeable, because it causes the ASR system to miss the first words of a speech segment in some cases. Thus, a technique to refine the segment boundaries should be incorporated to the segmentation system, in order to get rid of this problem.

Although Table 1 shows that the automatic segmentation has less insertions than the manual segmentation, this percentage can be reduced by improving the speech/non-speech detection, because labeling non-speech as speech forces the ASR system to transcribe something that is not speech, generating insertions.

With respect to the ASR task, acoustic speaker adaptation has shown to improve recognition performance. Nevertheless, the strategy to select the acoustic model should be improved, as it is easily influenced by errors in the speaker segmentation: as commented in Sect. 6, when a speech segment has non-speech at the beginning, this non-speech is used to select the speaker model, compromising the model selection. It was also commented that the ASR system used in this work does not have a model for noise; this causes the system to, in some cases, treat the non-speech information as speech and generating a transcription for it, which leads to an increase of the insertions. Thus, a model for noise should be incorporated. Also, to reduce the influence of noise in the transcription, acoustic channel adaptation should be tested in order to overcome the influence of mismatch conditions (different environmental conditions between train and test data) in recognition.

Once a suitable diarization strategy for speaker recognition has been obtained, future work will be focused in developing an online implementation of this automatic transcription system.

Acknowledgements. This work has been supported by the Galician Regional Government (CN2011/019, 2009/062), Spanish Government (TEC009-14094-C04-04 and FPI grant BES-2010-033358), and the European Regional Development Fund.

References

1. 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: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 97–100 (2006)
2. Cardenal-Lopez, A., Dieguez-Tirado, F.J., Garcia-Mateo, C.: Fast LM look-ahead for large vocabulary continuous speech recognition using perfect hashing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 705–708 (2002)
3. CLUTO - software for clustering high-dimensional datasets, http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview
4. Garcia-Mateo, C., Dieguez-Tirado, J., Docio-Fernandez, L., Cardenal-Lopez, A.: Transcrigal: A bilingual system for automatic indexing of broadcast news In: Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation, pp. 2061–2064 (2004)
5. Herbig, T., Gerl, F., Minker, W.: Fast Adaptation of Speech and Speaker Characteristics for Enhanced Speech Recognition in Adverse Intelligent Environments. In: Proceedings of 6th International Conference on Intelligent Environments, pp. 100–105 (2010)
6. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: a Review. ACM Computing Surveys 31(3), 264–323 (1999)
7. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Novel Strategies for Reducing the False Alarm Rate in a Speaker Segmentation System. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4970–4973 (2010)
8. NIST Speech Recognition Scoring Toolkit, http://www.itl.nist.gov/iad/mig/tools/
9. Ortega, A., García, J.E., Miguel, A., Lleida, E.: Real-Time Live Broadcast News Subtitling System for Spanish. In: Proceedings of Interspeech, pp. 2095–2098 (2009)
10. Reynolds, D., Quatier, T., Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
11. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6(2), 461–464 (1978)
12. Setiawan, P., Suhadi, S., Fingscheidt, T., Stan, S.: Robust Speech Recognition for Mobile Devices in Car Noise. In: Proceedings of Interspeech, pp. 2673–2676 (2005)
13. The NIST Rich Transcription Evaluation Project Website, http://www.itl.nist.gov/iad/mig/tests/rt/
14. Wang, Y., Han, J., Li, H., Zheng, T.: A Novel Audio Segmentation Method Based on Changing Trend of Distance between Audio Scenes. Journal of Communication and Computer 3, 22–30 (2006)
15. Yaman, S., Tur, G., Vergyri, D., Hakkani-Tur, D., Harper, M., Wang, W.: Anchored Speech Recognition for Question Answering. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 265–268 (2009)

Preliminary Results of Alignment of Text and Audio in News and Songs

Darwin Patricio Córdova Lucero and Doroteo Torre Toledano

ATVS, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain
dar.cordova@estudiante.uam.es, doroteo.torre@uam.es

Abstract. This paper addresses the problem of forced alignment in news and songs in order to get the times where every word of the transcriptions begins and ends. For this purpose two methods are used. The first one is basically a forced alignment process of the audio and text based on pre-existent models. The second one is a model-free method in which new models are trained on the audio to align producing as a result the aligned text and audio. For analysis of the songs, we have considered two versions of the same song: one is an *a capella* song (only voice with no music) and the other, the full song (with instrumental music included). Three songs have been selected from different singers and different styles. Regarding news, we have analyzed four speakers (2 females and 2 males). Analyzing all the results, we observe that news is better aligned than songs, as expected. The two methods work similarly in both *a capella* songs and news, but in the case of songs that include the instrumental part, the model-free method is much better.

Keywords: Alignment, Songs and Lyrics, Language Learning, Broadcast News.

1 Introduction

This paper presents preliminary experiments on alignment of songs and lyrics and texts and audio news. One of the purposes of this paper is to analyze the difference in the behavior of the forced alignment in songs and broadcast news. To that end, for the analysis of the news, we will be using four speakers (two females and two males), and for the analysis of the songs, we will consider three English songs from three different styles of music: the first one is a very fast-speed song (rap), the second one, a normal-speed song (pop), and, finally, a very slow-speed song (ballad). Two versions of these songs will be considered, one including instrumental music and one *a capella*.

Other of the purposes of this paper is to compare two different ways of producing the alignments. One way is based on pre-existent Hidden Markov Models (HMMs), and another a model-free approach based on training HMM models from scratch using only the audio to align, or this audio complemented by a set of similar audios.

Our main goal is the alignment of songs and lyrics to feed new songs and aligned lyrics into a web-based system (www.inglesdivino.com) that plays songs and videos and shows each word pronounced aligned in real time, among many other possibili-

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 59–68, 2012.

ties. This system tries to help students to learn and improve their English in less tedious ways. Having songs and lyrics aligned is very useful, for example, for students who are beginning to learn a new language, because they normally get lost when they try to follow the lyrics as they listen to the audio recordings. With this technology that wouldn't happen, since every word will be highlighted as accurately as possible while it's pronounced. All the experiments so far have been done in English, but in the future we plan to expand them to more languages. In general, this system will be useful for learning any language.

This problem is closely related to other similar problems that share in common the need to have audio and text aligned: TV subtitling, entertainment (i.e. karaoke), design of games based on synchronized audio, etc.

The issue of song and lyrics alignment has found some interest in the research community in the last years. Good examples are [1] where pre-existent models are used, [2] where dynamic programming and a model-free method is used and [3] where music and speech try to be first segregated and then pre-existent models are adapted to speech with music. On the other hand, the issue of broadcast news subtitling has been more studied due to its clear application, in particular to allow deaf people to access the content of the news. Broadcast news subtitling can be faced in two different ways, by using speech recognition and obtaining transcription and alignment from audio, as done in [4], or by exploiting knowledge from the news transcription used by the speakers to align text and audio as done in [5]. In this paper we will always use the text for the alignment. We will be comparing the problem of songs and lyrics and the one of text and news alignment and finally we will compare two methods for performing the alignment.

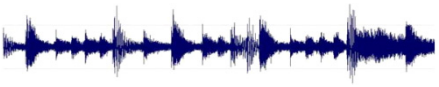
2 Proposed Methods

The alignments of text and audio will be performed using two methods: the first one is based on pre-existing English phonetic HMM models, and the other one (model-free method) is based on training HMM models from scratch using the audio to align (and possibly some similar complementary audios). In both cases, models will be used or trained using HTK [6]. Next subsections explain in more detail these methods.

2.1 Using Pre-Existing Models

In this method, we use English phonetic HMM models previously trained with 8 KHz English audio (TIMIT corpus [7]). The models have been created for each phone of English, with 40 Gaussians per state and 3 states per phone. For these experiments we use the models without any modification. In order to obtain our times of interest, we perform the following steps:

1. Prepare input data. In this case we need the audio recording and its transcription (a word level transcription).



Audio recording

HELLO WORLD

Transcription

Fig. 1. Audio recording and its transcription

2. Parameterize the audio. Here we convert the audio file to MFCCs (Mel Frequency Cepstral Coefficients).
3. Convert the word level transcription into a phone level grammar. For this, we use an English phonetic dictionary derived from the CMU pronouncing dictionary [8]. This phone level grammar will be used to create a network of HMM phone models.

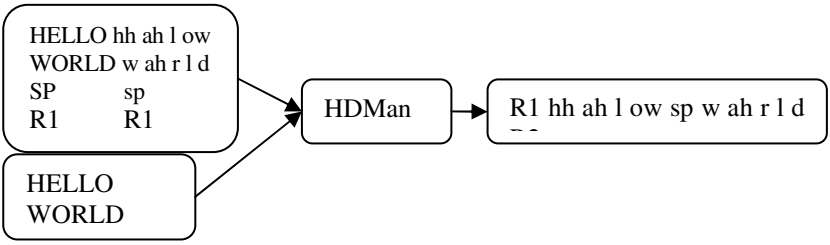


Fig. 2. Creation of phone-level grammar from word-level transcription and dictionary

4. Perform the alignment. The alignment is performed by the HVite tool. It will match the parameterized audio against the created network of HMMs and output the beginning and ending time for each phone and word.

Once the alignment is performed, we extract the beginning and ending time of each word, and then make the comparison with the manual reference. We will show the results in Section 3.

2.2 Model-Free Alignment: Aligning during Training

This method is based on the model training process. What we do here is to train phone models from the data we want to be aligned. In our case this data could be songs or news recording. During the training process, HVite and HERest are used for realigning and retraining the models, giving as a result a phone-level alignment of the input data. Compared to the previous method, this one has the advantage that it uses acoustic models that are completely adapted to the data to process with respect to speaker, presence of music, noises, etc. It is well known that the best results in recognition are achieved when we try to recognize the data used for training. That is precisely what we do in this method. Normally using test data for training is not fair, but in this particular application it is perfectly valid. We use as input data for model training the data (audio and text) we want to align, and then as a result of the training process we obtain the alignment. Of course, there are also disadvantages. The main one is that

using only the audio and text to align means using a very limited amount of data. We will try to alleviate this by adding other audios and texts from the same speaker and in similar conditions (to the extent that it is possible) to improve the training and alignment process. The following describes the steps of this method:

1. Prepare input data as in the previous method. We prepare the audio recording and its transcription (a word level transcription). The main novelty here is that we may be interested in preparing additional transcriptions and audios with similar features (speaker, acoustic conditions, etc.) to help in the training and alignment process by adding more data.
2. Parameterize the audio as in the previous method, converting it to MFCCs.
3. Convert the word level transcriptions into phone level grammar, as in the previous method, using again an English phonetic dictionary derived from the CMU pronouncing dictionary [8].
4. With all the necessary data prepared, we proceed to train the acoustic models of each phone appearing in the grammar we have previously defined. We start defining a prototype of a model and creating "flat start" monophones using the HTK HCompV tool. Then, these "flat start" monophones are re-estimated using the HERest tool. The purpose of this is to load all the "flat start" monophones and re-estimate them using the MFCC files generated from our training data (audios of our songs or broadcast news) and create a set of new models. We do this re-estimation four times.
5. In the final step a realignment of the training data is performed using the HVite tool. This tool can consider all pronunciations for each word (in the case where a word has more than one pronunciation in the grammar), and then output the pronunciation that best matches the acoustic data. HVite gives us a first alignment of the data. We use this alignment to re-estimate the models and get more accuracy. We re-estimate (with HERest) four more times using the output of the HVite (the first alignment). After this process, once all the re-estimation has been done, we have the models ready and use them to realign the training data. From the alignment obtained in this process, we will extract the final times for comparing with the manual reference.

3 Results

3.1 Experimental Data

For the experiments with broadcast news we have chosen four segments from YouTube containing four speakers: two females and two males. The duration of the audios is around a minute and a half. Regarding songs, three songs have been selected to cover different styles: The first one is a very fast-speed song (rap), the second one is a normal-speed song (pop) and the last one a very slow-speed song (ballad). The experiments for the model-free method will be performed with audios with a sampling rate of 44100 Hz. Two experiments will be carried out, the first one consists of introducing as input data only the song or piece of news to be aligned, and the second one

consists of adding extra audios to help in the training process that produces the alignment. In other words, in the last case, apart from the audio we want to align we introduce more audios from the same speaker or the same singer. These extra audios are only used to improve the accuracy of the alignment.

3.2 Results with the Model-Free Method

We will first show the results obtained with the model-free method. Results referring to the method based on pre-existing models will be shown in Section 3.3.

Results are presented showing the percentage of words with segmentation errors smaller than certain values of tolerances, which were chosen to be 50, 100 and 200 ms, because the target application is relatively robust to segmentation errors and most probably errors of 100 ms could remain unnoticeable. These evaluation metrics are similar to those used in [9]. Tables 1 and 2 show a comparison of results obtained on the experiments using only a single audio or additional audios for broadcast news speakers and singers.

These results show that, although there are some cases where the model-free method works well even with one audio, it is when we have access to other audios from the same speaker or from the same singer where the method reaches better performance. To illustrate this improvement when we add more input data in song and lyrics alignment, Figure 3 shows an example histogram of the absolute value of the errors found in the alignment when no added data is used and when only two additional audios are used. As it can be seen, the error in the alignment reduces considerably when adding more data.

It is important to note that there are speakers (speaker 4) and songs (song 3) that are particularly problematic for this method. In both cases we found that speech was slow, which seems to be particularly problematic for this method.

Table 1. Percentage of words (%) in broadcast news with errors smaller than three values of tolerance (50, 100 and 200 ms) with only one audio and with additional audios

	Single Audio			Additional Audios		
Tolerance	50 ms	100 ms	200 ms	50 ms	100 ms	200 ms
SPEAKER 1 (female)	(188 words and 1 audio)			(895 words and 4 audios)		
	7.45	19.68	19.68	29.79	50.53	93.09
SPEAKER 2 (female)	(223 words and 1 audio)			(1181 words and 4 audios)		
	24.50	49.67	91.06	28.81	55.30	96.69
SPEAKER 3 (male)	(319 words and 1 audio)			(1486 words and 3 audios)		
	1.57	3.13	10.97	35.11	57.68	96.87
SPEAKER 4 (male)	(318 words and 1 audio)			(950 words and 2 audios)		
	2.52	5.03	9.12	3.46	5.35	8.18

Table 2. Percentage of words (%) in songs with errors smaller than three values of tolerance (50, 100 and 200 ms) with only one audio and with additional audios

	Single Audio			Additional Audios		
Tolerance	50 ms	100 ms	200 ms	50 ms	100 ms	200 ms
Singer 1 (fast-speed song)	(1 song and 794 words)			(2 songs and 1801 words)		
	27.71	56.55	88.54	28.72	59.07	92.44
Singer 2 (normal-speed song)	(1 song and 398 words)			(2 songs and 767 words)		
	38.94	65.08	80.90	41.96	73.12	92.46
Singer 3 (slow-speed song)	(1 song and 172 words)			(2 songs and 412 words)		
	1.14	1.71	4.00	0.00	0.57	1.71

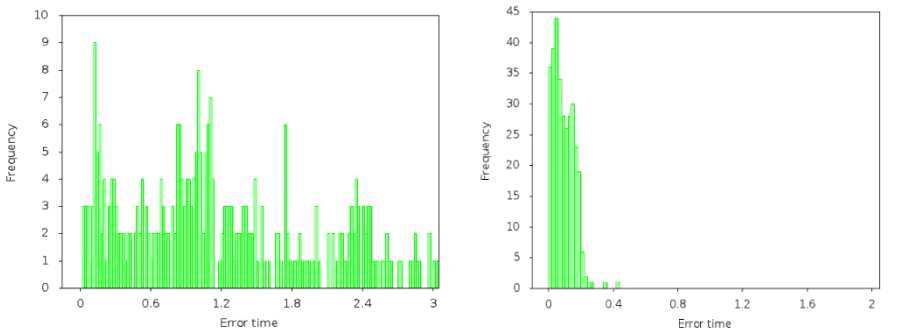


Fig. 3. Comparison of time errors (in seconds) in the cases where there the alignment is performed using only one audio (left), and when two more audios are added as the input data (right).

3.3 Comparison of Methods

Now we compare the results obtained with the model-free method with the results obtained using the method based on pre-existing models. Since the previous results have been obtained with audios with a sampling rate of 44100 KHz, and taking into account that for the method based on pre-existing models it is necessary to work with audios of 8 KHz (due to availability of trained models in our particular case), we need to resample our audios to 8000Hz to compare their alignments in a fair way. Now we will show the results obtained with the method based on pre-existing models, the results obtained with the model-free method with 8000Hz audios, and results obtained also with the model-free method, but with a sampling rate of 44100Hz.

Table 3. Comparison of different methods and sampling frequency for broadcast news. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms). For the model-free method we use always additional audios.

SPEAKER 1 (female)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	28.72	59.57	90.96
	Model-free method (8000Hz)	28.72	48.40	80.32
	Model-free method (44100 Hz)	29.79	50.53	93.09

SPEAKER 2 (female)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	30.46	65.23	86.42
	Model-free method (8000 Hz)	26.49	54.30	95.70
	Model-free method (44100 Hz)	28.81	55.30	96.69

SPEAKER 3 (male)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	34.80	67.40	94.36
	Model-free method (8000Hz)	34.80	56.43	95.92
	Model-free method (44100 Hz)	35.11	57.68	96.87

SPEAKER 4 (male)	Results			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	26.73	55.66	88.05
	Model-free method (8000Hz)	0.00	1.26	7.55
	Model-free method (44100 Hz)	3.46	5.35	8.18

Table 4. Comparison of different methods and sampling frequency in *a capella* songs. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms). For the model-free method we use always additional audios.

Singer 1	Results (*a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	13.98	33.38	75.57
	Model-free method (44100 Hz)	28.72	59.07	92.44

Singer 2	Results (*a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	40.45	61.56	78.14
	Model-free method (44100 Hz)	41.96	73.12	92.46

Singer 3	Results (*a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz)	0.57	1.14	1.14
	Model-free method (44100 Hz)	0.00	0.57	1.71

The results obtained above for the singers are from *a capella* songs. We have performed a comparison of *a capella* songs with those that include instrumental music as well for one particular singer.

Table 5. Comparison of different methods and sampling frequency in songs with music. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms). For the model-free method we use always additional audios.

Singer 2	Results (with music and *a capella*)			
	Tolerance	50 ms	100 ms	200 ms
	Pre-exising models (8000 Hz, with music)	7.04	16.33	44.22
	Model-free method (44100 Hz, with music)	27.64	52.01	80.40
	Model-free method (44100 Hz, *a capella*)	41.96	73.12	92.46

Finally, we perform an analysis on how the number of songs (all with instrumental music) used as input data improves the final result. Again we perform this with only one song.

Table 6. Comparison of results using different number of additional audios (songs) for singer 2. Table shows percentage of words (%) with errors smaller than three values of tolerance (50, 100 and 200 ms).

Singer 2 (44100 Hz, with music)	Tolerance / Number of songs	50 ms	100ms	200ms
	1	25.13	48.99	76.38
	2	27.64	52.01	80.40
	3	26.88	55.03	82.01
	4	28.14	56.28	85.93
	5	30.65	53.77	82.66
	6	29.65	51.76	79.40
	7	32.91	59.80	84.67
	8	31.16	48.47	84.17

4 Discussion

As expected, results in broadcast news are better than those obtained in songs. Results show also (Table 3) that in the case of broadcast news pre-existing models are quite robust even in the case of very slow speech (as in speaker 4). On the other hand, the model-free approach completely fails at aligning very slow speech, while its results for other speakers are similar as those obtained with the pre-existing models method. Therefore it seems that the model-free method is not a good alternative to the pre-existing models method for broadcast news. We must point out that, in order to make the comparison fairer we report results using 8 kHz for both the pre-existing models approach and the model-free method. While using 8 kHz is required (due to the models available in our case) in the pre-existing method, it is not necessary in the

model-free method. Table 3 shows that the model-free method can take advantage of this extended bandwidth yielding results slightly better than the pre-existing models and the model-free methods with limited bandwidth for the three first speakers. We can see (in Table 1 and 3) that the results for the *speaker 4* are very poor due to the slow speech, as mentioned before. In this case, when we add an additional audio, the results get even worse. When we analyzed why we found that in the added audio, there was a small segment of around 20 seconds in which the voice of a different speaker appears. This example points out a real danger that we must deal with in a real-life scenario with the model-free method.

Although our results are still very preliminary, they seem to indicate that, unless the speech to align is very slow (as in the cases of speaker 4 and song 3), the model-free method tends to work better than the method based on pre-existing models in songs and particularly when music is included. Results seem to indicate that, the faster is a song, the better results we obtain. Songs which are very slow have very bad results. In our experiments we have made several experiments with songs: first we have performed the alignment of *a capella* songs, then we have compared with the case in which the instrumental music is included, and finally we have studied to what extent the introduction of additional songs improves the results on audios with instrumental music included.

With respect to the behaviour with songs with different types of audio (*a capella* or not), in the case of *a capella* songs, the model-free method performs better, but it is when we introduce instrumental music, when the difference is more evident in favour of the model-free method. It is logical since the pre-existing models are trained with speech only, while in the model-free method the music and environmental conditions are naturally incorporated during the training process.

Table 6 analyzes how much the introduction of additional audios improves results. Results show that there is a tendency towards improvement of results, however, this tendency is not monotonic and there are maximums and minimums suggesting that some audios may help while others actually decrease performance. In this particular case we find the first maximum (in 100 an 200ms) with four audios, but in other experiments we have found that maximum with only one additional audio.

5 Conclusions and Future Work

One of the main conclusions of the paper is that the use of the model-free method can be an alternative for performing alignments to the method using pre-existing models, particularly in the case of songs. This method is more robust to audio and speaker particularities and could benefit from the possibility of adding more similar data for training. This possibility, however, has some risks that have to be dealt with in the future such as the risk of including speech from other speaker or including songs from the same singer, but very different from the one being aligned. This model-free method is particularly interesting when instrumental music is present in the song to align. One curiosity that we found is that results tended to be better for fast songs than for slow songs, which may be counterintuitive. Our results, however, should be taken

with cares since they are still preliminary and to be more conclusive more experimentation is required.

As future work we would like to deepen our analysis, to extend the experiments including a larger number of songs and news fragments and to find ways to improve the alignment of slow-speed songs and speech in the model free-method. We would also like to extend our study of the influence of the number of songs to be included in the model-free method.

Acknowledgements. This research has been partially supported by the Ministry of Education of Spain under project TEC2009-14719-C02-02 (PriorSpeech) project and by the Regional Government of Madrid under MA2VICMR project.

References

1. Mesaros, A., Virtanen, T.: Automatic Alignment of Music Audio and Lyrics. In: Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx 2008), Espoo, Finland, September 1-4 (2008)
2. Lee, K., Cremer, M.: Segmentation-Based Lyrics-Audio Alignment Using Dynamic Programming. In: Proc. ISMIR, pp. 395–400 (2008)
3. Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T., Okuno, H.G.: Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals. In: Proceedings of the Eighth IEEE International Symposium on Multimedia, ISM 2006 (2006)
4. Meinedo, H., Abad, A., Pellegrini, T., Neto, J., Trancoso, I.: The L2F Broadcast News Speech Recognition System. In: Proc. FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, pp. 93–96 (2010)
5. Ortega, A., Garcia, J., Miguel, A., Lleida, E.: Real-time live broadcast news subtitling system for spanish. In: Proc. Interspeech 2009, Brighton (September 2009)
6. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Povey, D., Valtchev, V., Woodland, P.: The HTK Book, Version 3.4 (March 2009)
7. TIMIT Acoustic-Phonetic Continuous Speech Corpus, LDC Catalog Number LDC93S1, Available through the Linguistic Data Consortium, `http://www.ldc.upenn.edu`
8. CMU Pronouncing Dictionary, `ftp://ftp.cs.cmu.edu/project/speech/dict/` (accessed June 25, 2012)
9. Toledano, D.T., Hernández, L.A., Villarubia Grande, L.: Automatic Phonetic Segmentation. IEEE Transactions on Speech and Audio Processing 11(6) (November 2003)

Aligning Very Long Speech Signals to Bilingual Transcriptions of Parliamentary Sessions*

Germán Bordel, Mikel Penagarikano,
Luis Javier Rodríguez-Fuentes, and María Amparo Varona Fernández

GTTS, Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country UPV/EHU
Barrio Sarriena, 48940 Leioa, Spain
german.bordel@ehu.es
http://gtts.ehu.es

Abstract. In this paper, we describe and analyse the performance of a simple approach to the alignment of very long speech signals to acoustically inaccurate transcriptions, even when two different languages are employed. The alignment algorithm operates on two phonetic sequences, the first one automatically extracted from the speech signal by means of a phone decoder, and the second one obtained from the reference text by means of a multilingual grapheme-to-phoneme transcriber. The proposed algorithm is compared to a widely known state-of-the-art alignment procedure based on word-level speech recognition. We present alignment accuracy results on two different datasets: (1) the 1997 English Hub4 database; and (2) a set of bilingual (Basque/Spanish) parliamentary sessions. In experiments on the Hub4 dataset, the proposed approach provided only slightly worse alignments than those reported for the state-of-the-art alignment procedure, but at a much lower computational cost and requiring much fewer resources. Moreover, if the resource to be aligned includes speech in two or more languages and speakers conmute between them at any time, applying a speech recognizer becomes unfeasible in practice, whereas our approach can be still applied with very competitive performance at no additional cost.

Keywords: speech-to-text alignment, multilingual speech, automatic video subtitling.

1 Introduction

The work presented in this paper was motivated by a contract with the Basque Parliament for subtitling videos of bilingual (Basque/Spanish) plenary sessions. The task consisted of aligning very long (around 3 hours long) audio tracks with syntactically correct but acoustically inaccurate transcriptions (since all the silences, noises, disfluencies, mistakes, etc. had been edited).

* This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06, by the Government of the Basque Country, under program SAIOTEK (project S-PE11UN065), and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially financed by FEDER funds).

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 69–78, 2012.

The above described task may have been easily solved by means of forced alignment at word level, allowing some mismatch between speech and text to cope with imperfect transcripts or alternative pronunciations [1] [2]. However, forced alignment cannot be directly performed on long audio tracks, due to memory bounds. The algorithm proposed in [3] comes to solve this limitation, by applying a speech recognizer, then looking for anchor phrases (sequences of words matching part of the text), splitting the text at such points and recursively applying the same algorithm on the resulting fragments, until their length is small enough to apply forced alignment.

However, the mix of Basque and Spanish in the parliamentary sessions made language and lexical models needed by the speech recognizer difficult to integrate. Therefore, an alternative procedure was developed, which started from a hybrid set of phonetic units covering Basque and Spanish. Acoustic-phonetic models were estimated and a phone decoder was built based on data from both languages. The alignment algorithm operated on two sequences of phonetic units: the first one produced by the phone decoder and the second one obtained by means of a grapheme-to-phoneme transcriber mapping ortographic transcriptions to sequences of hybrid (Basque/Spanish) phonetic units. Finally, time stamps provided by the phone decoder were mapped to ortographic transcriptions through phonetic alignments. This approach worked pretty well for the intended application, as shown in [4].

With the aim to compare our approach to that presented in [3], we carried out a series of experiments on the 1997 English Hub4 dataset (see [5] for details). Following the evaluation criteria proposed in [3], we found that 96% of the word alignments were within 0.5 seconds the true alignments, and 99.2% within 2 seconds the true alignments. In the reference approach [3], better figures are reported (98.5% and 99.75%, respectively) but at a much higher computational cost, and as we noted above, it could not be easily applied to multilingual speech. In this paper, we summarize the above described efforts and devote more space to analyse and discuss the alignment errors, which may light the way to further improvements.

The rest of the paper is organized as follows. In Section 2, we provide the key features of our simple speech-to-text alignment approach. The experimental setup is briefly described in section 3. Results are presented and commented in Section 4. Finally, conclusions are given in Section 5, along with a discussion on possible ways of improving the method.

2 The Speech-to-Text Alignment Method

To synchronize speech and text, we map both streams into a common representation, then align the resulting sequences and relate positions in the original sources by mapping back from the common representation. A suitable candidate for such common representation is the phonetic transcription, which features a small vocabulary size and a small granularity. We assume that phone decoding is performed without any language/phonotactic models, so that the alignment

will be language independent, provided that the set of phonetic units covers all the languages appearing in the speech stream.

2.1 Phone Inventories

In this paper, we consider two different datasets: (1) Hub4, monolingual (English), for which a set of 40 phonetic units was defined, based on the TIMIT database [6]; and (2) plenary sessions of the Basque Parliament, bilingual (Basque/Spanish), for which a set of 27 phonetic units covering both languages was defined (see Table 1). Note that Basque and Spanish share most of their phones, with few differences. We selected 26 units for Basque and 23 units for Spanish, meaning that just one *foreign* sound was added to Basque (θ in IPA coding) and four *foreign* sounds to Spanish (ʃ, ʦ, ʦʲ and sʲ in IPA coding)[1]. Also, though not specified in Table 1, the sounds corresponding to graphemes 'll' in Spanish and 'il' in Basque are assimilated to IPA dʒ.

Phones are represented so that the original ortographic transcriptions can be fully recovered, which is needed at the end of the process. Internally, articulatory codes related to the physiology of the production of each phone are used. Externally, those codes are mapped to IPA codes. Since Basque/Spanish phonetics is very close to its orthography, we also use a highly readable *single-character* specific coding (GTTS-ASCII, see Table 1).

2.2 From Speech to Phones

Audio streams were converted to PCM, 16 kHz, 16 bit/sample. The acoustic features consisted of 12 Mel-Frequency Cepstral Coefficients plus the energy and their first and second order deltas (a common parameterization in speech recognition tasks). Left-to-right monophone continuous Hidden Markov Models, with three looped states and 64 Gaussian mixture components per state, were used as acoustic models.

For the Hub4 experiments, a phone decoder was trained on the TIMIT database [6] and then re-trained on the Wall Street Journal database [7]. The phone decoder yielded error rates in the range 40-60%, depending on the acoustic conditions of the Hub4 subset considered for test (see Figure 1).

Defining a common set of phonetic units covering both Basque and Spanish allowed us to train a single phone decoder to cope with the mixed use of Spanish and Basque in the Basque Parliament. The material used to train the phonetic models was the union of the Albayzin [8] and Aditu [9] databases. Albayzin consists of 6800 read sentences in Spanish from 204 speakers and Aditu consists of 8298 sentences in Basque from 233 speakers. The phone decoder trained this way yielded around 80% phone recognition rate in open-set tests on Albayzin and Aditu, and only above 60% on the Basque Parliament sessions, probably due to acoustic mismatch (background noise, speaker variability, etc.) [4].

[1] Note, however, that the sound θ is pronounced by Basque speakers in words imported from Spanish, and that sounds considered foreign in the central Castilian Spanish (such as ʦ) are widely used in other Spanish dialects.

Table 1. Phone inventory for Basque (Euskera) and Spanish, with examples. IPA codes (Unicode) are shown, as well as a highly readable single-character coding (GTTS-ASCII). Internally, the grapheme-to-phoneme transcriber uses the articulatory codes (*physio codes*) shown in the first column.

Physio CODE	Computational coding		Spanish		Euskera	
	IPA Unicode (HEX)	GTTS ASCII	Orthogr.	Example	Orthogr.	Example
111	i (0069)	i	i	pico	i	ipar
115	u (0075)	u	u	duro	u	umore
132	e (0065)	e	e	pero	e	hemen
135	o (006F)	o	o	toro	o	hori
173	a (0061)	a	a	valle	a	kale
21112	m (006D)	m	m	madre	m	ama
21142	n (006E)	n	n	nunca	n	neska
21172	ɲ (0272)	N	ñ	año	in	arraina
21211	p (0070)	p	p	padre	p	apeza
21212	b (0062)	b	b v	bolsa vino	b	begia
21241	t (0074)	t	t	tomo	t	etorri
21242	d (0064)	d	d	dónde	d	denda
21281	k (006B)	k	c qu k	casa queso kilo	k	ekarri
21282	g (0067)	g	g	gata	g	gaia
21321	f (0066)	f	f	fácil	f	afaria
21331	θ (03B8)	z	c z	cinco paz	--	--
21341	s (0073)	s	s	sala	s	hasi
21351	ʃ (0283)	x	--	--	x	xoxoa
21381	x (0078)	j	j	mujer	j	ijito
21624	r (0072)	R	r rr	rosa torre	rr	arrunta
21742	ɾ (027E)	r	r	puro	r	dirua
21942	l (006C)	l	l	lejos	l	lana
243	tʃ (02A7)	X	ch	mucho	tx	txikia
244	dʒ (02A4)	y	i y	hielo cónyuge	i dd	leoia onddo
24111	ts' (02A6 02BC)	C	--	--	tz	atzo
24122	ts (02A6)	S	--	--	ts	mahatsa
21342	s' (0073 02BC)	c	--	--	z	zoroa

2.3 From Text to Phones

In the Hub4 experiments, phonetic transcriptions were extracted from the CMU English pronouncing dictionary [10]. In the case of Basque parliament sessions, a multilingual transcriber architecture was defined, including a specific transcription module for each target language. Each transcription module consists of a dictionary, a set of transcription rules and a *number-and-symbols to text converter* (for numbers, currencies, percentages, degrees, abbreviations, etc). In this work, two modules for Basque and Spanish were used (including their respective dictionaries), and a third auxiliary module was defined, consisting of a dictionary covering all the words falling out of the vocabulary of both languages.

Phonetic transcriptions are generated as follows. First, each input word is searched in the three available dictionaries: Basque, Spanish and out-of-vocabulary words. If the word appears in a single dictionary, the phonetic transcription provided by that dictionary is output. If the word appears in more than one dictionary, the transcriber uses the context to determine the language being used and outputs the phonetic transcription for that language. Finally, if the word doesn't appear in any dictionary, the transcriber outputs a rule based transcription based on the subsystem corresponding to the most likely language. New transcriptions generated by applying rules are added to the corresponding dictionary and reported to be supervised. This mechanism makes dictionaries to grow incrementally and works as a misspelling detector, allowing for the refinement of rules.

2.4 Alignment of Very Long Sequences

A globally optimal solution to the alignment of two symbol sequences is given by the Needleman-Wunsch algorithm [11]. In this work, we apply a variant of this algorithm. Let X and Y be two sequences of n and m symbols, respectively. A $n \times m$ matrix C is filled with the minimum accumulated edition cost, also known as Levenshtein's distance, using an auxiliary $n \times m$ matrix E to store the edition operations that minimize the cost at each step. Four possible edition operations are considered: deletions, insertions, substitutions and matches. The three first operations have cost 1 and matches have cost 0. Note that what we call *best alignment* depends on the set of weights assigned to deletions, insertions, substitutions and matches. We use the Levenshtein distance because, after some experimentation on the set of parliamentary sessions (which is our target application), it gave the best results. Finally, the path generating the minimum cost is tracked back from $E(n, m)$ to $E(1, 1)$, which defines the optimal mapping (i.e. the optimal alignment) between X and Y.

The above described method is prohibitive for very long sequences due to the matrix memory allocation. However, we can still use a *Divide and Conquer* approach known as Hirschberg algorithm [12], where the original problem is optimally split into two sub-problems with half the size, by doing all the matrix calculations but storing only one row that goes half matrix forward from the start, and one row that goes half the matrix backward from the end. This method is recursively applied until the amount of memory needed to apply the non-recursive approach can be allocated. This algorithm reduces dramatically the required memory, increasing less than 2 times the computation time. Besides, since it can be easily parallelized, it can be run even on a desktop computer (e.g. less than 1 minute for a 3-hour signal in an 8-thread Intel i7 2600 processor).

3 Experimental Setup

3.1 The 1997 Hub4 Dataset

The 1997 Hub4 dataset consists of about 3 hours of transcribed broadcast audio, classified into 6 categories according to acoustic conditions, plus a seventh *Other*

category and a set of *Unclassified* segments. The last two subsets (amounting to 8.5% of the phone units) were not considered in this work. The six remaining categories were defined as follows: F0 (clean speech), F1 (spontaneous speech), F2 (telephone-channel speech), F3 (speech with background music), F4 (degraded speech), and F5 (speech from non-native speakers). Their proportions are shown in Figure 1.

3.2 Basque Parliament Plenary Sessions

We started processing plenary sessions of the Basque Parliament by September 2010. The dataset considered in this work consists of 80 sessions, amounting to 407 hours of video. Due to limitations of video recording media, each video lasts no more than 3 hours, although sessions are 4-8 hours long. Therefore, each session consists of two or three (exceptionally up to four) sections. Videos (originally recorded in high definition using professional media) are converted to RealMedia format for the Basque Parliament web, the audio stream being downsampled to 22050 Hz, 16 bit/sample.

The Session Diary is available almost immediately as a single PDF document, because text transcriptions are produced on the fly by a team of human operators (who are instructed to exclude non-relevant events such as silences, noises, disfluencies, etc.). The session diary is made up of blocks related to operator shifts (approximately 15 minutes per block) with some undefined overlap between them. Also, after each voting procedure in the Basque Parliament, results are not transcribed verbatim as read by the president but just tabulated. All these issues make the synchronization between videos and diaries even more difficult. To address them, we designed specific (in part heuristic, in part supervised) solutions which are out of the scope of this paper.

3.3 Evaluation Measure

Following [3], the alignment accuracy is measured in terms of the deviation of the starting point of each word from the available ground truth. In the case of Hub4, the reference positions were obtained by forced alignment (on a sentence by sentence basis) at the phone level, using acoustic models closely adapted to the HUB4 dataset. In the case of parliamentary sessions, we manually annotated the starting point of each word for a continuous fragment containing 876 words. Finally, to evaluate the alignment accuracy, we provide the percentage of words whose starting point is within a tolerance interval with regard to the reference position.

4 Results

4.1 Results on the Hub4 Dataset

Results on the Hub4 dataset are summarized in Figure 1. As in [3], tolerance intervals of 0.5 and 2 seconds around the reference positions are considered

to measure alignment accuracy. We found that 4.06% of the words deviated more than 0.5 seconds from their reference positions, whereas only 0.82% of the words deviated more than 2 seconds from their reference positions. These figures are slightly worse than those reported in [3] (1.5% and 0.25%, respectively), but the computational savings are quite remarkable, both in terms of time and infrastructure (data, models, etc.).

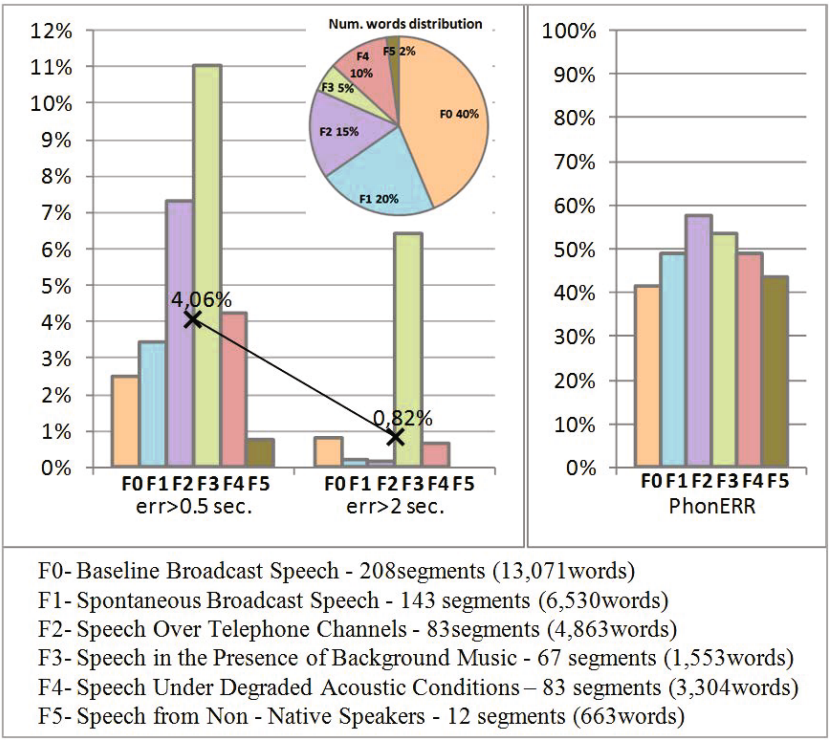


Fig. 1. Results for the 1997 Hub4 dataset: proportions of data in each category, phone decoding error rates and alignment accuracy for tolerance intervals of 0.5 and 2 seconds

As shown in Figure 1, the alignment accuracy strongly depends on the acoustic condition considered for test. On the other hand, the alignment accuracy for a given condition is not only related to the phone recognition accuracy. For instance, the highest alignment error rate was found for the F3 condition (speech with background music), whereas the highest phone recognition error rate was found for the F2 condition (telephone-channel speech). The large difference between the alignment accuracies for the F2 and F3 conditions when considering a 2-second tolerance interval (despite having very similar phone recognition accuracies) is even more significant.

4.2 Results on the Basque Parliament Sessions

As noted above, the alignment accuracy was measured on a continuous fragment of a parliamentary session including 876 words. Table 2 shows the alignment accuracy for different tolerance intervals between 0.1 and 0.5 seconds.

Table 2. Alignment accuracy on a fragment of a session of the Basque Parliament, for different tolerance intervals (in seconds)

Tolerance (seconds)	0.1	0.2	0.3	0.4	0.5
Alignment accuracy	67.69%	88.58%	92.01%	94.41%	95.43%

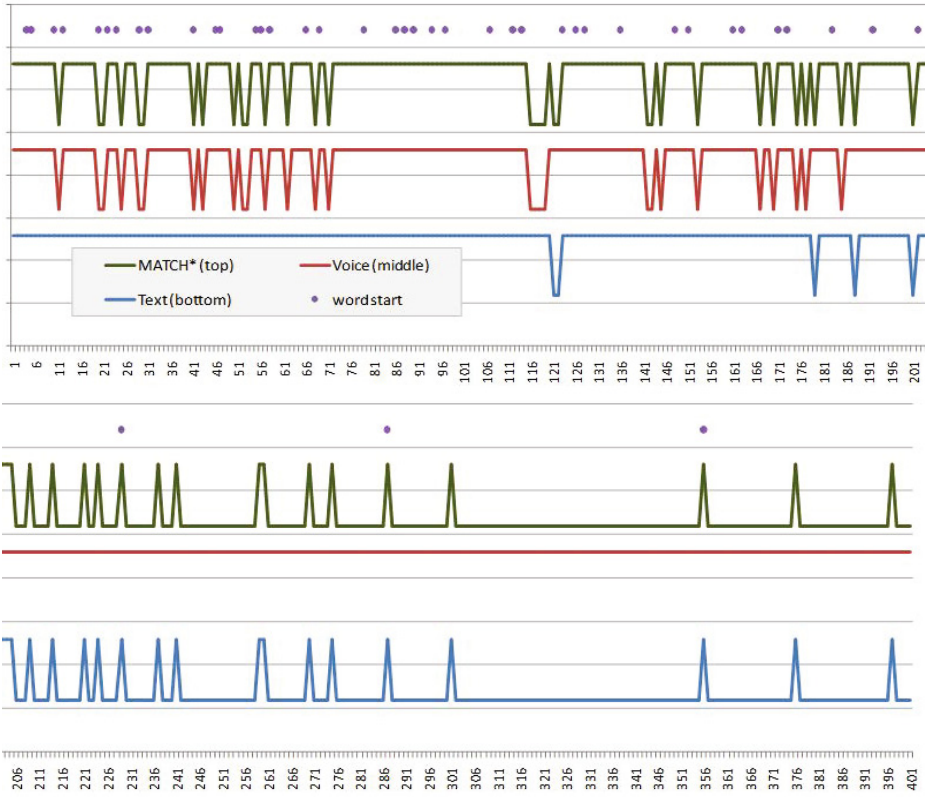


Fig. 2. Patterns found in the alignment path. The bottom line (blue) is up for phones in the text, whereas the middle line (red) is up for phones in the recognized speech. Lines go down at insertions for the bottom line and at deletions for the middle line. The top line (green) represents the AND function of the two other lines, so that it is up for matches and substitutions, and down for insertions and deletions.

Note that 95% of the words deviated less than 0.5 seconds from the reference position considered as ground truth, which is enough for the subtitling application that motivated this work. After aligning speech and text, and following a number of rules related to lengths, times and punctuation, the synchronized text stream is split into small segments suitable for captioning. Only the first word of each segment is taken into account to synchronize text and speech, so that errors are perceived by users as these segments being presented with some advance or delay. Since errors involve both advances and delays in random order, having a long run of advances or delays is quite unlikely. In any case, a deviation of 0.5 seconds is not perceived as an error, specially when the caption appears in advance. For instance, after a long silence, when captions blank, users can easily accept a caption appearing in advance but not a delayed caption. Based on this analysis and taking into account that the captioning procedure can be configured to behave in different ways when there are more than one equivalent partial alignment, we tuned the application so that the mismatched segments had a tendency to show captions before the audio.

5 Conclusions and Future Work

In this paper, we have presented a simple and efficient method to align long speech signals to multilingual transcriptions, taking advantage of a single set of phonetic units covering the sounds of the target languages. We have compared the accuracy of the proposed approach to that of a well-known state-of-the-art alignment procedure, finding a small degradation on the Hub4 dataset, but remarkable savings in both computation time and the required infrastructure. On the other hand, the proposed method can deal with multilingual speech, which is not the case of the state-of-the-art approach used as reference. Alignment results have been also shown for plenary sessions of the Basque Parliament, for which a captioning system has been built based on the proposed algorithm.

Possible ways of increasing the alignment accuracy in future developments include: (1) adapting the acoustic models used in phone decoding to the particular resources to be aligned; and (2) replacing the kernel in the Needleman-Wunsch algorithm (currently representing a Levenshtein distance) with a more informative kernel, e.g. by using continuous weights based on phone confusion probabilities.

Also, by analysing the alignment path, we can search for patterns that may eventually help to automatically reconsider some word synchronizations. Figure 2 represents a section of an alignment path for a Basque parliament session. Two different halves can be identified: the upper half corresponds to a correct alignment, whereas the bottom half corresponds to a wrong alignment due to a missing transcription. In the first half, words are detected at distances that are basically in accordance to phone lengths (upper dots). In the second half, a long run of decoded phones (middle line up) matches to few text phones (bottom line mostly down), meaning that words in the text are sparsely matched to phones from non-transcribed speech. In the first half, we also find that the insertion penalty applied by the phone decoder is too high, since there are much more deletions than insertions in the recognized sequence.

The relation between matches and substitutions and the time span for each word provide key information about the probability of having a perfect match. Places where there is no reference transcription can be detected as long runs of phone insertions, that is, as words spanning in excess through the alignment path. The opposite situation (extra text), which rarely appears in manual transcriptions, would generate long runs of phones in the other axis, that is, a high number of deletions. Both events produce border effects that should be identified and compensated. The *attraction* or *repulsion* that these regions induce on the recognized sequence of phones will depend on the number of deleted or inserted words and will be smoothed by the constraint that both sequences match.

This analysis suggests that, given that most of the alignment is right, we should focus on the problematic areas to isolate the alignment errors and correct the border effects by means of forced alignment. Curiously, this idea is complementary to the algorithm proposed in [3].

References

1. Vonwiller, J., Cleirigh, C., Garsden, H., Kumpf, K., Mountstephens, R., Rogers, I.: The development and application of an accurate and flexible automatic aligner. The International Journal of Speech Technology 1(2), 151–160 (1997)
2. Moreno, P., Alberti, C.: A factor automaton approach for the forced alignment of long speech recordings. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4869–4872 (April 2009)
3. Moreno, P., Joerg, C., Thong, J., Glickman, O.: A recursive algorithm for the forced alignment of very long audio segments. In: Fifth International Conference on Spoken Language Processing (1998)
4. Bordel, G., Nieto, S., Penagarikano, M., Rodriguez Fuentes, L.J., Varona, A.: Automatic subtitling of the Basque Parliament plenary sessions videos. In: Proceedings of Interspeech, pp. 1613–1616 (2011)
5. Bordel, G., Penagarikano, M., Rodriguez Fuentes, L.J., Varona, A.: A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In: Interspeech 2012, Portland (OR), USA, September 9-13 (2012)
6. 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, Philadelphia (1993)
7. Garofolo, J.S., Graff, D., Paul, D., Pallett, D.S.: CSR-I (WSJ0) Complete. Linguistic Data Consortium, Philadelphia (2007)
8. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Marino, J., Nadeu, C.: Albayzin speech database: design of the phonetic corpus. In: Proceedings of Eurospeech, Berlin, Germany, September 22-25, pp. 175–178 (1993)
9. Basque Government, "ADITU program", Initiative to promote the development of speech technologies for the Basque language (2005)
10. Weide, R.: The Carnegie Mellon pronouncing dictionary (cmudict.0.6). Carnegie Mellon University (2005)
11. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)
12. Hirschberg, D.: A linear space algorithm for computing maximal common subsequences. Communications of the ACM 18(6), 341–343 (1975)

Factor Analysis Segmentation and Classification in Broadcast News Domain

Diego Castán, Alfonso Ortega Giménez, and Eduardo Lleida

ViVoLab Aragon Institute of Engineering Research (I3A)
University of Zaragoza
{dcastan,ortega,lleida}@unizar.es
http://www.vivolab.es/

Abstract. This paper proposes a study of a Factor Analysis (FA) segmentation and classification system. Our approach is inspired by language recognition systems where every input sequence is a language. Following this idea, a study between the classic segmentation systems based on HMM/GMM and FA is done over the output of a perfect segmentation system (oracle boundaries). It can be seen how FA improves the classification results compared to HMM/GMM. Also, the first experiments of an on-building FA segmentation system are reported suggesting the need to improve the channel compensation over some classes.

Index Terms: Factor Analysis, Channel Compensation, Broadcast News Segmentation.

1 Introduction

Due to the increase in audio or audiovisual content, it becomes necessary to use automatic tools for different tasks such as analysis, indexation, search and retrieval. Given an audio document, the first step is audio segmentation producing a delineation of a continuous audio stream into acoustically homogeneous regions. When the audio segmentation is followed by a classification system the result is a system that is able to divide an audio file into different predefined classes chosen for a specific task.

Several approaches have been proposed for audio segmentation in different scenarios. For example, in the task of automatic transcriptions of broadcast news [1] the data contain clean speech, telephone speech, music segments and speech overlapped with music and noise so the segmentation generates a boundary for every speaker change and environment/channel condition change with no explicit cues. In [2] segmentation is based on five different classes: silence, music, background sound, pure speech, and non-pure speech. The solution is based on SVM combination. In [3] the audio stream from broadcast news domain is segmented into 5 different types including speech, commercials, environmental sound, physical violence and silence. [4] presents a review of different solutions and the acoustic features used in each one of them and also a new algorithm for computing various time-domain and frequency-domain features, for speech and

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 79–88, 2012.

music signals separately, and estimating the optimal speech/music thresholds. In [5], a system of three components (segmentation, clustering and classification) is used to recognize an entire half hour show with no prior knowledge of acoustic conditions and speakers.

In the context of the Albayzin-2010 evaluation campaign an audio segmentation task was proposed in [6]. Almost all the participants of the evaluation used hierarchical systems, including the winning system [7] based on a hierarchical arquitecture that used different sets of features for every level. For this evaluation database, in [8] we proposed a system that uses a 2-level hierarchical architecture where the second level is based on FA minimizing the segmentation error over this database.

In this paper, a comparison between Factor Analysis and HMM-GMM is reported. The first group of experiments is based on the classification task over the segments with oracle boundaries. In the second group of experiments, the systems must identify the begining and the end of each segment so a segmentation/classification error is reported.

The remainder of the paper is organized as follows: database and metric of Albayzin 2010 evaluation is presented in section 2. Section 3 shows the factor analysis theoretical approach based on language recognition systems. Classifications and segmentation results are presented in section 4. Finally, the conclusions and the future work are presented in section 5.

2 Albayzin 2010 Audio Segmentation Evaluation

2.1 Database

The database used for the Albayzin2010 evaluation consists of a Catalan broadcast news database from the public TV news channel that was recorded by the TALP Research Center from the UPC, and was manually annotated by Verbio Technologies. The database includes approximately 87 hours of annotated audio (24 files of 4 hours long).

Five different audio classes were defined for the evaluation: music(MU), speech(SP), speech with music(SM), speech with noise(SN) and others(OT) but this class is not evaluated in final test. The distribution of the classes within the database is the following: Clean speech: 37%; Music: 5%; Speech over music: 15%; Speech over noise: 40%; Other: 3%.

The database for the evaluation was split into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3).

2.2 Metric

The metric is defined as a relative error averaged over all acoustic classes (ACs):

$$Error = average_i \frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)} \quad (1)$$

where dur($miss_i$) is the total duration of all deletion errors (misses) for the *ith* AC, dur(fa_i) is the total duration of all insertion errors (false alarms) for the *ith* AC, and dur(ref_i) is the total duration of all the *ith* AC instances according to the reference file. The incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A forgiveness collar of 1 sec (both + and -) is not scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an AC begins/ends.

The proposed metric is slightly different from the conventional NIST metric for speaker diarization, where only the total error time is taken into account independently of the acoustic class. Since the distribution of the classes in the database is not uniform, the errors from different classes are weighed differently (depending on the total duration of the class in the database). Therefore the participants have to detect correctly not only the best-represented classes (speech and speech over noise, 77% of total duration), but also the minor classes (like music, 5%). Detection error rates (DET) curves are also provided in the hierarchical segmentation systems for comparison.

3 Factor Analysis Framework

3.1 Statistics

The Factor Analysis approach has been successfully used in speaker recognition [9] and more recently in language recognition [10]. The main advantage of Factor Analysis compared to other classification methods is its ability to compensate for the session variability that can be found in the data due to several factors like background noise, recording devices, etc.

As in language identification, this work examines the problem of assigning a class label to each segment using FA models trying to compensate the within-class variability. Additionally, this task has to deal the with the problem of detecting boundaries between segments of different classes where every segment may have a different length. These segments are going to be mapped to sufficient statistics of fixed size by using a Universal Background Model (UBM) which is a class-independent GMM trained with the EM-algorithm on the feature vectors of the training data. Following the classic terminology of the bibliography, we refer mean-vector and diagonal precision matrix of the UBM as μ_k and P_k where k is the Gaussian component index. All further processing is based only on the statistics, rather than the original feature vectors. Let $P_{ksi} = P(k|\phi_{si})$ denote the posterior probability of UBM component k, given feature vector ϕ_{si}, computed with the standard method for GMM observations, assuming frame-independence. For segment s, with frames indexed $i = 1, 2, ..., N_s$, we define the zero and first-order statistics respectively as:

$$n_{sk} = \sum_{i=1}^{N_s} P_{ksi} \tag{2}$$

$$f_{sk} = \sum_{i=1}^{N_s} P_{ksi} P_k^{1/2} (\phi_{si} - \mu_k) \tag{3}$$

For convenience, we stack the first-order vectors for all components into a single supervector, denoted as f_s. We also center and reduce our statistics relative to the UBM, so that we can assume the UBM as having zero mean and unity precision for all components. After this transformation the formulas below no longer require UBM parameters.

3.2 Channel Compensation

Data from a particular class segment is modeled by a GMM defined by means $m_1, m_2, ..., m_C$, weights $w_1, w_2, ..., w_C$ and covariances $\Sigma_1, \Sigma_2, ..., \Sigma_C$ where C is the number of Gaussians. The Factor Analysis model is the adaptation of the UBM model where the supervector of means is not fixed and it can vary from segment to segment to account for differences in the channel. These GMMs have segment and class dependent component means but fixed component weights and covariances chosen to be equal to the UBM weights and covariances. Specifically, we use a Factor Analysis model for the mean of kth component of the GMM for segment s:

$$m_{sk} = t_{c(s)k} + U_k x_s \tag{4}$$

where $c(s)$ denotes the class of segment s; t_{sk} is the channel independent class location vector; U_k is the factor loading matrix which is the subspace of channel variability and x_s is a vector of L segment-dependent channel factors generated by a normal distribution. Channel factor vector x_s can be seen as the coordinates of the channel dependent class segment vector in the subspace defined by U_k . As in the case of the first-order statistics, we stack component-dependent vectors into supervectors m_s and t_c and we stack the component-dependent U_k matrices into a single tall matrix U, so that equation 4 can be expressed more compactly as:

$$m_s = t_{c(s)} + U x_s \tag{5}$$

where U is known as the channel matrix and it represents the within-class variability. Let $T = [t_{mu}, t_{ot}, t_{sm}, t_{sn}, t_{sp}]$ where T represents the locations of classes in the GMM space, so our metamodel for class-segment-dependent GMM is parametrized by (T, U) which are describing prior distribution of parameters m.

The parameters $\Theta = \{T, U\}$ can be estimated using the EM algorithm iteratively. Data from many segments are used, where the channel factors of each segment is treated as a hidden variable. In the E-step posterior distributions of x are estimated for each segment, using current parameters Θ_{old}. In the M-step we find parameters Θ that maximize the auxiliary function $Q(\Theta, \Theta_{old})$. The simple case is considered where location vectors t_{ck} are obtained by using a single iteration of relevance-MAP adaptation from the UBM. This adaptation is expressed in terms of statistics as:

$$t_{ck} = \frac{\sum_s f_{sk}}{r + \sum_s n_{sk}} \tag{6}$$

where the sums are over all segments s belonging to the class c and r is the relevance factor ($r = 14$ in our experiments). With the class locations fixed, x is re-estimated for each segment s and then U_k for every component k.

Given the channel matrix U and the statistics f_{sk} and n_{sk} for a segment s, a class-independent maximum-a- posteriori(MAP) point-estimate of the channel factors x_s can be performed, relative to the UBM as it can be seen in [10]. This estimate is computed as:

$$\hat{x}_s = \left(I + \sum_k n_{sk} U_k' U_k\right)^{-1} U_k' f_s \tag{7}$$

The effect of the channel factors can be approximately removed from the first-order statistics:

$$\hat{f}_{sk} = f_{sk} - n_{sk} U_k \hat{x}_s \tag{8}$$

where $\hat{f}_{sk}$ is the compensated first-orden statistic.

3.3 Scoring

In [11], different scoring methods are studied. The log-likelihood ratio (LLR) scoring shows a significant speedup without any loss in performance due to the simplification of scoring shown in [9] by omitting non-linear terms. To get the score, the compensated first-orden statistics are used to calculate the class locations :

$$\hat{t}_{ck} = \frac{\sum_s \hat{f}_{sk}}{r + \sum_s n_{sk}} \tag{9}$$

Again, the location supervectors are packed into the columns of a matrix denoted as $\hat{T}$ and thus the score is computed as:

$$\boldsymbol{\lambda}_s = \frac{\hat{T}' \hat{f}_{sk}}{\sum_k n_{sk}} \tag{10}$$

This type of scoring can be seen as a dot product between the compensated test vector and the different class vectors. As a result, a calibration for the dot product is needed. In our approach, a normal distribution $N(\mu, \Sigma)$ (one Gaussian) is trained using the set of scores vector where each class is represented by one dimension of the Gaussian. This Gaussian transforms the general scores to N_s multi-class log-likelihoods where N_s is the number of target classes [12].

4 Experimental Results

4.1 Factor Analysis as a Classifier of Segments with Oracle Boundaries

To evaluate the benefits of using FA, a baseline using the same configuration of the winning system in the Albayzin 2010 evaluation [7] is presented over the output of a perfect segmentation system to be able to evaluate the classification error. This system uses a hierarchical HMM/GMM approach to classify the frames between *MU/NOMU* on the first level, *SM/NOSM* on the second level and *SP/SN* on the last level as it is shown in Fig. 1. The audio features extracted for this system is a combination of 15MFCC + C0 + Δ + $\Delta\Delta$ + 12Chroma. The mean and the standard deviation are computed over a 1 second window with an overlap of 0.5 seconds. Previous experiments showed us that it is better to use less components in the models of the classes with less data (SM and MU) and more Gaussians per HMM state for the classes with more data (SP and SN). Table 1 shows the average error of the four classes for a different number of states of the HMM and a different number of Gaussians. Note that the number of Gaussians for the SP/SN classes is four times greater than that for the MU/SM classes.

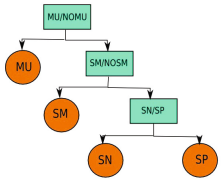


Fig. 1. Block diagraman of the hierarchical system

Table 1. Classification error for oracle boundaries with HMM-GMM systems

Gauss / States	3	4	5	6	7	8	9
32G-128G	26.17	25.87	26.58	25.11	26.40	26.20	28.13
64G-256G	26.30	25.29	**24.56**	25.29	26.34	25.97	27.05

With the same audio features used for the HMM/GMM experiments, a UBM with 1024 Gaussians was trained over all the training set. The channel compensation was performed with 100 channel factors. In Table 2 the average error of the four classes over the training dataset, over the test dataset and over the test dataset with a GBE calibration stage are shown. The first row shows the results of the FA over the smoothed features with mean and standard deviation. The use of the mean and the standard deviation for FA seems to be not a good

Table 2. Classification error for oracle boundaries with FA systems using MFCCs and Chroma features

	TRAIN	TEST
With Mean-Std - UBM 1024G - 100ChnF	3.24	55.21
Without Mean-Std - UBM 1024G - 100ChnF	14.77	**22.91**

solution based on the results of Table 2 where a good accuracy over the training segments can be seen but a poor generalization over the test segments.

A very important issue for the channel compensation task is the early fusion of various types of features. According to the results shown in Table 3 stacking features seems not to be the best solution if we compare these figures with the results shown in Table 2.

Table 3. Classification error for oracle boundaries with FA systems using MFCCs

	TEST
MFCC16+Δ+$\Delta\Delta$ - UBM 1024G - 100ChnF	21.25
MFCC16+Δ+$\Delta\Delta$ - UBM 2048G - 100ChnF	**20.81**

Fig. 2 shows the error per class and the average error for those systems that have better results in each table. The effectiveness of the HMM/GMM system as a music detector compared with FA systems is evident. A possible explanation for this behavior is that the U matrix was trained for all the classes and the most important channel effect is the speech class because it represents 92% of the database so when channel compensation is applied over the music segments, a distortion is produced. On the other hand, the behavior of the FA in SN and SP classes is much better than HMM/GMM.

4.2 Factor Analysis as a Segmentation System

Using the same HMM/GMM hierarchical system with the same audio features that were used in the previous subsection, the error segmentation for different number of states were calculated and the results are presented in Table 4. In this case, it is clear that the best configuration for the segmentation task is with 8 HMM states instead of 6 states in the case of the classification task.

For the FA segmentation system we use the MFCC16+Δ+$\Delta\Delta$ audio features with 2048 Gaussians to train the UBM and 100 channel factor to model the channel compensation. As it can be seen, this system was used in the previous subsection yielding the best results in the classification task. The segmentation is produced with the classification of 3 seconds segments with an overlap of 1.5

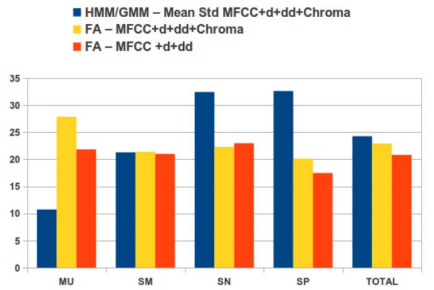


Fig. 2. Comparison of error rate per class between HMM/GMM and FA systems

Table 4. Segmentation error with HMM-GMM systems

Gauss / States	3	4	5	6	7	8	9
32G-128G	39.97	34.76	32.88	31.47	30.85	30.43	31.31
64G-256G	39.27	33.66	31.11	30.91	30.99	**29.37**	31.59

seconds. In this case the transition probabilities between segments are not used so there is no contextual information at all. With this framework, the difference between the classification error and the segmentation error in the worst case (MU-NOMU because we have more error rate) is evident and is shown in Fig. 3(a) where DET curves for oracle and non-oracle segmentation are compared. In Fig. 3(b) the DET curves for every branch of the hierarchical system are plotted.

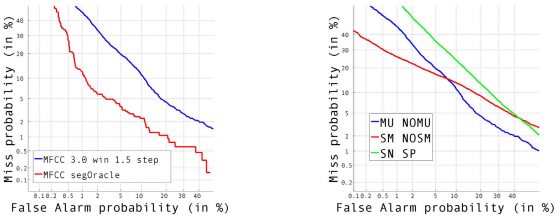


Fig. 3. (a) DET curves for oracle boundaries vs non-perfect segmentation in the music hierarchical level and (b) DET curves for every level of the hierarchical system

Table 5 shows the results plotted in Fig. 3(b) with the evaluation metric for every class. The error for classes like MU or SM is still very high compared to

the HMM/GMM error rate. Nevertheless, Fig. 4 shows the DET curves divided by the length of each segment. It can be seen that for long segments, the GMM is the best classifier but for short segments, FA system is much better than the GMM.

Table 5. Segmentation error per class for the best HMM/GMM system and FA system

	MU	SM	SN	SP	TOTAL
HMM/GMM-8states	15.93	23.43	38.66	39.48	29.37
Hierarchical FA	52.91	37.19	45.08	40.80	43.99

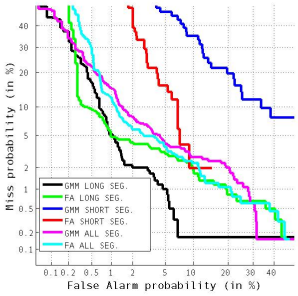


Fig. 4. DET curves GMM vs FA with different length of segments

5 Conclusion and Future Work

By means of classification experiments it has been shown that channel compensation helps to classify segments decreasing the error rate and improving the classification of all speech classes. These results justify the creation of a whole-FA segmentation system following the same hierarchical structure used for HMM/GMM. Although the segmentation error is very high, the better classification in short segments encourages to improve the system.

For future work, different window lengths and time advances will be implemented to try to improve the segmentation. Also, other scoring methods different than the linear scoring will be studied like those presented in [11]. In addition, the class dependent training of several U matrices will be investigated creating a new U matrix by stacking the different class dependent U matrices to decrease the error in the MU class which is critical for the metric of the evaluation.

References

[1] Chen, S.S., Gopalakrishnan, P.S.: IBM L CSR System for Transcription of Broadcast News Used in the 1997 Hub4 English Evaluation. In: Proceedings of the Speech Recognition Workshop (1998)

[2] Lu, L., Zhang, H.-J., Li, S.Z.: Content-based audio classification and segmentation by using support vector machines. Multimedia Systems (2003)
[3] Nwe, T.L., Li, H.: Broadcast news segmentation by audio type analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), vol. 2. IEEE (2005)
[4] Lavner, Y., Ruinskiy, D.A.: Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation. EURASIP Journal on Audio, Speech, and Music Processing (2009)
[5] Siegler, M.A., Jain, U., Raj, B., Stern, R.M.: Automatic Segmentation, Classification and Clustering of Broadcast News Audio. Signal Processing, 4–6
[6] Butko, T., Nadeu, C., Schulz, H.: Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results. Evaluation (2010)
[7] Gallardo, A., San Segundo, R.: UPM-UC3M system for music and speech segmentation. In: Proc. of FALA (2010)
[8] Castan, D., Vaquero, C., Ortega, A., Martinez, D., Lleida, E.: Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain. In: Interspeech (2011)
[9] Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. IEEE Transactions on Audio, Speech and Language Processing (2007)
[10] Brummer, N., Strasheim, A., Hubeika, V., et al.: Discriminative acoustic language recognition via channel-compensated GMM statistics. In: Tenth Annual Conference of the International Speech Communication Association (2009)
[11] Glembek, O., Burget, L., Dehak, N., Brummer, N., Kenny, P.: Comparison of scoring methods used in speaker recognition with Joint Factor Analysis. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (2009)
[12] Benzeghiba, M.F., Gauvain, J.-L., Lamel, L.: Cnrs L: Cedex BPO. Language Score Calibration using Adapted Gaussian Back-end. In: Interspeech (2009)

Prosodic and Phonetic Features for Speaking Styles Classification and Detection

Arlindo Veiga[1,2], Dirce Celorico[2], Jorge Proença[2],
Sara Candeias[2], and Fernando Perdigão[1,2]

[1] Electrical and Computer Eng. Department, University of Coimbra, Portugal
[2] Instituto de Telecomunicações - Pole of Coimbra, Coimbra, Portugal
{aveiga,dircelorico,jproenca,saracandeias,fp}@co.it.pt

Abstract. This study presents an approach to the task of automatically classifying and detecting speaking styles. The detection of speaking styles is useful for the segmentation of multimedia data into consistent parts and has important applications, such as identifying speech segments to train acoustic models for speech recognition. In this work the database consists of daily news broadcasts in Portuguese television, on which two main speaking styles are evident: read speech from voice-over and anchors, and spontaneous speech from interviews and commentaries. Using a combination of phonetic and prosodic features we can separate these two speaking styles with a good accuracy (93.7% read, 69.5% spontaneous). This is performed in two steps. The first step separates the speech segments from the non-speech audio segments and the second step classifies read versus spontaneous speaking style. The use of phonetic and prosodic features provides alternative information that leads to an improvement of the classification and detection task.

Keywords: speaking styles, phonetic and prosodic features, speech classification, event detection.

1 Introduction

A considerable worldwide mobilization of efforts has been emerging in order to develop speech technology solutions. In a world where access to information is mainly acquired through multimedia services, the emergence of speech technologies and the lack of solutions based on such products lead to an increase in the need to generate segmented speech corpora.

Manual segmentation, even when carried out by linguistic and annotator experts, has many disadvantages, such as the amount of time spent, the lack of unanimously conventional criteria, the susceptibility to incoherence and to human errors. Automatic detection of speaking styles for segmentation purposes of multimedia data is one of the goals of the researchers in automatic speech processing, aiming to find a way of obtaining data in a more cost-effective way. The analysis and characterization of the speaking styles with a reliable feature set is also itself a topic for research in the

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 89–98, 2012.

speech communication domain. Sociolinguistic and psycholinguistic studies, such as [1] and [2] were pioneers in that direction.

Before starting to analyze any speech corpora in terms of style, a definition of what a speaking style is must be taken into account. There are some studies, as [3], which have defined different axes to better capture the nature of speaking styles. However, the concept of speaking style is not a closed definition and terms such as 'clear speech', 'slow' or 'fast speech', 'spontaneous', 'informal' or 'casual speech', 'planned' or 'read speech', among others have been used and defined in many ways, almost as numerous as the authors who have dealt with the subject (see an overview in [4]). A closer glance in the literature also shows that there has not been one single specific feature capable of characterizing changes in speaking styles. On the one hand, acoustic characteristics were used by [5] in the context of the automatic speech recognition to differentiate spontaneous and read speech. On the other hand, studies such as [6] have hypothesized that prosody could be intertwined with phonetics for a better comprehension of the speech structure, as well as an improvement of the speech style classification, [7] and [8]. For Portuguese, we can mention works that had attempted to provide evidence of the hesitation events [9], [10] or the degree of relative hypo articulation of the surface forms [11], [12], in the continuous speech. Another study, [13], takes evidence of the rhythm of the speech to compare European and Brazilian Portuguese speaking styles.

With this work we aim to distinguish the two most evident speaking styles in broadcast news: spontaneous speech and read speech. For this task we decided to explore the combination of a set of phonetic and prosodic features. For the same purpose, we also intend to further explore the possibility of extending the performance of this set of features with previous results concerning the characterization of the hesitation events [14].

In section 2 we briefly describe the data source used in our experiments. Section 3 provides information about methodology procedure and section 4 presents the results. Final conclusions are drawn in section 5.

2 Corpus Characterization

Audio signal extracted from broadcast news (BN) of a Portuguese television channel podcast was used for training, test and evaluation purposes. The audio was downsampled from 44.1 kHz to 16 kHz sampling rate and the video information is discarded. A total of 30 daily news programs were considered for this study, with a total duration of about 27 hours. The sound material contains studio and out of studio recordings, as well as sessions recorded from the telephone. Utterances by anchors and professional speakers, commentators, reporters, interviewers and interviewees are present in the audio. Prepared (reading) speaking style is dominant but, most of the time, speech is over background speech, noise or music. There are also non-speech events like music, jingles, laughter, coughing or clapping.

All the audio has been carefully examined and annotated manually, using the Transcriber software tool [15]. Four levels of annotation were considered. At the signal level

(first level), the labels are 'speech', 'music', 'jingle', 'noise', 'cough', 'laugh', 'claps', etc. The acoustical environment is the second level with the following labels: 'clean', 'music', 'stationary noise', 'speech overlapping', 'crowd noise', 'mixed-' or 'indistinct' background. In this level speech over a telephone system was also annotated. The speaking style is identified in the third level, labeled as 'Lombard', 'read' and 'spontaneous' speech. Spontaneous speech is still differentiated into low-, middle- and high-spontaneity, taking into consideration the occurrence of several fluency events such as hesitations. The speaker information corresponds to the fourth level, in which the speaker is identified whenever possible. The occurrence of foreign languages in the signal was also included in the labels. Table 1 shows all annotation levels. Only speech segments were annotated with levels 2, 3 and 4. All the non-speech annotations are grouped into a single label for the speech/non-speech classification.

Table 1. Corpus annotation levels

1 - Signal Level	2 – Acoustical Env.	3 – Speech Style	4 - Speaker
Speech	Clean (no noise)	Prepared	Anchor1
Silence	Music overlap	Lombard	Anchor2
Music	Speech overlap	Unprepared (High)	Journalist(M/F)
Jingle	Stationary noise	Unprepared (Average)	Male
Noise	Crowd noise	Unprepared (Low)	Female
Clapping, etc.	Mixed background		VIP(1,2,3…)
	Indistinct background		
	Telephone		

3 Methods

The proposed method of speech styles classification and detection combines a set of phonetic and prosodic features in order to improve their automatic classification and detection. The classification task uses acoustic signal and manual segments and the detection task uses only acoustic signal. The detection task performs automatic segmentation based on the Bayesian Information Criterion (BIC) [16] and uses the same classifiers as the classification task.

The phonetic features are based on the duration of the phones automatically recognized from the segment signal. The phone recognition system used is based on hidden Markov models using phone acoustic models and a simple bigram model that constrains the allowable recognized phone sequence. Several phonetic features are computed using the phone durations and the recognition likelihood. The size of parameter vector is 214, corresponding to 5 statistical duration measures (mean, median, standard variation, maximum and minimum) and likelihood for each of the 35 phones (6x35=210) plus speaking and silent rates and durations.

The prosodic features are based on the pitch (F0) and harmonic to noise ratio (HNR), acquired by the Praat tool [16]. The features for each speaking-style segment consist of first and second order statistics of the F0/HNR envelope in every voiced portion of the segment as well as the parameters of a polynomial fit of order 1 and 2

of that envelope. Other measures, such as F0/HNR reset rate (rate of voiced portions), speaking and silence duration rates, were also taken. In total the resulting prosodic parameter vector has 108 features. Consequently the vector representing the combination has a total of 322 features.

3.1 Classification

Two Support Vectors Machines (SVM) classifiers are trained: one for speech/non-speech classification and one to distinguish the two speaking styles. The evaluation process uses the k-fold cross-validation paradigm (k=5). The results from the folds can be averaged to produce a single estimation.

The SVM was trained with Sequential Minimal Optimization (SMO) [18] and for each classifier we chose the complexity parameter C in the SMO algorithm that achieves the best-weighted average value given in cross validation. In order to implement this classification analysis we use the software WEKA (Waikato Environment for Knowledge Analysis, [19]) which is an open-source machine-learning tool, commonly used to train and test SVM classifiers.

In the broadcast news corpora there are different background environments that can cover a wide range of noises and events that can be classified as noise. Also the presence of jingles, music, claps, and similar occurrences increase the difficulty of a correct automatic classification of audio segments. Thus, our approach is based on a two-step classification. Given an audio segment, we first classify it as non-speech or as a speech segment, regardless of the style and environment type. The second step consists of a classification that is applied only for the speech audio segments. This double step classification is an obvious choice since SVMs are binary classifiers. This allows a more precise speaking style classification since, hopefully, there are only speech segments to classify in this step.

3.2 Automatic Segmentation

Automatic detection task requires an automatic segmentation process followed by the classification.

The acoustic segmentation method is based on the modified Bayesian Information Criterion (BIC) approach where the symmetric Kullback-Leibler distance is used in the first step to compute acoustic dissimilarity and the BIC measure is used on the second step to validate the detected segment boundaries. This approach was presented on [16] and is identified as distBIC. Sliding two consecutives windows with fixed size, and modeling each window by a Gaussian distribution, the distance between the two models is computed. High acoustic dissimilarity between windows results in a high distance value; so, the segment boundary is computed by finding significant local maximum on distance value. A threshold with hysteresis between the local maximum and enclosed local minimums is used to determine the significance of the local maximum. The threshold value is based on standard deviation of distance.

The second step validates the segment boundaries by computing the delta BIC. In this step, three Gaussian distributions are used for each boundary candidate: one to

model the left windows, other to model right windows and another to model the acoustic data in both windows. A negative delta BIC value means that it is better to model left and right windows as two Gaussian distributions instead of one.

In the automatic segmentation implementation, 16 Mel-frequency cepstral coefficients (MFCCs) plus logarithm of energy are used as acoustic features. Before the distBIC process, energy information is used to mark low energy segments with duration above 0.5 seconds. A threshold of 0.6 of distance standard deviation was used to select significant local maximum. The window size was 2 seconds and the sliding step was 0.1 second. These values are chosen empirically to maximize the segmentation accuracy.

4 Results and Discussion

The evaluations are carried out on two steps classification approach using phonetic and prosodic features as well as a combination of both features.

Results with classification using manual segmentation are presented first. This task illustrates the performance of the trained classifiers. Two measures are presented to evaluate the detection task performance: the segmentation accuracy and F1-score; and the agreement time between references and classified labels. Table 2 shows statistics for the type of segments used for this evaluation.

Table 2. Segment count and duration statistics

Type of segment	Number of segments	Average duration (± std deviation) (s)
Speech	7971	11.0 (± 9.4)
Non-Speech	2529	4.1 (± 5.3)
Read Speech	4989	10.6 (± 8.5)
Spontaneous Speech	1738	12.0 (± 10.4)

4.1 Speech/Non-speech Classification

The first classification stage aims to distinguish speech segments from the remaining audio-segmented events. In the database the events considered as non-speech have a total duration of approximately three hours. Using the method described above we have obtained the values presented in Table 3.

Table 3. Speech/Non-speech classification results. “Acc” stands for accuracy.

Type of features	Acc.	Speech	Non-speech
Phonetic	93.8%	96.7%	82.0%
Prosodic	93.8%	97.5%	81.9%
Combination	94.4%	97.6%	84.0%

The speech and non-speech groups have distinctive audio characteristics that manifest in distinct phonetic and prosodic characteristics. This may explain why the classification performance presents similar values for the two sets of features. Some of the misclassified events observed are due to the fact that some audio segments labeled as speech have intense music or noise background leading to a classification as non-speech segment. However, the association of phonetic and prosodic features leads to a better performance, showing the advantage of their combination.

4.2 Read/Spontaneous Classification

For the second step in the classification task we only considered the audio segments that were labeled as speech. In this particular group a small set of speech segments (Lombard and spontaneous-low speech) were not included because they lack characteristics that can clearly fit them in the read/spontaneous distinction. Results in the prediction of correctly classified instances are shown in table 4.

Table 4. Read/Spontaneous speech classification results. "Acc" stands for accuracy.

Type of features	Acc.	Read	Spontaneous
Phonetic	83.2%	92.8%	55.4%
Prosodic	86.4%	95.0%	61.6%
Combination	87.4%	93.7%	69.5%

Table 4 shows that the prosodic features resulted in better classifications over the phonetic ones, and that the combination approach provided a significant increase in spontaneous speech detection, leading to a slightly global improvement.

4.3 Detection Performance

The detection task is carried out by two processes: automatic segmentation and classification. The classification is performed similarly as above but using the new segmentation and a different performance metric.

4.3.1 Automatic Segmentation Performance

The automatic segmentation is evaluated using the manual segmentation as reference mark and the performance is based on accuracy and F1-score. The accuracy shows the rate of correctly detected reference marks. A detected mark is assigned as correct if there is one reference mark inside a predefined threshold collar. Figure 1 shows the accuracies with collars from 0.5 to 2.0 seconds.

In the segmentation procedure the two types of errors that can occur are misses - where a reference mark was not detected - and insertions - where a detected mark did not have a corresponding reference mark. Since the accuracy only relates to the miss error and the recall to the insertion error, the use of the F1-score that combines accuracy and recall can be used to show a global system performance (Fig. 2).

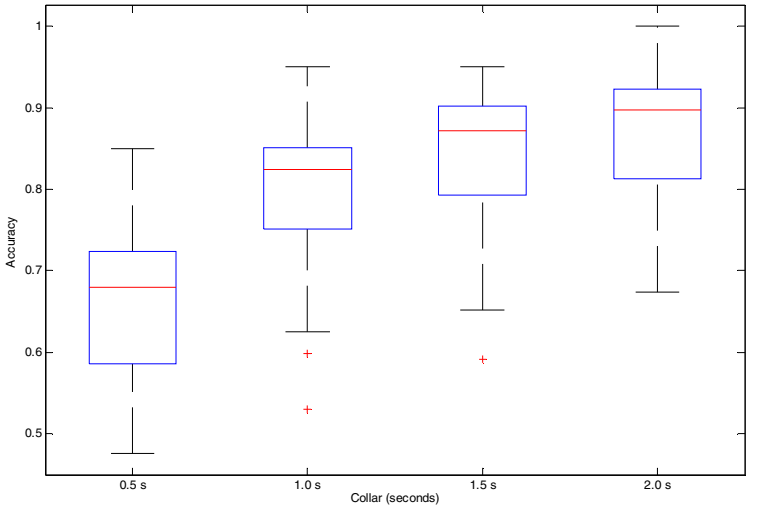


Fig. 1. Segmentation performance (Accuracy) for the audio signals using 4 collar values

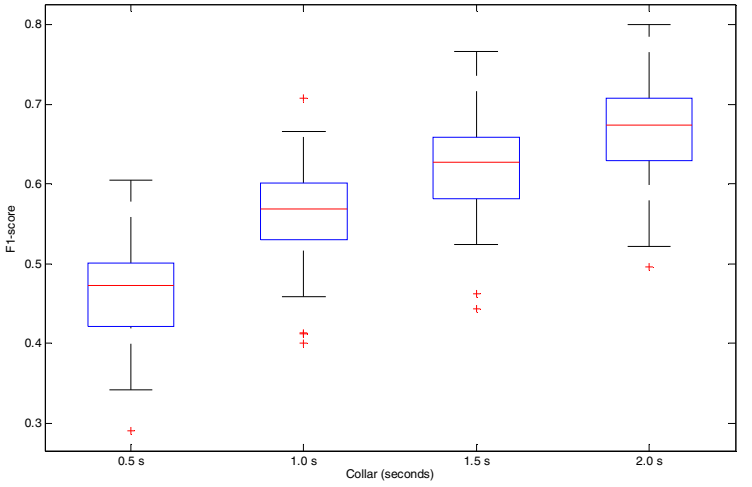


Fig. 2. Segmentation performance (F1-score) for the audio signals using 4 collar values

It is harder to recover from a miss error than an insertion error. Therefore, the automatic segmentation was tuned in order to minimize miss errors at the cost of insertion ones, causing lower F1-score values.

4.3.2 Classification Performance

The classification for automatic detection performance is based on agreement time between reference label and automatic classified label.

Table 5. Speech/Non-speech detection results. "AT" stands for agreement time.

Type of features	AT.	Speech (87772 sec.)	Non-speech (10287 sec.)
Phonetic	91.5%	94.9%	62.2%
Prosodic	93.2%	97.0%	61.0%
Combination	93.3%	96.6%	64.9%

The segments classified as speech on the first step (table 5) are used as input for the second classifier. The final results are presented in table 6.

Table 6. Read/Spontaneous speech detection results. "AT" stands for agreement time.

Type of features	AT.	Read (52622 sec.)	Spontaneous (20917 sec.)
Phonetic	76.7%	91.9%	38.6%
Prosodic	81.1%	93.0%	51.2%
Combination	83.3%	92.7%	59.6%

Similarly to the classification procedure, the combination of features leads to the best detection performance.

4.4 Final Remarks

The above results are based on only a few and simple measures of the speech signal: statistics of phone durations and likelihoods and F0/HNR statistics. It is significant that such few measures could provide these reasonable results.

The results suggest that the F0 envelopes and phone durations should have consistent patterns that could differentiate audio segments. In fact, we have also tried to separate the audio segments into other classes, for instance, Jingle and Music (representing approximately 40% of the non-speech segments) versus other kind of segments. We achieved a detection of 76% of Jingle/Music, with 99% accuracy on other segments. The main characteristics of Music or Jingles are the long F0 patterns and probably, the long recognized phones.

The detection of different acoustic speech environments was also examined but the results were poor (only 60% of global accuracy) which demonstrate that the considered features are not appropriate for this classification. For this acoustic environment detection, GMMs (Gaussian Mixture Models), for example, could lead to much better results [21].

5 Conclusions

The result of this study provides a practical example, using broadcast news, of the importance of considering both prosody and phonetics in the characterization of speech in terms of its structure and style. Under the term of speaking styles, spontaneous and read speech were characterized automatically. Prosodic features were the best in classifying the two styles, but a combination of phonetic- and prosodic-based

classification provided even better results, and therefore both seem to have important and alternative information. The use of both phonetic and prosodic features together has been already explored by the authors for the detection of hesitation events (fillers) from the speech signal [9].

Using an approach through automatic audio segmentation based on BIC we also obtained reasonable results. This encourages our long-term objective: to automatically segment all audio genres and speaking styles. In other works we have already implemented several important features, such as hesitations detection [14], aspiration detection using word spot techniques, speaker identification using GMM [20] and jingle detection based on audio fingerprint [22] that can be incorporated to achieve this goal.

Furthermore, we believe that the study of the relevance of each SVM feature (ranking) is important, allowing to consider a smaller set of features that will perform classification with the same level of accuracy but with a reduction in time and resources.

The identification of both prosodic and phonetic features to characterize different speaking styles in children's speech is another intended extension of this study.

Acknowledgements. This work is funded by FCT and QREN projects (PTDC/CLE-LIN/11 2411/2009; TICE.Healty13842) and partially supported by FCT (Instituto de Telecomunicações multiannual funding PEst-OE/EEI/LA0008/2011). Sara Candeias is supported by the FCT grant SFRH/BPD/36584/2007. We would also like to thank RTP (Rádio e Televisão de Portugal) for allowing us to use their multimedia files.

References

1. Labov, W.: Sociolinguistic Patterns. University of Pennsylvania Press (1973)
2. Goldman-Eisler, F.: Psycholinguistics: experiments in spontaneous speech. Academic Press, London (1968)
3. Eskenazi, M.: Trends in speaking styles research. In: EUROSPEECH 1993, PP. 501–509, Berlin (1993)
4. Llisterri, J.: Speaking styles in speech research. In: ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language, Dublin, Ireland (1992)
5. Nakamura, M., Iwano, K., Furui, S.: Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Computer Speech and Language 22, 171–184 (2008)
6. Deshmukh, O.D., Kandhway, K., Verma, A., Audhkhasi, K.: Automatic evaluation of spoken English fluency. In: Proc. of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009), Taipei, Taiwan, pp. 4829–4832 (2009)
7. Biadsy, F., Hirschberg, J.: Using Prosody and Phonotactics in Arabic Dialect Identification. In: Proc. of Interspeech 2009, Brighton, UK (2009)
8. Sanchez, M.H., Vergyri, D., Ferrer, L., Richey, C., Garcia, P., Knoth, B., Jarrold, W.: Using prosodic and spectral features in detecting depression in elderly males. In: Proc. of Interspeech, Florence, Italy, pp. 3001–3004 (2011)
9. Veiga, A., Candeias, S., Lopes, C., Perdigão, F.: Characterization of hesitations using acoustic models. In: Proc. of the 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, pp. 2054–2057 (2011)

10. Moniz, H., Trancoso, I., Mata, A.: Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In: Proc. of Interspeech 2009, Brighton, UK, pp. 1719–1722 (2009)
11. Braga, D., Freitas, D., Teixeira, J.P., Barros, M.J., Latsh, V.: Back Close Non-Syllabic Vowel [u] Behavior in European Portuguese: Reduction or Suppression. In: Proc. of ICSP 2001 (International Conference in Speech Processing), Seoul (2001)
12. Candeias, S., Perdigão, F.: A realização do schwa no Português Europeu. In: Proc. of the II Workshop on Portuguese Description-JDP, 8th Symposium in Information and Human Language Technology (STIL 2011), Cuiabá, Mato Grosso, Brasil (2011)
13. Barbosa, P., Viana, M., Trancoso, I.: Cross-variety Rhythm Typology in Portuguese. In: Proc. of Interspeech 2009, Brighton, UK (2009)
14. Veiga, A., Candeias, S., Celorico, D., Proença, J., Perdigão, F.: Towards Automatic Classification of Speech Styles. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F., et al. (eds.) PROPOR 2012. LNCS (LNAI), vol. 7243, pp. 421–426. Springer, Heidelberg (2012)
15. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In: Proc. of the First International Conference on Language Resources and Evaluation (LREC), pp. 1373–1376 (1998)
16. Delacourt, P., Wellekens, C.J.: DISTBIC: A speaker-based segmentation for audio data indexing. Speech Communication 32, 111–126 (2000)
17. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (Version 5.1.05), Computer program (retrieved May 1, 2009)
18. Platt, J.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft Research, MSRTR-98-14 (1998)
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 (2009)
20. Reynold, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
21. Akbacak, M., Hansen, J.H.L.: Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems. IEEE Transactions on Audio, Speech, and Language Processing 15(2), 465–477 (2007)
22. Lopes, C., Veiga, A., Perdigão, F.: Using Fingerprinting to Aid Audio Segmentation. In: Proc. of the VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, FALA 2010, Vigo (2010)

Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit

David Martínez, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, and Jesús Villalba

Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{david,lleida,ortega,amiguel,villalba}@unizar.es

Abstract. The paper presents a set of experiments on pathological voice detection over the Saarbrücken Voice Database (SVD) by using the MultiFocal toolkit for a discriminative calibration and fusion. The SVD is freely available online containing a collection of voice recordings of different pathologies, including both functional and organic. A generative Gaussian mixture model trained with mel-frequency cepstral coefficients, harmonics-to-noise ratio, normalized noise energy and glottal-to-noise excitation ratio, is used as classifier. Scores are calibrated to increase performance at the desired operating point. Finally, the fusion of different recordings for each speaker, in which vowels /a/, /i/ and /u/ are pronounced with normal, low, high, and low-high-low intonations, offers a great increase in the performance. Results are compared with the Massachusetts Eye and Ear Infirmary (MEEI) database, which makes possible to see that SVD is much more challenging.

Keywords: Pathological Voice Detection, Saarbrücken Voice Database, GMM, Fusion, MultiFocal toolkit.

The detection of laryngeal pathologies through an automatic voice analysis is one of the most promising tools for speech therapists, mainly due to its noninvasive nature and its objectivity for making decisions. The performance of these systems is nevertheless not perfect, and nowadays it is used as an additional source of information for other laryngoscopial exams [1].

Researchers have focused their efforts on finding new features that could discriminate between normal and pathological voices or even assess their quality, but also on finding different approaches for classification. Some of the most useful features are considered to be acoustic parameters such as mel-frequency cepstral coefficients (MFCC) [17, 1], amplitude and frequency perturbation parameters [9], and noise related parameters [2, 6], but there are different alternatives like nonlinear analysis [3, 4]. Regarding to the classifiers, well-known approaches in speech processing like hidden Markov models (HMM) [7], Gaussian mixture models (GMM) [1], multilayer perceptrons (MLP) [8], or support vector machines (SVM) [9], have been studied.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 99–109, 2012.

Most of the works in the literature make use of the MEEI database, openly commercialized by *Kay Elemetrics* [12]. In other cases, private databases collected in local hospitals are the alternative. However, recently a new open and freely downloadable database, the SVD [13], has been recorded by the Institute of Phonetics of Saarland University. On it, sustained /a/, /i/ and /u/ vowels, pronounced with normal, low, high and low-high-low intonations, and a spoken sentence in German, are found, what make of this database a very complete set to conduct experiments, and easy to reach by all the community. No previous results for voice pathology detection have been found on it, and with this work we also aim at proposing a baseline for future research.

Experiments related with pathological voices can be focused on three main tasks. While the most simple and direct idea is to classify voices as pathological or normal, like in [2, 6, 7, 10, 11], another goal is to assess voice quality according to a perceptual scale, like GRBAS [24, 25], DSI [27, 28] or VPA [26], among others. A third typical problem is to identify a specific pathology, like for example, functional dysphonia in [28], nodules and other laryngeal injuries in [9], or polyps, keratosis leukoplakia, adductor spasmodic dysphonia and vocal nodules in [5].

The work developed in [11] explores different configurations for a GMM classifier to detect pathological voices, fed with MFCC, harmonics-to-noise ratio (HNR) [14], normalized noise energy (NNE) [15] and glottal-to-noise excitation ratio (GNE) [16]. The performance is tested on the MEEI database, and only files with recordings of vowel /ah/ sustained are used. A 30-fold strategy is followed and several random partitions are created to average results. The best performance is obtained with 3 Gaussians and 16 MFCCs, and the area under the curve (AUC) [18] is 0.98459, with the 95.49% of the pathological files and 90.70% of the normal files correctly classified. In addition, the same study is performed in different environments, such as MP3 compression and telephone channel distortion.

In this work, we have tried to follow the guidelines marked in [11] to discriminate between normal and pathological voices, extrapolating the techniques to the SVD, and taking benefit of the MultiFocal toolkit [19] for fusing different subsystems. MultiFocal toolkit is a toolkit developed by Niko Brümmer, and is widely used among the speaker and language recognition community.

The rest of the paper is organized as follows: in Section 1, the MEEI database and the SVD are presented; in Section 2, the features extracted from the audio are described; in Section 3, it is detailed how a GMM classifier works; in Section 4, a brief description of MultiFocal toolkit is given: in Section 5, the experiments that have been performed are presented and analyzed, for the two databases mentioned above; and in Section 6, the conclusions of this work are drawn.

1 Databases

1.1 MEEI, *Kay Elemetrics*

The same configuration adopted in [2, 6, 11] has been taken for the present work. There, 226 recordings of the whole database were used, corresponding to

vowel /ah/ sustained. From this subset, 173 files belong to pathological patients and 53 to normal speakers. Male and female speakers covering ages from 21 to 59 are uniformly distributed in both groups. The mean length of pathological recordings is 1 second and the one of normal recordings is 3 seconds. All files are converted to a common sampling frequency of 25 kHz and 16-bit resolution.

1.2 SVD, Saarland University

This database has been recently made freely available online [13]. It is a collection of voice recordings from more than 2000 persons, where a session is defined as a collection of:

- recordings of vowels /a/, /i/, /u/ produced at normal, high, low and low-high-low pitch.
- recording of sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?").

That makes a total of 13 files per session. In addition, the electroglottogram (EGG) signal is also stored for each case in a separate file. The length of the files with sustained vowels is between 1 and 3 seconds. All recordings are sampled at 50 kHz and their resolution is 16-bit. 71 different pathologies are contained, including both functional and organic. For our experiments only files with sustained vowels and people older than 18 are used. A total of 1970 sessions are kept, after discarding those where some of the recordings were missing or damaged. 1320 (609 males and 711 females) sessions belong to pathological speakers and 650 (400 males and 250 females) to normal speakers.

2 Features

The features used in this work are divided in two groups, according to their nature: acoustic features, represented by the MFCC, where the aim is to characterize the frequency content of the signal; and noise related features, represented by HNR, NNE and GNE, where the aim is to measure how good the quality of the signal is, or simply, how noisy it is.

2.1 Acoustic Features

MFCC are a family of parameters widely used for many tasks related with speech processing. It makes a frequency analysis of the signal based on the human perception of the sounds. This idea matches well with the fact that an experienced speech therapist can detect the presence of a disorder just by listening to the signal [10].

In the extraction procedure, after downsampling to 25 kHz, a 40 ms window with 50% overlap has been used, with a bank of 30 Mel filters, to obtain 15 MFCC plus log-energy. The first two and last two frames have been discarded to avoid possible errors in the edges of the recordings, like peaks due to the on and off switches. Finally, the coefficients are mean and variance normalized within each file.

2.2 Noise Related Features

Harmonics-to-Noise Ratio. HNR was introduced to measure in an objective manner the perceptual feeling of hoarseness in the voice [14]. To calculate it, the signal is firstly downsampled to 16 kHz, and split into 25 ms length frames, with 10 ms shift. In each frame, a comb filter is applied to the signal to compute the energy in the harmonic components. To the logarithm of this quantity, the log-energy of the noise is substracted to get the HNR.

Normalized Noise Energy. In a similar process to the calculation of the HNR, and also with the signal downsampled to 16 kHz and with 25 ms length frames and 10 ms shift, the noise estimation is calculated and normalized by the total energy of the signal. This was first used in [15] and it assumes that pathological voices are noisier than normal voices.

Glottal-to-Noise Excitation Ratio. The goal of this parameter is to compare the amount of signal due to vocal folds vibration with the amount of signal due to noise produced by air turbulences produced during phonation [16]. It is a good measurement of breathiness, although not the only factor that can cause it. To compute it, the signal is first downsampled to 10 kHz, and frames of 40 ms length with 20 ms shift are taken. For each frame, the spectrum is divided into bands of 2000 Hz with centers separated 500 Hz. For each of these bands, the Hilbert envelope in time domain is calculated and the correlation of this envelope with the envelopes of the bands separated more than half of the bandwidth (in this case, bands must be at least 1000 Hz) is computed. The GNE is the maximum of all correlations. If the voice is not pathological, the correlation should be high, because all bands should be excited at the same time when the glottis is closed.

3 GMM Classifier

The features extracted from the signal are used to train a generative GMM model [22] for each class. This model is the basis for many speech processing tasks, like speech, speaker, or language recognition. It is a generalization of the Gaussian model, and it permits to generate much more complicated likelihood functions.

For D-dimension features $\mathbf{x}$ calculated in a frame-by-frame basis, a GMM probability density function has the form

$$p(\mathbf{x}|\omega, \mu, \mathbf{\Sigma}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k), \tag{1}$$

where K is the number of Gaussians in the model, ω_k is the weight of the k*th* Gaussian, and $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ is the Gaussian function with mean μ_k and covariance Σ_k.

To train this model the expectation-maximization (EM) algorithm [23] has been used. K random Gaussians are generated for initialization and 10 iterations of the algorithm are performed. Full-covariance matrices are used. Then for each test file y, the likelihoods for pathological and normal classes are calculated, calibrated as explained in next section, and the log-likelihood ratio between them is obtained as

$$LLK(y) = \log p(y|pathological) - \log p(y|normal), \tag{2}$$

which will decide to which class the file belongs.

4 Calibration and Fusion with MultiFocal Toolkit

In pathology detection, the traditional metrics to evaluate the performance of the classifiers are:

- the area under the receiver operating characteristic (ROC) curve (AUC) [18]
- equal-error-rate (EER), point in ROC where the probability of miss is equal to the probability of false alarm
- correct classification rate (CCR), percentage of trials correctly classified
- error rate (ER), percentage of trials wrong classified, complementary to CCR
- sensitivity (S), ability of the classifier to detect the target class
- specifity (E), ability of the classifier to detect the impostor class

AUC and EER are calibration-insensitive evaluation metrics, and the last four are calibration-sensitive metrics. Since we are interested in a specific operating point, that is, in the hard decisions made by our classifier, we find more interesting to evaluate the real performance of our classifier with a calibration-sensitive metric than with AUC and EER. AUC and EER can be useful in early stages of our system development, when hard decisions are not of immediate interest and we are only interested in the goodness of uncalibrated scores [20].

In addition, two more metrics that we consider meaningful are the detection cost function (DCF) or empirical Bayes risk, and its minimum value for the selected operating point (minDCF) [19]. DCF is defined as

$$DCF = \pi C_{miss} P_{miss} + (1 - \pi) C_{fa} P_{fa}, \tag{3}$$

where π is the prior probability of the target class, in our case the pathological voice, C_{miss} and C_{fa} are the costs of a miss, that is, a pathological voice classified as normal, and a false alarm, that is, a normal voice classified as pathological, respectively, and P_{miss} and P_{fa} are the probabilities of a miss and a false alarm. It is a calibration-sensitive metric, since it depends on the current threshold. However, minDCF is calibration-insensitive, and it gives the minimum cost that could have been obtained with optimal calibration, at every operating point. It is calculated by varying the threshold from $-\infty$ to ∞ for each operating point, and then picking the minimum.

MultiFocal is a toolkit developed in Matlab primarily designed for calibrating and fusing scores of a language recognition task [19]. The aim of using this toolkit

is twofold: *i)* to calibrate scores so cost effective Bayes decisions can be made, by setting the threshold to the *Bayes decision threshold*, η,

$$\eta = \log \frac{C_{fa}}{C_{miss}} - logit(\pi); \tag{4}$$

ii) to fuse scores coming from different recognizers to obtain a better recognizer.

The idea behind calibration is that scores are converted in such a way that the Bayes decision threshold can be used for making the best possible decisions. Equivalently, the user could tune the threshold manually to minimize the error metric.

To calculate calibrated log-likelihoods, MultiFocal optimizes another calibration-sensitive metric, C_{llr} [21]. C_{llr} is defined as

$$C_{llr} = -\frac{1}{T}\sum_{t=1}^{T} \omega_t \log_2 P_t, \tag{5}$$

where T is the number of trials, ω_t is a weight to normalize the class proportions in the evaluation trials,

$$\omega_t = \frac{\pi_{c(t)}}{Q_{c(t)}} \quad (6)\ , \qquad Q_i = \frac{nr.\,of\,trials\,of\,class\,H_i}{T}, \tag{7}$$

c(t) is the true class of trial t, π_i the prior of class i, P_t is the posterior probability of hypothesis $H_{c(t)}$ of true class given the vector of calibrated log-likelihoods, $\boldsymbol{l'}(x_t)$, at trial t,

$$P_t = P(H_{c(t)}|\boldsymbol{l'}(x_t)) = \frac{\pi_{c(t)} e^{l'_{c(t)}(x_t)}}{\pi_{c(t)} e^{l'_{c(t)}(x_t)} + \pi_{\bar{c}(t)} e^{l'_{\bar{c}(t)}(x_t)}}, \tag{8}$$

being $\bar{c}(t)$ the impostor class at trial t, and x_t the observation at t. As we are in a 2-class problem, $c(t) \in$ {'0','1'}, being '0' the label for the target class, and '1' the label for the impostor class.

C_{llr} has the sense of a cost and it is measured in terms of bits of information. $0 \leq C_{llr} \leq \infty$, being 0 for perfect recognition.

Well-calibrated log-likelihoods, $\boldsymbol{l'}(x_t)$, are the final output of our calibration procedure. They are obtained as,

$$\boldsymbol{l'}(x_t) = \alpha \boldsymbol{l}(x_t) + \boldsymbol{\beta}, \tag{9}$$

where $\boldsymbol{l}(x_t)$ is the uncalibrated log-likelihood obtained from the classifier. Then, through the mimization of C_{llr}, we obtain the scalar α, that scales our outputs, and the vector β, that shifts our outputs. The optimization is made via a discriminative logistic regression.

More generally, to fuse K systems what we want is our calibrated log-likelihoods to be a linear combination of the uncalibrated log-likelihoods of the K systems,

$$\boldsymbol{l'}(x_t) = \sum_{k=1}^{K} \alpha_k \boldsymbol{l_k}(x_t) + \boldsymbol{\beta}. \tag{10}$$

As we can check, the fusion is a generalization of the calibration of a single system (where K=1), and since the fusion is also a calibration, because of the linearity of the operation, there is no need to pre-calibrate each input system, or to post-calibrate the fusion [19].

5 Experiments

The experiments conducted in this work are divided in two, according to the database tested. First, results for the MEEI datatabase will be shown. We have followed a similar procedure to [11]. This will be useful to compare our classifier with a state of the art system. After this check, our classifier will be used to test the SVD. This database will make possible to show how the fusion of different sources of information increase the performance of the system. The results will be given in terms of the traditional metrics described in Section 4, AUC, EER, CCR, ER, S and E, and also in terms of DCF and minDCF as additional information to check how good the calibration is. Finally, a confidence interval (CI) at 95% confidence ($\alpha = 0.05$) will be given for each experiment. Note that it is computed over the CCR.

For all our experiments, $C_{fa} = C_{miss} = 1$, $\pi_0 = \pi_1 = 0.5$, and threshold equal to the Bayes decision threshold, in our case η=0.

5.1 Results on MEEI

The database is divided in 30 folds, in the same manner as in [11]. For every test fold, the remaining 29 are used for training. Then, an average performance measure is extracted from the 30. GMMs with 3 components are trained. This is the optimal number found in [11]. One difference with [11] is that all recordings of the same speaker are grouped into the same fold, in such a way that one speaker is not in the training and test subsets at the same time. This will avoid recognizing the speaker instead of the pathology. Note that a slight drop in performance could be seen as a consequence, compared to the experiments in [11].

In table 1, results between normal and pathological classes are shown, where every recording is considered as a trial. The features are 19 dimensional, including 15 MFCCs plus log-energy, HNR, NNE and GNE, previously normalized in mean and variance. A comparison between results with and without calibration is done. As we can see, the calibration-insensitive metrics, AUC and EER, do not change. However, S, E, CCR and ER, indicate that something has changed. With calibration our system detects better both normal and pathological classes. The reason is that the scores have been transformed in such a way that the posterior probabilities of the true class are maximized, what at the same time minimizes the Bayes risk, because now the Bayes decision threshold will be the optimal one to make decisions. Note that we have trained the calibration with the data under test, and this gives the optimal values for α and β. In a real system, these values should be trained before any clinical evaluation, and the performance would not probably be optimal. However, in this way an upper bound of the results is obtained, and this is more reliable for comparison with other systems tested over

Table 1. Evaluation metrics for the experiments on MEEI database. Averages over 30 folds.

Metric	AUC	EER	CCR	ER	S	E	CI	DCF	minDCF
Calibrated	0.943	**0.048**	**0.948**	**0.052**	0.949	**0.950**	0.071	**0.050**	**0.033**
Uncalibrated	0.943	0.048	0.923	0.077	0.950	0.850	0.123	0.099	0.033
Work in [11]	**0.985**	-	0.943	0.057	**0.955**	0.907	**0.034**	0.069	-

the same data, since there is no dependence on any development data used for training the calibration.

We consider DCF a reasonable criterion to choose one classifier or another, because it weights both kind of errors, misses and false alarms, at the desired operating point. minDCF will tell us how good our system could have been with a perfect calibration. In turn, AUC and EER are evaluation metrics of the performance of our system considering all operating points. Also in table 1, DCF and minDCF can be found for the experiments made with MEEI. In [11] they have actually worked at the operating point given by EER, but we do not know the values for C_{fa}, C_{miss} and π. Assuming they are the same as ours (which can be an unfair but reasonable assumption since they work at EER and their effective prior is 0.5, what would give those values of C_{fa}, C_{miss} and π, for η =0), in terms of DCF, our system performs better. However, their AUC is very good, and it would probably give better estimated DCF if a calibration had been performed.

5.2 Results on SVD

In this case, 12 subsets of data are created by grouping separately the recordings belonging to /a/, /i/, and /u/, pronounced with normal, low, high, and low-high-low intonation. Then, for each one of these subsets a 30-fold strategy is followed. Also 3 components are trained in the GMM. In this case, grouping of the same speaker into the same subset is not guaranteed, what could give optimistic results. In short, the same procedure as in Section 5.1 is followed for each subset. In table 2, results for the same 19 dimension features as above are shown in terms of AUC, EER, CCR, ER, S, E, CI, DCF, and minDCF. Only calibrated results are shown. The behavior of the classifier for each vowel and intonation is interesting. It seems that the recognition rate is slightly better for /a/, but the differences are small. It can be checked that for /a/, normal and low intonations help the most, for /i/ all intonations behave similar, and for /u/, the normal intonation is the least discriminative.

Next, a partial fusion is made for each vowel, where the 4 intonations of each one are fused. This is made for every fold and then all folds are averaged. In table 2 this is in the last line of each vowel. It can be seen that the results are improved with regard to the case with every vowel and intonation tested individually, and that the classifier built with /a/ outperforms the ones with /i/ and /u/. However, looking at the confidence interval, no definitive conclusions should be made.

Table 2. Evaluation metrics for the experiments on SVD. Averages over 30 folds. Intonations are N: normal; L: low; H: high; and LHL: low-high-low.

Metric	AUC	EER	CCR	ER	S	E	CI	DCF	minDCF
Vowel /a/									
N	**0.747**	**0.321**	**0.670**	**0.330**	0.636	**0.739**	**0.112**	**0.313**	**0.270**
L	0.743	0.335	0.656	0.340	0.650	0.680	0.114	0.334	0.286
H	0.722	0.336	0.666	0.334	**0.655**	0.687	0.112	0.328	0.285
LHL	0.702	0.353	0.645	0.355	0.640	0.655	0.114	0.352	0.304
Fusion /a/	**0.804**	**0.277**	**0.718**	**0.282**	**0.701**	**0.752**	**0.108**	**0.273**	**0.234**
Vowel /i/									
N	0.702	0.350	**0.645**	**0.355**	0.627	0.682	**0.114**	**0.345**	0.305
L	**0.705**	**0.348**	0.642	0.358	0.620	**0.687**	0.115	0.347	**0.303**
H	0.700	0.354	0.640	0.359	0.629	0.664	0.115	0.352	0.305
LHL	0.679	0.373	0.639	0.361	**0.652**	0.612	0.116	0.368	0.322
Fusion /i/	**0.783**	**0.283**	**0.710**	**0.290**	**0.694**	**0.741**	**0.110**	**0.282**	**0.247**
Vowel /u/									
N	0.706	0.354	0.634	0.366	0.615	0.671	0.116	0.357	0.307
L	0.712	**0.342**	0.646	0.354	0.624	**0.692**	0.115	0.342	0.301
H	0.713	0.348	0.640	0.356	0.619	0.684	0.116	0.348	0.293
LHL	**0.715**	0.344	**0.666**	**0.334**	**0.678**	0.642	**0.114**	**0.340**	**0.293**
Fusion /u/	**0.797**	**0.282**	**0.715**	**0.284**	**0.702**	**0.741**	**0.108**	**0.278**	**0.242**
Fusion	**0.879**	**0.206**	**0.794**	**0.206**	**0.778**	**0.826**	**0.095**	**0.198**	**0.165**

Finally, a global fusion with all vowels and intonations is made for each fold, and as before, all folds are averaged to obtain a single figure for each of the metrics. This is in the last line of table 2. Actually, we do not fuse different systems, but the same system trained on different data. A huge increase in performance is obtained. Comparing with the best result without fusion (/a/ normal), the increase in perfomance is 17.67% for the AUC, 35.83% for the EER, and 36.74% for the DCF. We believe that the main reason is the fact of having much more data, and containing different information, because they come from different vowels and intonations. Again, these results are optimal in terms of the fusion, since the fusion parameters have been trained on the test data. If we compare this fusion with the partial fusion of each vowel, it can be seen that all vowels contribute to the improvement, because the global fusion outperforms the partial ones.

6 Conclusions

A new voice database, the SVD, has been evaluated for the task of pathology detection. The amount of recordings of different sounds and intonations included in this database makes possible to conduct different and interesting experiments. This is an open and free database available online. A robust GMM with 3 components, trained on MFCC, HNR, NNE and GNE, has been used as classifier, and the effect of calibration has been shown. Finally, a fusion of the classifiers trained on /a/, /i/ and /u/, pronounced with normal, low, high and low-high-low intonations, has been performed, showing that every sound gives different information to the system and their combination offers a huge improvement: 17.67% for the AUC and 36.75 % for the DCF. As future work we plan to test the effect of the number of Gaussians on the performance and other methods for fusing,

such as a simple concatenation of files of the same session. In addition, this work is thought to be a starting point for a further research with the SVD database, which is currently being perceptually classified according to the GRABS scale by a speech therapist of the *Bioingeniería y Optoelectrónica* (ByO) group at the Universidad Politécnica de Madrid.

Acknowledgements. Thanks to Manfred Pützer and his team for the great work of recording the SVD.

This work was funded by the Spanish Ministry of Science and Innovation under projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

References

[1] Godino Llorente, J.I., et al.: Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. IEEE Tr. Biomed. Eng. 53(10) (2006)
[2] Sáenz-Lechón, N., et al.: Methodological Issues in the Development of Automatic Systems for Voice Pathology Detection. Biomed. Signal Proc. and Control 1(2) (2006)
[3] Jiang, J.J., Zhang, Y.: Nonlinear Dynamic Analysis of Speech from Pathological Subjects. Electron. Lett. 38(6) (2002)
[4] Zhang, Y., Jiang, J.J.: Nonlinear Dynamic Analysis in Signals Typing of Pathological Human Voices. Electron. Lett. 39(13) (2003)
[5] Markaki, M., Stylianou, Y.: Using Modulation Spectra for Voice Pathology Detection and Classification. In: Proc. IEEE EMBS Annual Intern. Conf., Minneapolis, MN (2009)
[6] Parsa, V., Jamieson, D.G.: Identification of Pathological Voices Using Glottal Noise Measures. J. Speech, Lang. and Hearing Res. 43(2) (2000)
[7] Gavidia-Ceballos, L., Hansen, J.H.L.: Direct Speech Feature Estimation Using an Iterative EM Algorithm for Vocal Fold Pathology Detection. IEEE Tr. Biomed. Eng. 43(4) (1996)
[8] Tadeusiewicz, R., et al.: The Evaluation of Speech Deformation Treated for Larynx Cancer Using Neural Network and Pattern Recognition Methods. In: Proc. EANN 1998 (1998)
[9] Gelzinis, A., et al.: Automated Speech Analysis Applied to Laryngeal Disease Categorization. Comput. Methods Programs Biomed. 91 (2008)
[10] Arias-Londoño, J.D., et al.: On Combining Information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for Automatic Detection of Pathological Voices. Logop. Phoniatrics Vocology (2010)
[11] Sáenz Lechón, N.: Contribuciones Metodológicas para la Evaluación Objetiva de Patologías Laríngeas a partir del Ánalisis Acústico de la Voz en Diferentes Escenarios de Producción. PhD Thesis (2010)
[12] Kay Elemetrics Corp., Disordered Voice Database, Version 1.03 (CD-ROM), MEEI, Voice and Speech Lab, Boston, MA (October 1994)
[13] Barry, W.J., Pützer, M.: Saarbrücken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/
[14] Yumoto, E., et al.: Harmonics-To-Noise Ratio as an Index of the Degree of Hoarseness. J. Acoust. Soc. Am. 71 (1982)

[15] Kasuya, H., et al.: Normalized Noise Energy as an Acoustic Measure to Evaluate Pathologic Voice. J. Acoust. Soc. Am. 80(5) (1986)
[16] Michaelis, D., et al.: Glottal-to-Noise Excitation Ratio. A New Measure for Describing Pathological Voices. Acustica/Acta Acustica 83 (1997)
[17] Davis, S.B., Mermelstein, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Tr. Acoust. 28(4) (1980)
[18] Hanley, J.A., McNell, B.J.: The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. Radiology 143 (1982)
[19] Brümmer, N.: FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual, http://sites.google.com/site/nikobrummer/focalmulticlass
[20] Brümmer, N.: The BOSARIS ToolkitUser Guide: Theory, Algorithms and Code for Binary Classifier Score Processing, http://sites.google.com/site/bosaristoolkit
[21] Brümmer, N., du Preez, J.A.: Application-Independent Evaluation of Speaker Detection. Computer Speech and Language 20(2-3) (2006)
[22] Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Models. IEEE Tr. on Speech and Audio Proc. 3 (1995)
[23] Dempster, A.P., et al.: Maximum Likelihood from Incomplete Data Via the EM Algorithm. J. of the Royal Statistical Society 39, Series B (1977)
[24] Hirano, M.: Clinical Examination of Voice. Springer, New York (1981)
[25] Sáenz-Lechón, N., et al.: Automatic Assessment of Voice Quality According to the GRBAS scale. In: Proc. 28th IEEE EMBS Annual Intern. Conf. (2006)
[26] Carding, P., et al.: Formal Perceptual Evaluation of Voice Quality in the United Kingdom. Logop. Phoniatrics Vocology 25 (2000)
[27] Wuyts, F., et al.: The Dysphonia Severity Index: An Objective Measure of Vocal Quality Based on a Multiparameter Approach. J. Speech, Lang. and Hearing Res. 43 (2000)
[28] Hakkesteegt, M.M., et al.: The Relationship between Perceptual Evaluation and Objective Multiparametric Evaluation of Dysphonia Severity. J. of Voice 22 (2008)

Score Level versus Audio Level Fusion for Voice Pathology Detection on the Saarbrücken Voice Database

David Martínez, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel

Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{david,lleida,ortega,amiguel}@unizar.es

Abstract. The article presents a set of experiments on pathological voice detection over the Saarbrücken Voice Database (SVD). The SVD is freely available online containing a collection of voice recordings of different pathologies, both functional and organic. It includes recordings for more than 2000 speakers in which sustained vowels /a/, /i/, and /u/ are pronounced with normal, low, high, and low-high-low intonations. This variety of sounds makes possible to set different experiments, and in this paper a comparison between the performance of a system where all the vowels and intonations are pooled together to train a single model per class, and a system where a different model per class is trained for each vowel and intonation, and the scores of each subsystem are fused at the end, is conducted. The first approach is what we call audio level fusion, and the second is what we call score level fusion. For classification, a generative Gaussian mixture model trained with mel-frequency cepstral coefficients, harmonics-to-noise ratio, normalized noise energy and glottal-to-noise excitation ratio, is used. It is shown that the score level fusion is far more effective than the audio level fusion.

Keywords: Pathological Voice Detection, Saarbrücken Voice Database, GMM, Fusion, MultiFocal toolkit.

1 Introduction

The detection of laryngeal pathologies through an automatic voice analysis is one of the most promising tools for speech therapists, mainly due to its noninvasive nature and its objectivity for making decisions. The performance of these systems is nevertheless not perfect, and nowadays it is used as an additional source of information for other laryngoscopial exams [1].

Researchers have focused their efforts on finding new features that could discriminate between normal and pathological voices or even assess their quality, but also on finding different approaches for classification. Some of the most useful features are considered to be acoustic parameters such as mel-frequency cepstral coefficients (MFCC) [17, 1], amplitude and frequency perturbation parameters [9], and noise related parameters [2, 6], but there are different alternatives like nonlinear analysis [3, 4]. Regarding to the classifiers, well-known approaches in speech processing like hidden Markov model

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 110–120, 2012.

(HMM) [7], Gaussian mixture model (GMM) [1], multilayer perceptron (MLP) [8], or support vector machine (SVM) [9], have been studied.

Experiments related with pathological voices can be focused on three main tasks. While the most simple and direct idea is to classify voices as pathological or normal, like in [2, 6, 7, 10, 11], another goal is to assess voice quality according to a perceptual scale, like GRBAS [22, 23], DSI [25, 26] or VPA [24], among others. A third typical problem is to identify a specific pathology, like for example, functional dysphonia in [26], nodules and other laryngeal injuries in [9], or polyps, keratosis leukoplakia, adductor spasmodic dysphonia and vocal nodules in [5].

The main difficulty when facing a pathological voice related experiment is the database acquisition. Most of the works in the literature make use of the MEEI database, openly commercialized by *Kay Elemetrics* [12]. But the amount of recordings is limited and current approaches already give excellent performance, being difficult to evaluate improvements of new ideas. In other cases, private databases collected in local hospitals are the alternative. This is costly and generally they are not public. However, recently a new open and freely downloadable database, the SVD [13], has been recorded by the Institute of Phonetics of Saarland University. On it, sustained /a/, /i/ and /u/ vowels, pronounced with normal, low, high and low-high-low intonations, and a spoken sentence in German, are found, what make of this database a very complete set to conduct experiments, and easy to reach by all the community. In [27] a first approach is conducted to test this database on a pathological voice detection task, with GMMs as classifier, MFCC, harmonics-to-noise ratio (HNR) [14], normalized noise energy (NNE) [15], and glottal-to-noise excitation ratio (GNE) [16], as features. Calibration and fusion of scores coming from the subsystems built with each vowel and intonation is performed with MultiFocal toolkit [18]. MultiFocal toolkit is a toolkit developed by Niko Brümmer, and is widely used among the speaker and language recognition community. The fusion results are shown to be very promising.

In this work, following the guidelines marked in [11] to discriminate between normal and pathological voices, a comparison between the fusion techniques used in [27] and a pool-of-data strategy is conducted, where instead of fusing different systems trained on each vowel and intonation, a single system trained on all vowels and intonations pooled together is used.

The classifier used in this work was also evaluated on MEEI database in [27]. The results were close to similar approaches in the bibliography, like [11]. Concretely, the Area Under the Curve (AUC) was 0.943 and 94.93% of pathological files and 95.00% of the normal files were correctly classified.

The rest of the paper is organized as follows: in Section 2, the SVD is presented; in Section 3, the features extracted from the audio are described; in Section 4, the classification, calibration and fusion procedures are explained; in Section 5, the experiments that have been performed are presented and analyzed, for the two strategies mentioned above; and in Section 6, the conclusions of this work are drawn.

2 Saarbrücken Voice Database

This database has been recently made freely available online [13]. It is a collection of voice recordings from more than 2000 persons, where a session is defined as a collection of:

– recordings of vowels /a/, /i/, /u/ produced at normal, high, low and low-high-low pitch.

– recording of sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?").

That makes a total of 13 files per session. In addition, the electroglottogram (EGG) signal is also stored for each case in a separate file. The length of the files with sustained vowels is between 1 and 3 seconds. All recordings are sampled at 50 kHz and their resolution is 16-bit. 71 different pathologies are contained, including both functional and organic. For our experiments only files with sustained vowels and people older than 18 are used. A total of 1970 sessions are kept, after discarding those where some of the recordings were missing or damaged. 1320 (609 males and 711 females) sessions belong to pathological speakers and 650 (400 males and 250 females) to normal speakers.

3 Features

The features used in this work are divided into two groups, according to their nature: acoustic features, represented by the MFCC, where the aim is to characterize the frequency content of the signal; and noise related features, represented by HNR, NNE and GNE, where the aim is to measure how good the quality of the signal is, or simply, how noisy it is.

3.1 Acoustic Features

MFCC are a family of parameters widely used for many tasks related with speech processing. It makes a frequency analysis of the signal based on the human perception of the sounds. This idea matches well with the fact that an experienced speech therapist can detect the presence of a disorder just by listening to the signal [10].

In the extraction procedure, after downsampling to 25 kHz, a 40 ms window with 50% overlap has been used, with a bank of 30 Mel filters, to obtain 15 MFCC plus log-energy. The first two and last two frames have been discarded to avoid possible errors in the edges of the recordings, like peaks due to the on and off switches.

3.2 Noise Related Features

Harmonics-to-Noise Ratio. HNR was introduced to measure in an objective manner the perceptual feeling of hoarseness in the voice [14]. To calculate it, the signal is firstly downsampled to 16 kHz, and split into 25 ms length frames, with 10 ms shift. In each frame, a comb filter is applied to the signal to compute the energy in the harmonic components. To the logarithm of this quantity, the log-energy of the noise is substracted to get the HNR.

Normalized Noise Energy. In a similar process to the calculation of the HNR, and also with the signal downsampled to 16 kHz and with 25 ms length frames and 10 ms shift, the noise estimation is calculated and normalized by the total energy of the signal. This was first used in [15] and it assumes that pathological voices are noisier than normal voices.

Glottal-to-Noise Excitation Ratio. The goal of this parameter is to compare the amount of signal due to vocal folds vibration with the amount of signal due to noise produced by air turbulences produced during phonation [16]. It is a good measurement of breathiness, although not the only factor that can cause it. To compute it, the signal is first downsampled to 10 kHz, and frames of 40 ms length with 20 ms shift are taken. For each frame, the spectrum is divided into bands of 2000 Hz with centers separated 500 Hz. For each of these bands, the Hilbert envelope in time domain is calculated and the correlation of this envelope with the envelopes of the bands separated more than half of the bandwidth (in this case, bands must be at least 1000 Hz) is computed. The GNE is the maximum of all correlations. For a normal voice, the correlation should be high, because all bands should be excited at the same time when the glottis is closed.

4 Classification, Calibration and Fusion

4.1 Classification

The features extracted from the signal are used to train a generative GMM [21] for each class. For D-dimension features $\mathbf{x}$ calculated in a frame-by-frame basis, a GMM probability density function has the form

$$p(\mathbf{x}|\omega, \mu, \mathbf{\Sigma}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k), \tag{1}$$

where K is the number of Gaussians in the model, ω_k is the weight of the kth Gaussian, and $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ is the Gaussian function with mean μ_k and covariance Σ_k.

For each test file y, the likelihoods for pathological and normal classes are calculated, calibrated as explained in section 4.2, and the log-likelihood ratio between them is obtained as

$$LLK(y) = \log p(y|pathological) - \log p(y|normal), \tag{2}$$

which will decide to which class the file belongs.

The metrics to evaluate the performance used in this work are the AUC of the receiver operating characteristic (ROC), the equal-error-rate (EER), the detection cost function (DCF) or empirical Bayes risk, and its minimum value for the selected operating point (minDCF) [18]. DCF is defined as

$$DCF = \pi C_{miss} P_{miss} + (1 - \pi) C_{fa} P_{fa}, \tag{3}$$

for a false alarm cost C_{fa}, a miss cost C_{miss}, a prior probability for the target class π, a false alarm probability P_{fa}, and a miss probability P_{miss}. DCF is a calibration-sensitive metric, since it depends on the current threshold. However, minDCF is calibration-insensitive, and it gives the minimum cost that could have been obtained with optimal calibration, at every operating point. It is calculated by varying the threshold from $-\infty$ to ∞ for each operating point, and then picking the minimum. AUC and EER are also calibration-insensitive metrics. We are interested in the hard decisions made by our classifier to decide if a voice is pathological or not, therefore we find more interesting to use calibration sensitive metrics like DCF. AUC and EER can be useful in early stages of our system development, when hard decisions are not of immediate interest and we are only interested in the goodness of uncalibrated scores [19].

4.2 Calibration and Fusion

MultiFocal toolkit [18] is used for calibration and fusion. This toolkit developed in Matlab is primarily designed for calibrating and fusing scores of a language recognition task. The goal of the toolkit is twofold: *i)* to calibrate scores so cost effective Bayes decisions can be made, by setting the threshold to the *Bayes decision threshold*, η,

$$\eta = \log \frac{C_{fa}}{C_{miss}} - logit(\pi), \tag{4}$$

being the pathological voices our target class; and *ii)* to fuse scores coming from different recognizers to obtain a better recognizer. In our experiments, $C_{fa} = C_{miss} = 1$, $\pi_0 = \pi_1 = 0.5$, and threshold equal to the Bayes decision threshold, in our case η=0.

The idea behind calibration is that our scores are converted in such a way that the Bayes decision threshold can be used for making the best possible decisions. Equivalently, the user could tune the threshold manually to minimize the error metric.

To calculate calibrated log-likelihoods, $\boldsymbol{l'}(x_t)$, MultiFocal optimizes another calibration-sensitive metric, C_{llr}, through a discriminative logistic regression, and obtains a scalar α, and a vector β [20]. Then

$$\boldsymbol{l'}(x_t) = \alpha \boldsymbol{l}(x_t) + \boldsymbol{\beta}, \tag{5}$$

where $\boldsymbol{l}(x_t)$ is the uncalibrated log-likelihood obtained from the classifier.

More generally, to fuse K systems what we want is our calibrated log-likelihoods to be a linear combination of the uncalibrated log-likelihoods of the K systems,

$$\boldsymbol{l'}(x_t) = \sum_{k=1}^{K} \alpha_k \boldsymbol{l_k}(x_t) + \boldsymbol{\beta}. \tag{6}$$

As we can check, the fusion is a generalization of the calibration of a single system (K=1), and since the fusion is also a calibration, due to the linearity of the operation, there is no need to pre-calibrate each input system, or to post-calibrate the fusion [18].

5 Experiments

The experiments conducted in this work are divided in two, according to the strategy followed to combine all the different varieties of sounds. The first strategy is the one implemented in [27], where an individual subsystem is trained for each vowel and intonation, and a discriminative score level fusion of all the subsystems is performed with MultiFocal toolkit. We will revise this experiment in the first part of section 5.1. In the second part of this section, an experiment has been configured with a single common GMM trained with all the files of all vowels and intonations, and the fusion is made with the scores tested over this common model. This will make possible to see if the different sounds are correlated and we can take benefit from one to the other. The second strategy trains a unique model per class by concatenating all the files with the different sounds belonging to the same speaker, and in the test phase, all the files belonging to the same speaker are also concatenated and tested over this model as a single file. This is called audio level fusion.

The features used as input for our classifiers will be 19-dimensional, including 15 MFCC + log-energy + HNR + NNE + GNE, all of them mean and variance normalized within each file. A 30-fold strategy is followed in all experiments, in which for every test fold, the remaining 29 are used for training. Then, an average performance measure is extracted from the 30, in the same manner as in [11].

Table 1. Metrics with score level fusion and individual models for each vowel and intonation for 3, 6, 8, 16 and 36 Gaussians. Average over 30 folds.

Metric	AUC	EER	DCF	minDCF
3 G	0.879	0.206	0.198	0.165
6 G	0.891	0.226	0.228	0.169
8 G	0.886	0.191	0.187	0.158
16G	0.890	**0.190**	**0.184**	**0.148**
36 G	**0.899**	0.274	0.274	0.191

5.1 Score Level Fusion

Score Level Fusion with Individual Training Subsets In this case, 12 subsets of data are created by grouping the recordings of all speakers belonging to the same vowel and intonation. That is, different subsets for the vowels /a/, /i/, and /u/, pronounced with normal, low, high, and low-high-low intonation are created. With each subset a different model is trained, and in the test phase, every vowel and intonation is tested against its model and the scores are fused at the end. A graphical view of the model is in figure 1. This is the same experiment made in [27], and the results for each vowel and intonation can be consulted in this reference. Averaged fused results for 3, 6, 8, 16 and 36 Gaussians are included in table 1. As we can see, there is not a significant improvement when increasing the number of Gaussians. Remember that every model is trained with data of only one vowel and intonation, and thus the number of modes in each dataset is not big. The model with 16 Gaussians gives slightly better perfomance than the others.

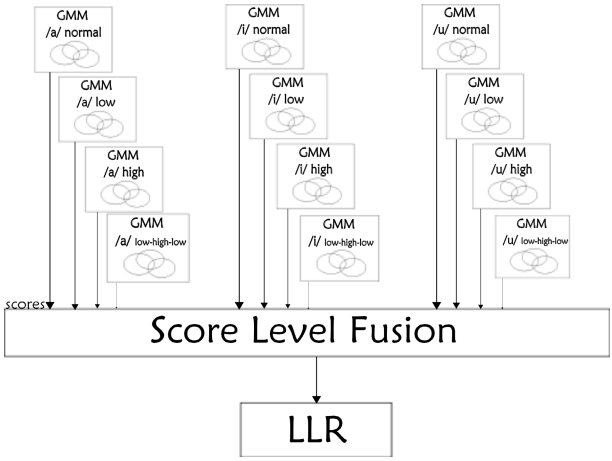


Fig. 1. Score level fusion scheme with one model per vowel and intonation

The analysis of each vowel and intonation individually is also in [27], but we will summarise the main points in this paragraph, since the same trends can be seen in the rest of experiments of this article. The first evidence is that vowel /a/ pronounced with normal intonation performs the best, while /i/ pronounced with low-high-low intonation performs the worst. In particular, for /a/ normal AUC is 0.747, EER is 0.321, DCF is 0.313 and minDCF is 0.270. Studying the fusion of all intonations for each vowel, it can be seen that vowel /a/ outperforms vowels /u/ and /i/, and vowel /u/

outperforms vowel /i/. However, the overlap in the figures according to the confidence interval, says that no definitive conclusions should be drawn. It is also noticeable the important improvement after this partial fusion. For example, for 3 Gaussians, the relative improvement in AUC is 7.63%, and in DCF is 12.78%, comparing /a/ with all pronountiations fused and /a/ normal. If we look at the intonations, the variability is high: among the results with /a/, the normal intonation works better, for /i/ the low intonation works better, and for /u/ the low-high-low intonation is the one that performs the best. As we see, no specific intonation is better than others, and it depends on the vowel pronounced by the subject. With the global fusion with all vowels and intonations a huge increase in performance is obtained. Comparing with the best result without fusion (/a/ normal), the increase in perfomance, again for the case of 3 Gaussians, is 17.67% for the AUC and 36.74% for the DCF. We believe that the main reason is the fact of having much more data, and containing different information, because they come from different vowels and intonations. Note that these results are optimal in terms of the fusion, since the fusion parameters have been trained on the test data. If we compare this fusion with the partial fusion of each vowel, it can be seen that all vowels contribute to the improvement, because the global fusion outperforms the partial ones.

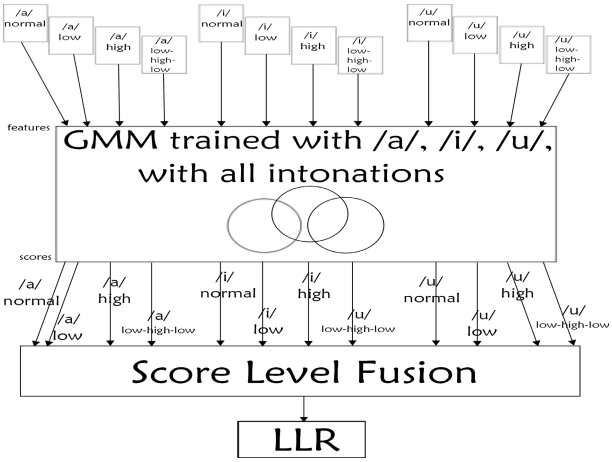


Fig. 2. Score level fusion scheme with a common model for all vowels and intonations

Score Level Fusion with a Common Training Subset. This experiment is similar to the one in the previous section, where a late fusion with MultiFocal toolkit is done, but training a common GMM with all training data, including all vowels and intonations. Then, in the test phase, every vowel and intonation is tested against this common model and the scores are fused at the end. The scheme can be seen in figure 2. In table 2 averaged results are shown for the metrics described in Section 4.1. Only fusion results are shown for 3, 6, 8, 16 and 36 Gaussians. In this case, as the model includes all the sounds, it could be expected to reach a higher number of Gaussians robustly trained. However, as it can be seen in the tables, the best results are obtained with 16 Gaussians. The individual behavior of the different vowels and intonations is similar to the previous case, being the best results obtained for /a/ pronounced with normal intonation.

Table 2. Metrics with score level fusion and a common model for all vowels and intonations for 3, 6, 8, 16 and 36 Gaussians. Average over 30 folds.

Metric	AUC	EER	DCF	minDCF
3 G	0.868	0.210	0.210	0.174
6 G	0.865	0.230	0.226	0.179
8 G	0.867	0.214	0.214	0.177
16G	**0.892**	**0.192**	**0.190**	**0.154**
36 G	0.879	0.201	0.198	0.169

More interesting is the comparison with the results of the previous experiment. Comparing tables 1 and 2, in both experiments the optimal number of Gaussians is 16, and the differences between each other are not meaningful. In short, we obtain no benefit when training the models with all data, mixing all vowels and intonations. That can be a sign of being each of the sounds really independent of each other, because they do not take advantage of a bigger model trained with the pool of the data.

5.2 Audio Level Fusion

In this case, we train a single GMM with data coming from all vowels and pronunciations, as in the second experiment of section 5.1. In the test phase, all the files belonging to the same speaker are concatenated and evaluated as just one single file. The difference with the previous case is that now a single score per speaker is obtained and no score level fusion is needed. A graphical scheme can be seen in figure 3. In table 3 we have the averaged results in terms of AUC, EER, DCF and minDCF for 3, 6, 8, 16 and 36 Gaussians. For comparison with the first case of section 5.1 (also in [27]), note that there we needed 12 (models) $\times$ [3 (Gaussians per model) $\times$ [19 (means) + 190 (Σ)]+3 (weights)] + 12 (fusion weights) + 2 (fusion offsets) = 7574 parameters for 3 Gaussians, and 90734 parameters for 36 Gaussians. Now, in the most demanding case, the one with 36 components, we need 36 (weights) + 36 $\times$ [19 (means) + 190 (Σ)] + 1 (calibration weight) + 2 (calibration offset) = 7563 parameters. We see that

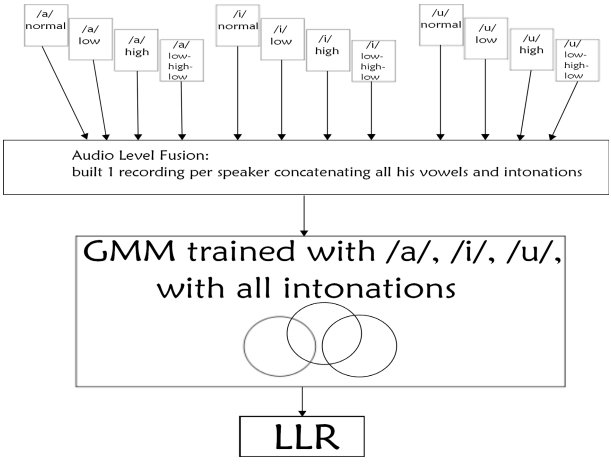


Fig. 3. Audio level fusion scheme

Table 3. Audio level fusion metrics for 3, 6, 8, 16 and 36 Gaussians. Average over 30 folds.

Metric	AUC	EER	DCF	minDCF
3 G	0.768	0.308	0.310	0.250
6 G	0.777	0.303	0.305	0.256
8 G	0.771	0.311	0.302	0.259
16G	0.788	0.303	**0.284**	**0.242**
36 G	**0.790**	**0.303**	0.286	0.242

the audio level fusion makes possible to save parameters, since the most demanding case is similar to the least demanding case of the score level fusion. The interest lies in checking if it is better an early fusion, as the one presented in this section, or a late fusion, as the one presented in the previous subsection.

As it can be checked in the different experiments, the performance is much better when fusing at score level that when concatenating the different files of the same session and doing the test of a single file. For example, with 16 Gaussians, the relative improvement of using score level fusion with individual training subsets for each vowel and intonation with respect to the audio level fusion is 12.94% in terms of AUC and 35.21% in terms of DCF. Since the number of parameters is very similar between the case of score level fusion with a 3 component GMM and the case of audio level fusion with 36 component GMM, more similar results could be expected. One explanation could be that the audio level fusion is more sensitive to errors in any of the concatenated files, whereas in the score level fusion, the fact of having different weights for each of the sounds makes the system more flexible and able to give stronger weights to the sounds working better.

As in section 5.1, the fact of having longer files with larger sound variability could take benefit of a model with more components. But also in this case it can be appreciated that the 16 Gaussian model gives slightly better results than the rest, supporting the theory presented before, stating that every vowel and intonation is independent of the rest and there is no benefit of a bigger model trained with all sounds. Note also that the audio level fusion gives improvements with regard to the case where only one of the vowels and intonations is considered.

6 Conclusions

SVD is an open and free database available online. The amount of recordings of different sounds and intonations contained in this database makes possible to conduct different and interesting experiments. In this article a comparison between voice pathology detection experiments carried out on the SVD with a score level fusion and an audio level fusion is presented. The score level fusion refers to the process in which every file with a different vowel and intonation is tested separately and the scores are fused at the end. The audio level fusion refers to the concatenation of the files with different sounds into a single file that is evaluated to obtain directly a single likelihood. A robust GMM, trained on MFCC, log-energy and HNR, NNE and GNE, has been used as classifier. The score level fusion gives better results than the audio level fusion. For a model trained with 16 Gaussians, the AUC is a 12.94% higher and DCF a 35.21% lower in the former than in the latter. The improvement of the score level fusion results with respect to the evaluation of a single vowel and intonation alone is also huge: a 17.67% in AUC and 36.75% in DCF, with respect to /a/ pronounced in normal intonation, which

is the one with the best individual performance. It is also interesting to see that the optimal number of Gaussians is 16 both in the score level and in the audio level fusion. In addition, training a bigger model with the different sounds pooled into the same file gives no further benefit, an indication of independence among different sounds.

Acknowledgements. This work was funded by the Spanish Ministry of Science and Innovation under projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

References

[1] Godino Llorente, J.I., et al.: Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. IEEE Tr. Biomed. Eng. 53(10) (2006)
[2] Sáenz-Lechón, N., et al.: Methodological Issues in the Development of Automatic Systems for Voice Pathology Detection. Biomed. Signal Proc. and Control 1(2) (2006)
[3] Jiang, J.J., Zhang, Y.: Nonlinear Dynamic Analysis of Speech from Pathological Subjects. Electron. Lett. 38(6) (2002)
[4] Zhang, Y., Jiang, J.J.: Nonlinear Dynamic Analysis in Signals Typing of Pathological Human Voices. Electron. Lett. 39(13) (2003)
[5] Markaki, M., Stylianou, Y.: Using Modulation Spectra for Voice Pathology Detection and Classification. In: Proc. IEEE EMBS Annual Intern. Conf., Minneapolis, MN (2009)
[6] Parsa, V., Jamieson, D.G.: Identification of Pathological Voices Using Glottal Noise Measures. J. Speech, Lang. and Hearing Res. 43(2) (2000)
[7] Gavidia-Ceballos, L., Hansen, J.H.L.: Direct Speech Feature Estimation Using an Iterative EM Algorithm for Vocal Fold Pathology Detection. IEEE Tr. Biomed. Eng. 43(4) (1996)
[8] Tadeusiewicz, R., et al.: The Evaluation of Speech Deformation Treated for Larynx Cancer Using Neural Network and Pattern Recognition Methods. In: Proc. EANN 1998 (1998)
[9] Gelzinis, A., et al.: Automated Speech Analysis Applied to Laryngeal Disease Categorization. Comput. Methods Programs Biomed. 91 (2008)
[10] Arias-Londoño, J.D., et al.: On Combining Information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for Automatic Detection of Pathological Voices. Logop. Phoniatrics Vocology (2010)
[11] Sáenz Lechón, N.: Contribuciones Metodológicas para la Evaluación Objetiva de Patologías Laríngeas a partir del Ánalisis Acústico de la Voz en Diferentes Escenarios de Producción. PhD Thesis (2010)
[12] Kay Elemetrics Corp., Disordered Voice Database, Version 1.03 (CD-ROM), MEEI, Voice and Speech Lab, Boston, MA (October 1994)
[13] Barry, W.J., Pützer, M.: Saarbrücken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/
[14] Yumoto, E., et al.: Harmonics-To-Noise Ratio as an Index of the Degree of Hoarseness. J. Acoust. Soc. Am. 71 (1982)
[15] Kasuya, H., et al.: Normalized Noise Energy as an Acoustic Measure to Evaluate Pathologic Voice. J. Acoust. Soc. Am. 80(5) (1986)

[16] Michaelis, D., et al.: Glottal-to-Noise Excitation Ratio. A New Measure for Describing Pathological Voices. Acustica/Acta Acustica 83 (1997)
[17] Davis, S.B., Mermelstein, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Tr. Acoust. 28(4) (1980)
[18] Brümmer, N.: FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual, http://sites.google.com/site/nikobrummer/focalmulticlass
[19] Brümmer, N.: The BOSARIS ToolkitUser Guide: Theory, Algorithms and Code for Binary Classifier Score Processing, http://sites.google.com/site/bosaristoolkit
[20] Brümmer, N., du Preez, J.A.: Application-Independent Evaluation of Speaker Detection. Computer Speech and Language 20(2-3) (2006)
[21] Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Models. IEEE Tr. on Speech and Audio Proc. 3 (1995)
[22] Hirano, M.: Clinical Examination of Voice. Springer, New York (1981)
[23] Sáenz-Lechón, N., et al.: Automatic Assessment of Voice Quality According to the GRBAS scale. In: Proc. 28th IEEE EMBS Annual Intern. Conf. (2006)
[24] Carding, P., et al.: Formal Perceptual Evaluation of Voice Quality in the United Kingdom. Logop. Phoniatrics Vocology 25 (2000)
[25] Wuyts, F., et al.: The Dysphonia Severity Index: An Objective Measure of Vocal Quality Based on a Multiparameter Approach. J. Speech, Lang. and Hearing Res. 43 (2000)
[26] Hakkesteegt, M.M., et al.: The Relationship between Perceptual Evaluation and Objective Multiparametric Evaluation of Dysphonia Severity. J. of Voice 22 (2008)
[27] González, D.M., Solana, E.L., Giménez, A.O., Artiaga, A.M., Villalba, J.: Voice Pathology Detection on the Saarbrücken Voice Database with Calilbration and Fusion of Scores Using MultiFocal Toolkit. In: Toledano, D.T., Giménez, A.O., Teixeira, A. (eds.) IberSPEECH 2012. CCIS, vol. 328, pp. 99–109. Springer, Heidelberg (2012)

Using HMM to Detect Speakers with Severe Obstructive Sleep Apnoea Syndrome

Ana Montero Benavides[1], José Luis Blanco[1], Alejandra Fernández[2], Rubén Fernandez Pozo[1], Doroteo Torre Toledano[2], and Luis Hernández Gómez[1]

[1] Signal, Systems & RadioCommunications Department
Universidad Politécnica de Madrid, Spain
{ana.montero,jlblanco,ruben,luis}@gaps.ssr.upm.es
[2] ATVS Biometric Recognition Group
Universidad Autónoma de Madrid, Spain
alejandra.fernandezh@estudiante.uam.es, doroteo.torre@uam.es

Abstract. Nowadays definitive diagnosis of obstructive sleep apnoea (OSA) syndrome is expensive and time-consuming. Previous research on voice characteristics of OSA patients has shown that resonance, phonation and articulation differences arise when compared to healthy subjects. In this contribution we study different speech modeling techniques to detect patients with severe OSA envisioning the future classification of patients according to their priority of need identifying the most severe cases and reducing medical costs.

Hidden Markov Models (HMMs) are used, as generally applied in text-dependent speech recognition, for detecting voices of OSA patients. Specific acoustic properties of continuous speech are modeled attending to different linguistic contexts which reflect discriminative physiological characteristics found in OSA patients. Experimental results on the discrimination of apnoea voices are presented over a database including both severe OSA and healthy speakers. An 85% correct classification rate is achieved by using whole-sentence HMMs, outperforming previous schemes proposed in the literature.

Keywords: HMM, GMM, Phoneme models, Linguistic context, Obstructive sleep apnoea syndrome.

1 Introduction

Obstructive sleep apnoea (OSA) is a common sleep disorder that affects 2-4% of adults and 11% of children [1]. The male-female ratio in the United States in clinical practice varies from 6:1 to 10:1 [2, 3]. OSA consists in the complete cessation of airflow for more than 10 seconds, as a result, a fall in oxygen saturation and arousal from sleep occurs [4]. The AHI (apnoea-hypopnea index), the measure used to quantify the severity of patients' condition, refers to the number of apnoeas and hypopneas occurring per hour of sleep. It is considered mild between 5 and 15 and severe above 30. Epidemiologic studies have shown a frequent prevalence of undiagnosed OSA.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 121–128, 2012.

Even mild OSA is associated with significant morbidity [5] and mortality [6]. OSA is also a risk factor for cardiovascular disease [7], often related to traffic accidents caused by somnolent drivers [7-9] and it can lead to a poor quality of life and impaired work performance.

Nowadays, the definitive diagnosis of OSA requires an expensive full overnight study, the polysomnography, considered the "gold standard" in sleep disorders. It consists in recording and processing neuroelectrophisiological and cardiorespiratory variables which is very time-consuming. In Spain, e.g. patients have to endure a waiting list of several years before the test is done. This delays the proper therapy. If not treated, OSA is a serious health risk [9]. The common treatment is continuous positive airway pressure (CPAP), i.e., providing a pneumatic splint to the airway during sleep. This prevents apnoea episodes and reduces snoring, one of the earliest symptoms of OSA. Alternative methods for early diagnosis of OSA are required. Speech-based methods for OSA detection are promising in this respect, due to their non-intrusive nature and their ability to provide quantitative data swiftly reducing the time for diagnosis. Our main goal is to classify patients according to the severity of their disease just by the analysis of some extracts of speech, complementary to existing OSA diagnosis methods and clinicians' judgment and as an aid for early detection of these cases.

Previous studies [10, 11] have confirmed that patients with OSA commonly have narrower and more collapsible upper airways (UAs) than patients without OSA, suggesting that OSA could be associated with anatomical and functional abnormalities of the UA. Unfortunately, not much research has been carried out on the acoustic properties of speakers suffering from OSA. Nonetheless, abnormalities in phonation, articulation and resonance have been found, although differences are somewhat unclear [12]. What seemed to be clear was that the apnoea group had abnormal resonances that might be due to altered structure or function of the UA, and this anomaly could result not only in a respiratory but also in a speech dysfunction.

The standard approach concerning pathological voices detection has most often been based on sustained vowels analysis, though the analysis of continuous speech offers more possibilities. Specific patterns present in OSA voices could be present in the transitions between different phonetic units [13]. According to this, we focus on continuous speech signals as in [14], where a remarkable discussion on how ASR techniques could be applied to the OSA detection is provided. Efforts have been devoted to the characterization of OSA patients' acoustic space to trace specific patterns connected to the apnoea syndrome [8, 12, 14, 15]. Nonetheless, no previous work has focused on introducing HMMs to improve OSA characterization.

The rest of this paper is organized as follows: Section 2 briefly presents the database we used in our study. The use of different HMM models to characterize and discriminate continuous speech from apnoea and healthy speakers are described in Section 3. Section 4 presents experimental results over the speech database and a discussion on them. Finally, some conclusions and future research are given in Section 5.

2 Apnoea Database

2.1 Data Collection

The apnoea database was recorded at the Respiratory Department of the Hospital Clínico Universitario of Málaga, Spain. It contains the voices from 80 male subjects with similar physical characteristics such as age and body mass index (BMI, height to square-weight ratio). 40 of these subjects present severe sleep apnoea (AHI > 30), while the other 40 from the control group present mild OSA (AHI < 10). Speech signals were recorded at 16 kHz sampling rate in an acoustically isolated booth. Recording equipment was a standard laptop computer equipped with a SP500 Plantronics headset microphone that includes A/D conversion and digital data exchange through USB-port.

2.2 Speech Corpus

The speech corpus includes four sentences that are repeated three times each in an alternate sequence by each speaker. Sentences were designed trying to cover all the relevant linguistic contexts where physiological OSA features could have higher impact on specific acoustic characteristics, keeping in mind the results from the perceptual study by Fox and colleagues [12]. The phrases include instances of the following specific phonetic contexts:

- In relation to resonance anomalies, sentences were designed to allow measuring differential voice features for each speaker (e.g. to compare the degree of vowel nasalization).
- Regarding phonation anomalies, we included consecutive voiced sounds to measure irregular phonation patterns related to muscular fatigue in apnea patients.
- To look at articulatory anomalies, we collected voiced sounds affected by preceding phonemes that have their primary locus of articulation near the back of the oral cavity.

More details about the corpus and database can be found in [16].

3 HMM Models to Characterize OSA Voices

In this contribution we consider apnoea/control classification as a two-class recognition problem using two different statistical models: one trained for the apnoea group and the other for the control group (i.e. *healthy* speakers). This approach was followed in [14] using Gaussian Mixture Models (GMMs) to fit apnoea and control acoustic spaces. The severe apnoea detection system described there will be used as a baseline system in this contribution to compare the new approaches using HMMs. It is important to point out that when using two GMMs for modeling the apnoea and control group, respectively, all the acoustic space from the four sentences in our data-

base is represented in a text-independent way. That is, no specific modeling is done to represent the characteristics of different sounds depending on their particular linguistic context.

Differently from using GMMs, in this work we explore different ways to characterize OSA voices using HMMs looking for a better modeling of the aforementioned different linguistic contexts. More specifically, we have followed two approaches: a) to use HMM phoneme models trained for each one of our four sentences, which will be denoted as *sentence-dependent phonemes*; and b) to train a whole HMM model for each sentence, that will be referred to as *whole sentence HMMs*. Two different sets of *whole sentences* and *sentence-dependent phonemes* HMM models are needed for representing the apnoea and control groups, and were tested in detecting severe OSA cases.

The rationale behind using these two types of HMM units is as follows. Context-dependent (CD) phonetic modeling is broadly used in large vocabulary ASR systems. The most frequent units for CD modeling are triphones which represent a phone in a particular left and right context. However, as the specificity of the model increases, the number of parameters to train also increases, and as the amount of triphone repetitions in our database is limited (e.g. the first triphone of the first sentence only appears once) we decided to test two different alternative approaches.

The first approach, *sentence-dependent phonemes,* was designed to use context-independent (CI) phonemes but training a different CI phoneme depending on each one of the four sentences in our corpus. So, e.g. the CI HMM model for phoneme /a/ in sentence 1 is different from CI HMMs for phoneme /a/ in sentences 2, 3 and 4 respectively. In that way HMM models of *sentence-dependent phonemes* should be able to model possible apnea/control voices differences related to the specific different phonetic contexts defined in the design of each one of the sentences (as previously described in sub-section 2.2).

While for the second approach, *whole sentence* models, different HMMs have been used for each sentence with a number of states close to the number of phonemes in the sentence (as will be described in Section 4, best results were obtained when using 40 states, and the average number of phonemes for the four sentences is 42). Therefore *whole sentence* HMM models can be seen very similar to the use of CD triphones, as they represent the specific linguistic structure of each sentence through the sequence of states, but with the benefit of containing a reduced number of states, thus providing an efficient way to train them (based on the limited amount of available data). Furthermore, another interesting feature of *whole sentence* HMMs is that they are not constrained to use explicit phoneme information and every HMM state trained can freely adjust to different units.

4 Experiments

In addition to the methodology described, throughout this work the standard leave-one-out cross-validation protocol was used. This protocol will sequentially discard all audio records from one sample speaker and use all the remaining apnea and control

samples for training the apnoea/control classifier. The excluded samples are then used as test data to evaluate each system's performance, and a figure of merit, e.g. DET (Detection Error Trade-off) curve and/or equal error rate (EER), is estimated to compare multiple schemes.

Audio files parameterization included 12 MFCC (Mel Frequency Cepstral Coefficients), without velocity or acceleration coefficients, to extract the information representing the speakers' acoustic space. Conventional MFCC parameterization provides relative independent coefficients and high discrimination between sounds, while being designed relying on human perception processing. Different parameterizations were tested, while the best results were obtained with 12 MFCCs. Due to the lack of space, the discussion on the optimization of these parameters for the apnoea detection problem will be presented in future publications.

Training and evaluation of HMMs were performed using the Hidden Markov Model Toolkit (HTK) [17], while the aforementioned text-independent baseline system used to compare our results is based on GMMs [14] and was trained using the BECARS open source toolkit [18]. In both situations the number of mixtures included in the trained models was limited to prevent models overfitting, according to the specific characteristics of the classification scheme. As for the baseline GMM system, 256 Gaussian components were used to adapt both the apnoea and control models from a universal background model by means of MAP adaptation. The use of a higher number of mixtures up to 2048 for the GMM was tested without obtaining a noticeable improvement in the classification accuracy.

For the *sentence-dependent phonemes* we used a total number of 61 HMM models to represent all the sentence-dependent phonemes in the four sentences. Each HMM phoneme model used a standard left-to-right 3 states topology and the best results were obtained using 6 mixtures per state (resulting in 1098 gaussians). HMM models for *sentence-dependent phonemes* were obtained using standard MLLR + MAP for adapting a set of CI phoneme models trained from the Spanish phonetically balanced corpus of Albayzin [19], to the apnea and control groups (only records from male subjects were used as our OSA database includes only male speakers).

To establish the proper topology for the *whole sentence* HMM models a maximum number of Gaussian mixtures close to 500 was considered. Bearing this in mind, different topologies were trained using standard MAP adaptation from a background model using global cepstral mean and variance for every Gaussian, and tested to identify the most suitable one. The best results were obtained when using *whole sentence* HMMs with the same number of states equal to 40 for each sentence and 13 mixtures of Gaussians for state (resulting in a total of 2080 gaussian components).

4.1 Results

Results using the leave-one-out procedure on the three automatic severe apnoea detection systems included in this contribution are shown in Figure 1. LLRs (i.e. Log-Likelihood Ratio (apnea/control)) scores, corresponding to each speaker (40 control + 40 severe OSA speakers) while uttering each phrase (4 sentences, 3 repetitions each one) were used to estimate the plotted DET curves in the Figure (a total of 960 tests), considering all possible operating points.

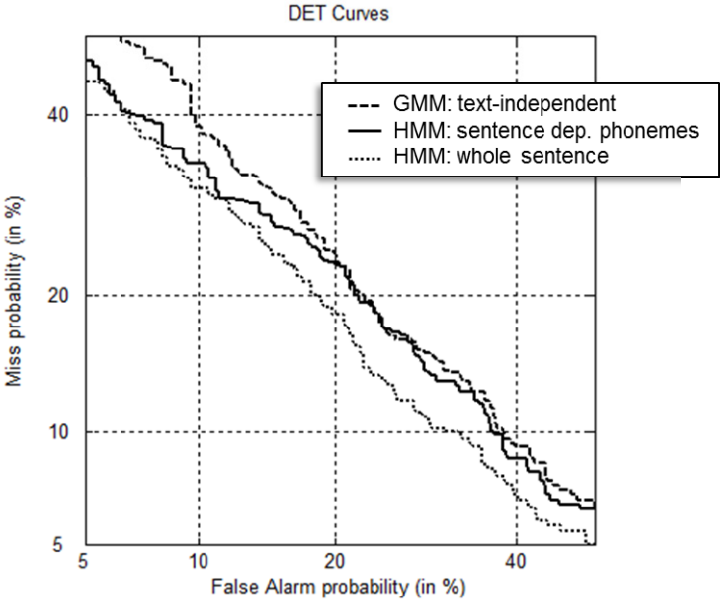


Fig. 1. DET curves corresponding to three severe apnoea detection systems: GMM-based baseline system, sentence-dependent phoneme HMMs and whole sentence HMMs

As we can see in Figure 1 the initial classification results obtained with the text-independent GMM models, which mostly disregard the linguistic context information, were enhanced as context information modeled by HMMs was added. So, when compared to the GMM baseline system, a relative moderate improvement is achieved using sentence-dependent phoneme-based HMM models. However, better results are achieved for whole sentence HMM models with 40 states, producing an overall 11.6% relative reduction in Equal Error Rate (EER) (see second column in Table 1). In order to have a global figure-of-merit using all the available audio data for each speaker in our database, LLR scores obtained from the three repetitions of the four sentences were fussed (i.e. averaged) for every speaker and used, again following a leave-one-out protocol, for severe OSA / control classification. ERRs for each classification scheme are presented in the third column of Table 1. These results show that the best classification performance is again achieved when using the whole sentence HMM classifier, providing a 25% relative reduction in the EER value compared to the baseline GMM system.

Table 1. EER values obtained for each classification scheme

Models	**EER on one sentence**	**EER combining all sentences**
GMM	21,46%	20,00%
sentence-dependent phonemes HMMs	21,25%	17,50%
whole sentence HMMs	18,96%	15,00%

5 Conclusions and Future Research

Experimental results on the discrimination of apnoea voices are presented over a database including both severe OSA and healthy speakers. This study offers an innovative perspective on how HMMs can be used to model CD phonetic information for detecting voices of speakers suffering from severe OSA syndrome. MFCC-based parameterization was used in all three apnoea detection systems compared in this contribution, just as it was done in [13-15]. From this acoustic representation, experimental results were obtained using HMMs for modeling both sentence-dependent phonemes and the whole sentence. Additionally, a GMM-based baseline system was also evaluated on this same task to compare the results. After analyzing all the experimental results we can conclude that CD models outperform the text-independent baseline GMM system. The highest correct classification rate of 85% is achieved when using whole-sentence HMMs. These sentence-dependent models seem to be able to combine the potential of a flexible representation for specific linguistic contexts in each sentence together with an efficient training procedure with a limited amount of training data. Further analysis should be carried out to confirm these results on a larger database, as well as to explore the use of other alternative HMM modeling schemes (such as triphones) and to explicitly exploit the different discriminative power of different linguistic contexts.

Acknowledgements. The activities described in this paper were funded by the Spanish Ministry of Science and Technology as part of the TEC2009-14719-C02-02 (PriorSpeech) project. The authors would like to thank Christopher Gaul for his contribution in the references.

References

1. Eliot, S., Janita, L., Cheryl, B., Carole, L.: Transit time as a measure of arousal and respiratory effort in children with sleep-disorder breathing. Pediatric Research 53, 580 (2003)
2. Redline, S., Kump, K., Tishler, P.V., Browner, I., Ferrette, V.: Gender differences in sleep disordered breathing in a community-based sample. American Journal of Respiratory and Critical Care Medicine 149, 722 (1994)
3. Block, A.J., Boysen, P.G., Wynne, J.W., Hunt, L.A.: Sleep apnea, hypopnea and oxygen desaturation in normal subjects. New England Journal of Medicine 300, 513 (1979)
4. Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., Badr, S.: The occurrence of sleep-disordered breathing among middle-aged adults. New England Journal of Medicine 328, 1230 (1993)
5. Bahammam, A., Delaive, K., Ronald, J., Manfreda, J., Roos, L., Kryger, M.H.: Health care utilization in males with obstructive sleep apnea syndrome two years after diagnosis and treatment. Sleep 22, 740 (1999)
6. He, J., Kryger, M.H., Zorick, F.J., Conway, W., Roth, T.: Mortality and apnea index in obstructive sleep apnea. Experience in 385 male patients. Chest 94, 9 (1988)
7. Coccagna, G., Pollini, A., Provini, F.: Cardiovascular disorders and obstructive sleep apnea syndrome (2006)

8. Lloberes, P., Levy, G., Descals, C., Sampol, G., Roca, A., Sagales, T., de la Calzada, M.D.: Self-reported sleepiness while driving as a risk factor for traffic accidents in patients with obstructive sleep apnoea syndrome and in non-apnoeic snorers (2000)
9. Puertas, F., Pin, G., María, J., Durán, J.: Documento de consenso nacional sobre el síndrome de apneas-hipopneas del sueño (SAHS), grupo Español De Sueño, GES (2005)
10. Ayappa, I., Rapoport, D.M.: The upper airway in sleep: physiology of the pharynx (2003)
11. Lan, Z., Itoi, A., Takashima, M., Oda, M., Tomoda, K.: Difference of pharyngeal morphology and mechanical property between OSAHS patients and normal subjects. Auris Nasus Larynx 33, 433 (2006)
12. Fox, A.W., Monoson, P.K., Morgan, C.D.: Speech dysfunction of obstructive sleep apnea. A discriminant analysis of its descriptors. Chest 96, 589 (1989)
13. Blanco, J.L., Fernández, R., Pardo, D.D., Hernández, L., López, E., Toledano, D.T.: Apnoea voice characterization through vowel sounds analysis using Generative Gaussian Mixture Models. AVFA (2009)
14. Fernández, R., Blanco, J.L., Hernández, L., López, E., Alcázar, J., Toledano, D.T.: Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. EURASIP J. Adv. Signal Process 6, 1 (2009)
15. Fernández, R., Blanco, J.L., Pardo, D.D., Hernández, L.A., López, E., Alcázar, J.: Early Detection of Severe Apnoea through Voice Analysis and Automatic Speaker Recognition Techniques. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2009. CCIS, vol. 52, pp. 245–257. Springer, Heidelberg (2010)
16. Fernández, R., Hernández, L.A., López, E., Alcázar, J., Portillo, G., Toledano, D.T.: Design of a multimodal database for research on automatic detection of severe apnoea cases. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), European Language Resources Association (ELRA), Marrakech (2008)
17. Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.4. Cambridge University Press (2006)
18. Blouet, R., Mokbel, C., Mokbel, H., Sanchez Soto, E., Chollet, G., Greige, H.: BECARS: A Free Software for Speaker Verification, pp. 145–148. ODYSSEY (2004)
19. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Nadeu, C.: Albayzín speech database: Design of the phonetic corpus. In: Proceedings of the 3rd European Conference on Speech Communication and Technology, Eurospeech 1993, vol. I, p. 175 (1993)

Acoustic Analysis of European Portuguese Oral Vowels Produced by Children

Catarina Oliveira[1], Maria Manuel Cunha[1], Samuel Silva[2], António Teixeira[2], and Pedro Sá-Couto[3]

[1] Escola Superior de Saúde / IEETA, Univ. de Aveiro, Portugal
[2] Dep. de Eletrónica, Telec. e Informática / IEETA, Univ. de Aveiro, Portugal
[3] Dep. de Matemática / CIDMA, Univ. de Aveiro, Portugal
{coliveira,sss,ajst,p.sa.couto}@ua.pt,
mariamanuelmoc@gmail.com

Abstract. This study investigates acoustic changes in the speech of European Portuguese children, as a function of age and gender. Fundamental frequency, formant frequencies and duration of vowels produced by a group of 30 children, ages 7 and 10 years, were measured. The results revealed that, for male speakers, F0, F1 and F2 decrease as age increases, although the age effect was not statistically significant for F0 and F1. A similar trend was observed for female speakers, but only in F2. Moreover, F0 and formant frequencies were found to be similar between male and female children. Between ages 7 and 10, vowel durations decreased significantly, and the values for females were higher than those for males. These results provide a base of information for establishing the normal pattern of development in European Portuguese children.

1 Introduction

There has been great interest in the study of acoustic characteristics of children's speech. The knowledge on the age-dependent changes of acoustic parameters (such as fundamental frequency, formant frequencies and segmental durations) has important implications to the development of speech applications suitable for children's voices [1,2] and to clinical assessment of speech disorders.

Several acoustic studies [3,4,5,6,7,8,9] have examined the effects of age and gender in the fundamental frequency (F0) and formant frequencies (F1 and F2) values. In general, these investigations showed that values of the three acoustic parameters decrease as age increases. Some authors [7,10,11] also reported longer segmental durations and greater spectral variability in children's speech than in adult's speech.

As regards the beginning of gender-related differences in F0 and formant frequencies, results of previous studies are not fully consensual. Busby and Plant [5] and Whiteside and Hodgson [6] reported gender differences in F0 and/or formant frequencies from the age of five, while Lee et al. [7] and Perry et al. [12] did not find significant differences related to gender in acoustic parameters until after

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 129–138, 2012.

eleven years old. Likewise, anatomical studies of the child vocal tract [13] only found differences in vocal tract morphology, between genders, starting at age ten.

Acoustic characteristics of children who speak languages other than English have rarely been investigated [8]. To our knowledge, no research on the acoustic analysis of European Portuguese (EP) children's speech has been published. The few studies currently available [14,15] provide only information on the acoustic characteristics of vowels produced by adult speakers.

The purpose of this study is to acoustically examine the vowels of EP produced by children of 7 and 10 years of age. Fundamental frequency (F0), first two formant frequencies (F1–F2) and vowel durations are measured and analysed as a function of age and gender. The acoustic parameters of EP children are also compared with adults data obtained from a study by Escudero et al. [15].

This study provides additional information on F0 and formant frequency characteristics of children who speak a language with a vowel space different from English and, in that sense, might help to better understand cross-linguistic similarities and language-particular features of vowel development.

This paper is organized as follows. Section 2 describes the adopted method, sections 3 and 4 present and discuss the obtained results and section 5 presents conclusions and ideas for future work.

2 Method

2.1 Participants

Thirty children, divided in two age groups (7 and 10 years old), were recruited from an elementary school in Aveiro (north of Portugal) for this study. For each age group there were seven boys and eight girls, so that the gender-dependence of the vowels could also be investigated. All children were native speakers of EP, with no history of hearing or speech disorders. Parents/guardians of children provided written consent before data collection.

2.2 Speech Materials

All experimental procedures were conducted in a quiet room in the children´s school. Voice samples were recorded at a sampling rate of 22 kHz with a headset condenser microphone connected to an external 24-bit sound system (Roland UA-25 EX Cakewalk).

Children produced each EP oral vowel ([a], [ɛ], [e], [i], [ɔ], [o], [u]) in a disyllabic CVCV sequence (e.g. ['**p**i**p**i]), where C was an identical voiceless stop consonant ([p], [t] or [k]). The nonce words were embedded in a carrier sentence "Digo ... para ti." ("I say ... to you").

The corpus is based on the one presented in a recent acoustic study of vowels produced by adult speakers of European and Brazilian Portuguese [15], in order to allow the comparison of results.

2.3 Data Collection

The sentences were presented to the children orthographically, on a computer screen, using the software tool ProRec [16]. The tester familiarized the participants with the materials and procedures prior to data acquisition. This was particularly important for the youngest subjects given their reduced reading experience. The children were encouraged to read the test items at a comfortable pitch and loudness level. If the children hesitated or misread a sentence, they were asked to repeat it.

Although more than thirty children were initially recorded, some voices with problems, i.e breathiness, roughness or muting (for boys), were excluded by a certified speech therapist (the second author), after evaluation of the voice samples. The children repeated each item three times. Thus, a total of 630 productions (30 participants $\times$ 7 vowels $\times$ 3 repetitions) were analysed.

2.4 Speech Analysis

Each vowel and flanking consonants were manually segmented and labelled, over the digitized sound wave, by using the software Praat [17].

Acoustic Analysis. The total duration was computed from the label files that contained the start and end points of each vowel. The consonantal context (voiceless stops) allowed that the onset and offset of the vowels could be easily determined.

The fundamental frequency (F0) of the seven vowels was estimated with the cross-correlation method available in the software Praat. Median F0 value was taken of the central 40 percent of each vowel.

Burg-LPC algorithm, as provided by Praat, was used to compile values for F1 and F2, at the central 40 percent of the vowel. The frequency range initially used was 50 Hz to 7500 Hz. This strategy yielded several unlikely values for some formants. Therefore, a procedure, adapted from Escudero et al. [15], was applied to optimize the formant ceiling for a certain vowel of a certain speaker. The first two formants were determined several times, for all ceilings between 5500 and 7500Hz in steps of 100 Hz. The chosen ceiling was the one that yielded the lowest variation (for more details see [15]). With this method, almost all outliers were removed.

Statistical Analysis. The statistical analysis was conducted with the SPSS software package (SPSS 19.0 – SPSS Inc., Chicago, IL, USA). For each dependent variable (F0, F1, F2, and duration), a three-way mixed analysis of variance (ANOVA) was performed, with vowel as within-subject factor and gender and age as between-subject factors. The ANOVA assumptions of residual normality and homogeneity of variance were validated. As regards sphericity, Epsilon Huynh-Feldt correction was used. In all statistical analysis, the level of significance was $p < 0.05$.

3 Results

3.1 Fundamental Frequency

The mean and standard deviation (SD) of F0 for the seven EP vowels are given in Table 1A, according to age and gender. For the male speakers, F0 decreases as age increases, while for female speakers the opposite tendency is observed. However, the three-factor ANOVA indicates there were no significant age ($F(1;26) = 0.701$; $p = 0.410$) or gender ($F(1;26) = 1.1$; $p = 0.315$) differences. Only a significant interaction age by gender was found ($F(1;26) = 4.4; p = 0.046$).

Table 1A denotes the following progression in the mean F0 value of individual vowels: [u] (258.8 ± 34.0Hz) followed by [i] (257.7 ± 34.0Hz), [o] (251.2 ± 33.4Hz), [e] (250.9 ± 32.5Hz), [ɔ] (242.6 ± 31.8Hz), [ɛ] (241.5 ± 32.7Hz) and [a] (239.0 ± 31.3Hz). There is a significant effect of vowels on F0 ($F(5.1;133.6) = 65.9$; $p < 0.001$).

As expected, mean F0 values are higher than those computed for adult speakers [15].

3.2 Formant Frequencies

Mean formant frequency values for F1 and F2, averaged across male and female subjects, for each vowel and age group, are provided in Table 1B and 1C. Figure 1 shows the mean F1 and F2 values for the boys and girls of each age group.

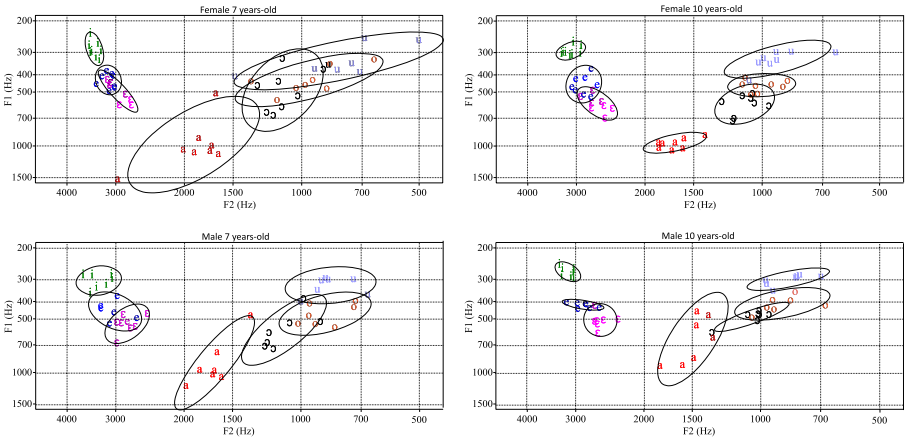


Fig. 1. First and second formants of the EP vowels produced by the seven females and eight males (age 7 on the left; age 10 on the right)

It can be observed, from Table 1B, that F1 values decrease with an increase in age, but only for male speakers. For female speakers, the F1 values are, in general, higher for 10-year-olds than for 7-year-olds, as it happens with F0. The exceptions are the F1 values for [a] and [u]. No statistical differences were found

for age ($F(1;26) = 2.5; p = 0.126$), but there was a significant gender effect ($F(1;26) = 4.4; p = 0.047$).

The effect of vowel is also statistically significant ($F(1.9;48.6) = 140.70; p = 0.001$). The mean value of F1 is higher for [a] (898.0 ± 237.0Hz), followed by [ɔ] (542.2 ± 106.8Hz), [ɛ] (531.8 ± 69.1Hz), [o] (444.1 ± 57.5Hz), [e] (439.5 ± 45.7Hz), [u] (324.6 ± 43.8Hz) and [i] (290.2 ± 25.9Hz).

There are significant age by gender ($F(1;26) = 8.1; p = 0.009$) and vowel by gender ($F(1.9;48.6) = 4.5; p = 0.017$) interactions.

As shown in Table 1C, F2 values for 7-year-olds (2002.4 ± 94.3Hz) are higher than those for 10-year-olds (1877.3 ± 94.3Hz), for the majority of vowels. The two exceptions are the vowels [o] and [u], for both male and female children. This F2 drop is statistically significant ($F(1;26) = 13.2; p = 0.001$).

In general, female speakers of both age groups exhibit high F2 values. The F2 differences between males and females are statistically significant($F(1;26) = 5.7; p = 0.024$).

The effect of vowel on F2 is statistically significant ($F(3.6;93.0) = 1205.6; p = 0.001$). The mean value of F2 is higher for vowel [i] (3244.4 ± 189.4Hz) followed by [e] (2978.5 ± 199.0Hz), [ɛ] (2758.7 ± 211.4Hz), [a] (1705.5 ± 298.8Hz), [ɔ] (1104.2 ± 118.5Hz), [o] (940.0 ± 149.4Hz) and [u] (866.9 ± 174.2Hz).

There is a significant vowel by age ($F(3.6;93.0) = 4.863; p = 0.002$) interaction.

Standard deviations in Tables 1B and 1C suggest that variability decreases with age, for both F1 and F2. This tendency is also clearly illustrated in Figure 1, especially for female children.

Figure 2 depicts the mean F1 and F2 values for the seven vowels, for male and female speakers, ages 7 and 10.

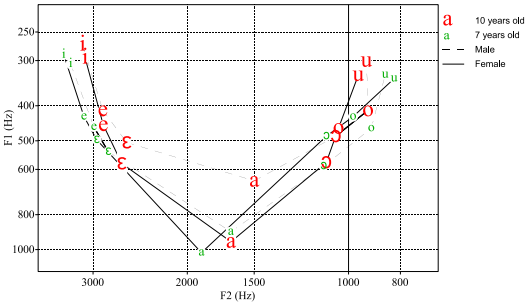


Fig. 2. Average F1 and F2 according to gender and age

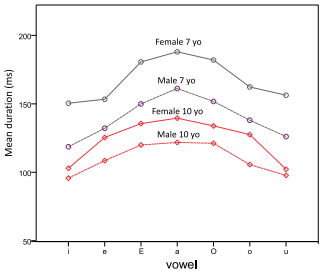


Fig. 3. Mean duration of vowels for each subject group (gender and age)

3.3 Duration

Mean duration of each EP vowel is depicted in Figure 3 in terms of age and gender. It can be observed that female subjects (145.4 ± 28.4ms) present higher vowel durations than male subjects (123.6 ± 28.5ms). ANOVA shows that the effect of gender on vowel duration is statistically significant ($F(1;26) = 4.4; p = 0.047$).

Seven-year-old children (152.1 ± 28.5ms) also display longer vowel durations than the older group (116.8 ± 28.5ms). Effect of age is significant ($F(1;26) = 11.5; p = 0.002$).

The effect of vowel on mean duration is also statistically significant ($F(6;156) = 64.4; p = 0.001$). The following pattern of vowel duration is observed: [a] (154.6 ± 40.3ms) > [ɔ] (147.3 ± 39.8ms) > [ɛ] (146.0 ± 35.8ms) > [o] (133.0 ± 35.4ms) > [e] (129.9 ± 27.9ms) > [u] (120.3 ± 36.0ms) >[i] (115.4 ± 34.5ms).

Despite the higher duration values, when compared to those given by Escudero et al. [15] for adults, similar vowel duration patterns are observed.

4 Discussion

4.1 Fundamental Frequency

In general, previous investigations [4,5,6,7,18] revealed that F0 decreases as age increases. Although not statistically significant, the current study also found a similar trend for male children. In contrast, the older group of female children showed a higher F0 than the younger group. Hasek et al. [3] have also reported a decrease in F0 for boys between 5 and 10 years old, but not for girls with the same age. It is possible that the beginning of pitch change may be different for male and female subjects.

The finding that F0 is similar for male and female children at 7 and 10 years is consistent with most of previous acoustic studies of children's speech, that have shown little F0 differences in children under 12 years of age [5,7,12]. This may be indicative of a modest growth of the vocal folds during pre-puberty [13]. Hasek et al. [3] and Whiteside and Hodgson [6] found that gender-related F0 differences begin to emerge earlier, but they examined only the /a/ vowel, while the other studies focused on F0 for all vowels.

The presented study revealed a tendency for the high vowels to have a higher fundamental frequency than the low vowels, as previously evidenced by Costa [19] and Escudero et al. [15] for adult speakers. This result suggests that children acquire the ability to control intrinsic F0 [20] relatively early [7].

4.2 Formant Frequencies

The values of F1 decreased with increase in age, but only for male speakers. Furthermore, F2 frequencies for older children were, in general, lower than those for younger children, both for male and female groups. Results for male subjects are in general agreement with those published in the literature [5,7,9]. For female

Table 1. Mean and standard deviation of vowel F0, F1 and F2 (values in Hz±SD) by gender (G; M-male, F-female) and age (A; 7-group of 7 years old, 10- group of 10 yeras old)

A – Mean and standard deviation of vowels' fundamental frequency (F0) by age and gender.

G	A	n	[a]	[e]	[ɛ]	[i]	[o]	[ɔ]	[u]
F	7	8	227.7 ± 31.4	238.9 ± 32.8	226.9 ± 34.3	245.6 ± 37.4	241.0 ± 39.8	230.4 ± 32.2	243.3 ± 35.0
	10	8	239.3 ± 22.2	253.5 ± 23.4	241.8 ± 24.0	260.8 ± 21.5	250.8 ± 22.8	243.5 ± 24.5	264.0 ± 22.4
M	7	7	260.9 ± 35.7	273.4 ± 37.2	264.9 ± 37.9	281.7 ± 37.9	274.1 ± 35.3	263.9 ± 34.9	282.1 ± 39.1
	10	7	229.8 ± 30.0	239.1 ± 29.3	234.2 ± 26.6	243.9 ± 29.4	240.3 ± 28.2	234.1 ± 30.5	247.2 ± 29.6

B – Mean and standard deviation of vowels' first formant (F1) by age and gender.

G	A	n	[a]	[e]	[ɛ]	[i]	[o]	[ɔ]	[u]
F	7	8	1020.1 ± 278.4	429.7 ± 41.0	498.0 ± 67.9	288.2 ± 29.9	433.2 ± 71.9	502.1 ± 128.2	340.0 ± 58.0
	10	8	959.5 ± 63.8	453.3 ± 53.7	580.3 ± 62.5	293.3 ± 17.1	457.4 ± 34.7	587.4 ± 79.0	329.2 ± 43.3
M	7	7	912.2 ± 230.0	457.6 ± 56.3	536.8 ± 71.5	306.4 ± 29.5	471.9 ± 65.3	589.0 ± 125.0	326.3 ± 40.5
	10	7	674.0 ± 199.3	416.7 ± 14.5	509.8 ± 53.0	272.6 ± 17.3	413.8 ± 61.3	489.5 ± 49.1	300.5 ± 22.5

C – Mean and standard deviation of vowels' second formant (F2) by age and gender.

G	A	n	[a]	[e]	[ɛ]	[i]	[o]	[ɔ]	[u]
F	7	8	1928.3 ± 438.2	3134.3 ± 123.2	2962.6 ± 171.8	3403.1 ± 92.8	977.3 ± 211.0	1126.4 ± 128.4	840.8 ± 289.2
	10	8	1683.3 ± 152.3	2876.1 ± 152.0	2656.3 ± 159.8	3104.5 ± 134.5	1007.8 ± 99.2	1113.3 ± 98.2	922.6 ± 132.4
M	7	7	1679.6 ± 194.5	3018.6 ± 233.8	2807.2 ± 180.9	3322.5 ± 222.0	873.7 ± 116.3	1113.1 ± 134.2	846.2 ± 110.6
	10	7	1502.1 ± 163.7	2877.3 ± 178.8	2594.1 ± 124.0	3144.9 ± 120.6	886.2 ± 120.5	1059.6 ± 127.3	853.6 ± 103.0

subjects, the growth of the vocal tract between 7 and 10 years of age seems to reflect only in F2 frequencies. Significant gender differences were noticeable for both F1 and F2 values.

As can be seen in Figure 2, gender and age differences are clearer and more consistent in the EP vowel [a]. Similar patterns were found in English and Korean-speaking children [5,8]. A possible explanation for the low vowels to be more sensitive to age- and gender- related distinctions is that these vowels are produced with a more open vocal tract and, therefore, the acoustic output would reflect the dimensions of the entire tube [8].

Comparing our results with the data from Escudero et al. [15], it is observed that the formant values of vowels produced by children of ages 10 are almost in adults range.

4.3 Duration

In this study, the group of age 7 showed significantly longer mean vowel durations than the older age group. Although this might be partially due to reading abilities, since younger children attended their first school year, the results obtained are consistent with the study of Lee et al. [7], that reported reduction of vowel duration as age increased.

Furthermore, the female speakers exhibited greater vowel durations than male speakers. These data are in agreement with those in Escudero et al. [15], Botinis et al. [21] and Hillenbrand et al. [11] for adult speakers. Although not statistically significant, Lee et al. [7] also found a similar trend for children. On the other hand, Rauber [22] did not find any significant difference in vowel duration due to gender. Ericsdotter and Ericsson [23] concluded that women use vowel duration contrasts better, producing shorter vowels (or similar to men) in unstressed positions and longer vowels in stressed positions.

Finally, it was observed that duration depends on vowel height, i.e. lower/more open vowels are longer than high/less open vowels. The tendency for low vowels to be intrinsically longer than high vowels is considered to be a phonetic universal phenomena and has been observed for many languages, including Portuguese [14,19,22,15]. The similar vowel duration patterns between children and adults suggest that EP children, like English children [7], are able to control intrinsic vowel duration since early ages.

5 Conclusions

The purpose of this study was to analyse acoustic characteristics of speech collected from children 7 to 10 years of age, in order to provide a base of information for establishing the normal pattern of development in EP children. Although not statistically significant, a trend for a decrease in F0 and formant frequencies with age has been observed for male speakers, which is an indicator of physical development of the voice source and vocal tract. The unexpected increase in F0 and F1, between the 7 and 10-year-olds, for female speakers, is not readily explained,

but may be associated with differences in the rate of growth of the larynx between male and female subjects [9]. Only further investigations, including additional acoustic data and direct articulatory measurements, could give clearer answers to this question.

This study also revealed a significant effect of age in vowel duration, consistent with previous findings [7]. However, it is likely that this result might be mainly determined by the elicitation method used, since young children have shown limited reading abilities. It would be important to use different speech-elicitation methods in future works.

As demonstrated by several studies (e.g. [13,7]), a rapid decrease in F0 and formant frequencies is observed after the ages included in this investigation, during adolescence (more so in males than females), and remarkable differences in male-female F0 and formant frequency patterns are evident by age 12. Therefore, more detailed data obtained from a larger number of subjects with a wider age range will be necessary, in order to fully understand developmental acoustic patterns of EP children and their relation to the underlying anatomical development.

Although several questions in EP speech development remain unanswered, the current study nevertheless provides some preliminary data on the normal pattern of development in EP children, against which disordered vowel productions can be compared.

Acknowledgements. This research was partially funded by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology in the context of the project FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011). We express our sincere gratitude to the children and teachers of Agrupamento de Escolas de Esgueira (Aveiro, Portugal), who participated in the study. Authors also thank Paul Boersma and Andréia Schurt Rauber for their help with the automatic estimation of the acoustic parameters.

References

1. Potamianos, Narayanan, S.: Robust recognition of children's speech. IEEE Transactions on Speech and Audio Processing 11(6), 603–616 (2003)
2. Giuliani, D., Gerosa, M.: Investigating recognition of children's speech. In: Proc. ICASSP, pp. 137–140 (2003)
3. Hasek, C.S., Singh, S., Murry, T.: Acoustic attributes of preadolescent voices. Journal of the Acoustical Society of America 68(5), 1262–1265 (1980)
4. Bennett, S.: Vowel formant frequency characteristics of preadolescent males and females. Journal of the Acoustical Society of America 69(1), 231–238 (1981)
5. Busby, P.A., Plant, G.L.: Formant frequency values of vowels produced by preadolescent boys and girls. Journal of the Acoustical Society of America 97(4), 2603–2606 (1995)
6. Whiteside, S.P., Hodgson, C.: Acoustic characteristics in 6- 10- year- old children's voices: some preliminary findings. Logopedics Phoniatrics Vocology 24(1), 6–13 (1999)

7. Lee, S., Potamianos, A., Narayanan, S.: Acoustics of children's speech: developmental changes of temporal and spectral parameters. Journal of the Acoustical Society of America 105(3), 1455–1468 (1999)
8. Lee, S., Iverson, G.K.: The development of monophthongal vowels in Korean: age and sex differences. Clinical Linguistics and Phonetics 22(7), 523–536 (2008)
9. Vorperian, H.K., Kent, R.D.: Vowel acoustic space development in children: a synthesis of acoustic and anatomic data. Journal of Speech, Language and Hearing Research 50, 1510–1545 (2007)
10. Most, T., Amir, O., Tobin, Y.: The hebrew vowel system: raw and normalized acoustic data. Language and Speech 43(3), 295–308 (2000)
11. Hillenbrand, J., Getty, L., Clark, M., Wheeler, K.: Acoustical characteristics of American English vowels. Journal of the Acoustical Society of America 97, 3099–3111 (1995)
12. Perry, T.L., Ohde, R.N., Ashmead, D.H.: The acoustic bases for gender identification from children's voices. Journal of the Acoustical Society of America 109(6), 2988–2998 (2001)
13. Vorperian, H.K., Kent, R.D., Lindstrom, M.J., Kalina, C.M., Gentry, L.R., Yandell, B.S.: Development of vocal tract length during early childhood: A Magnetic Resonance Imaging study. Journal of the International Phonetic Association 117, 338–350 (2005)
14. Martins, M.R.D.: Análise acústica das vogais tónicas em Português. Boletim de Filologia XXII, 303–314 (1973)
15. Escudero, P., Boersma, P., Rauber, A.S., Bion, R.A.H.: A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. Journal of the Acoustical Society of America 126(3), 1379–1393 (2009)
16. Huckvale, M.: Speech prompt & record system (ProRec) (2009)
17. Boersma, P., Weenink, D.: Praat: doing phonetics by computer. Computer Program (2010)
18. Viegas, F., Viegas, D., Atherino, C.C.T., Baeck, H.E.: Frequência fundamental das 7 vogais orais do Português em vozes de crianças. Revista CEFAC 12(4), 563–570 (2010)
19. Costa, F.: Intrinsic prosodic properties of stressed vowels in European Portuguese. In: Proc. 2nd Int. Conf. on Speech Prosody, Nara, Japan (2004)
20. Whalen, D.H., Levitt, A.G.: The universality of intrinsic F0 of vowels. Journal of Phonetics 23, 349–366 (1995)
21. Botinis, A., Fourakis, M., Panagiotopoulou, N., Pouli, K.: Greek vowel durations and prosodic interations. Glossologia 13, 101–123 (2001)
22. Rauber, A.S.: An acoustic description of Brazilian Portuguese oral vowels. Diacrítica, Ciências da Linguagem 22(1), 229–238 (2008)
23. Ericsdotter, C., Ericsson, A.M.: Gender differences in vowel duration in read Swedish: preliminary results. Working Papers, vol. 49 (2001)

Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese

Thomas Pellegrini[1], Isabel Trancoso[1,2], Annika Hämäläinen[3,4], António Calado[3], Miguel Sales Dias[3,4], and Daniela Braga[3,4]

[1] INESC-ID Lisboa
R. Alves Redol, 9, 1000-029 Lisbon, Portugal
thomas.pellegrini@inesc-id.pt
https://www.l2f.inesc-id.pt/wiki/index.php/Thomas_Pellegrini
[2] Instituto Superior Técnico, Lisbon, Portugal
[3] Microsoft Language Development Center, Lisbon, Portugal
[4] ADETTI ISCTE, IUL, Lisbon, Portugal

Abstract. Standard automatic speech recognition (ASR) systems use acoustic models typically trained with speech of young adult speakers. Ageing is known to alter speech production in ways that require ASR systems to be adapted, in particular at the level of acoustic modeling. This paper reports ASR experiments that illustrate the impact of speaker age on speech recognition performance. A large read speech corpus in European Portuguese allowed us to measure statistically significant performance differences among age groups ranging from 60- to 90-year-old speakers. An increase of 41% relative (11.9% absolute) in word error rate was observed between 60-65-year-old and 81-86-year-old speakers. This paper also reports experiments on retraining acoustic models (AMs), further illustrating the impact of ageing on ASR performance. Differentiated gains were observed depending on the age range of the adaptation data use to retrain the acoustic models.

Keywords: ASR, Portuguese, Elderly Speech.

1 Introduction

European countries, in particular Western European countries, are about to face a significant social change, brought by an unprecedented demographic change: the ratio of older people is steadily growing, while the ratio of younger people is shrinking. Between 2010 and 2030, the number of people aged 65 and over is expected to rise by nearly 30%-40% relative (according to the statistics of the European Commission from 2010).

Most elderly people would like to live in their own homes as long as possible ("ageing in place"). Thus, research and development of new technologies adapted to older people are becoming strategical, in order to increase their autonomy and independence. Due to the ageing process and the changes that come with it, this

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 139–147, 2012.

population faces specific difficulties to interact with computers and machines. To overcome this issue, speech appears to be the most natural and effective modality. Thus, speech recognition for the elderly is a key technology in many R&D projects related to the *Ageing Well* problematic.

Due to both cognitive and physiological age-related changes, elderly speech shows specific characteristics that make its processing significantly harder when using models built using speech from younger people. In particular, automatically recognizing the speech of older people is known to be challenging compared with automatically recognizing the speech of younger people, with performance decreases of around 9-12% absolute [1,2,3]. Various reasons are presented in the literature: ageing causes changes in the speech production mechanism, altering the vocal chords, the vocal cavities and the lungs; it also causes a decline in cognitive and perceptual abilities [4,5]. Seniors may also interact with machines in a different way than younger speakers do, by using everyday language and their own words to issue commands, even when instructions with a required syntax are given [6].

In the framework of an ongoing national Portuguese project named "AVoz"[1], an in-depth study of ASR for the elderly is conducted in order to improve the global performance in European Portuguese (EP). The goal of this paper is to illustrate the impact of age on ASR performance. Experiments on a large read speech corpus of elderly speech collected by the Microsoft Language Development Center (MLDC) from Lisbon[2] are reported. After an overview of our ASR system for EP, the MLDC elderly speech corpus is briefly described in Section 3. In Section 4, ASR results achieved on this database are reported.

2 Overview of Our ASR System

Our automatic speech recognition engine named Audimus [7,8] is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs). The MLPs perform a phoneme classification by estimating the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated to the single state of context independent phoneme HMMs.

Specifically, the system combines three MLP outputs trained with Perceptual Linear Prediction (PLP) features (13 static + first derivative), log-RelAtive SpecTrAl (RASTA) features (13 static + first derivative) and Modulation SpectroGram (MSG) features (28 static) [9]. Each MLP classifier incorporates two fully connected non-linear hidden layers. The number of units of each hidden layer as well as the number of softmax outputs of the MLP networks differs for every language. Usually, the hidden layer size depends on the amount of training data available, while the number of MLP outputs depends on the characteristic

[1] http://avoz.l2f.inesc-id.pt
[2] http://www.microsoft.com/pt-pt/mldc

phonetic set of each language. Finally, the decoder is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition, that maps observation distributions to words.

The baseline ASR system used in this work is exactly the ASR system for EP described in [10]. The acoustic models were initially trained with 46 hours of manually annotated broadcast news (BN) data collected from the public Portuguese TV, and in a second time with 1000 hours of data from news shows of several EP TV channels automatically transcribed and selected according to a confidence measure threshold (non-supervised training). The EP MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three state monophones of the EP language plus a single-state non-speech model (silence) and 385 phone transition units which were chosen to cover a very significant part of all the transition units present in the training data. Details on phone transition modeling with hybrid ANN/HMM can be found in [11].

The Language Model (LM) is a statistical 4-gram model that was estimated from the interpolation of several specific LMs: in particular a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts, collected from the Web from 1991 to 2005, and a backoff 3-gram LM estimated on a 531k word corpus of broadcast news transcripts. The final language model is a 4-gram LM, with Kneser-Ney modified smoothing, 100k words (or 1-gram), 7.5M 2-gram, 14M 3-gram and 7.9M 4-gram. The multiple-pronunciation EP lexicon includes about 114k entries.

These models, both AMs and the LM, were specifically trained to transcribe BN data. The Word Error Rate (WER) of our current ASR system is under 20% for BN speech in average: 18.4% obtained in one of our BN evaluation test sets (RTP07), composed by six one hour long news shows from 2007 [10].

Table 1. Number of speakers and speech durations according to the age ranges in the all corpus (after removing speakers with less than 2min of speech)

Age	# Speakers	Duration (h)
60-65	371	64.1
66-70	183	31.9
71-75	155	28.3
76-80	87	15.4
81-85	55	10.2
86-90	27	5.0
91-95	2	0.3
96-100	1	0.2

3 Elderly Speech Corpus

The speech corpus is comprised of about 150 hours of read speech (including silences) that was collected by MLDC. A total of 1038 speakers between 60 and

100 years of age read up to 160 prompts among a broad variety of prompts, from isolated digits to phonetically rich sentences. On average, this corresponds to 12 minutes of speech per speaker. For this work, speakers with less than 2 minutes of speech were removed from our datasets, so that the total speaker number was 881. Speaker age information is reported using 5-year ranges: 60-65, 66-70 and so on. Many more female than male speakers were recorded: 641 and 240 respectively. The number of speakers and the duration of the recordings according to the age ranges are presented in Table 1. Speakers in the 60-65 age range were the most numerous ones with a total of 64 hours of recordings, whereas only 5 hours were collected from speakers in the 86-90 age range. The corpus also provides speech from younger speakers, but with no precise information about their age (indication of a 0-59 age range), hence this data was not used in this work.

A test set comprised of about 10% of the corpus, totaling 15h of speech, was randomly selected. Speakers from this subset do not appear in the rest of the corpus. The proportions of the age range and gender in the full corpus were respected. Speech from the last two age ranges (91-95 and 96-100) was not considered since the corresponding durations were much shorter than for the other age ranges. Table 2 summarizes the characteristics of the subset.

Table 2. Test subset. Number of speakers and Speech durations according to the age ranges.

Age	# Speakers	Duration (h)
60-65	35	6h22
66-70	18	3h04
71-75	17	2h49
76-80	10	1h34
81-85	6	1h05

4 Results

In this section, performance results are reported, first gathered with our baseline system, second with the same system but with several sets of acoustic models that were adapted to each age range. The Out-Of-Vocabulary (OOV) rate with the 100K word vocabulary was 0.65% and the perplexity estimated with the 4-gram LM was 150 for the test set.

4.1 Age Impact on the Baseline System Performance

Table 3 presents the WERs obtained with our baseline system. For the entire test set, the WER was 35.3%. As stated earlier, the same system achieved a 18.4% WER with BN speech that, generally speaking, is much more difficult to transcribe than read speech. The much higher WER observed with the present corpus may be explained by the inappropriate LM that is suited for BN data

and not for this corpus, which is comprised of a diversity of prompts. Another reason may be the discrepancy of the AMs due to the age mismatch between the speech used to train the baseline MLPs and the elderly speech.

The difference in WER between male and female speakers, 33.5% and 36.0% respectively, was not found to be statistically significant by a one-sided t-test that gave a *p-value* of 0.5539. The greater diversity of female speakers may explain this difference.

Finally, the bottom part of the table reports the WERs according to the subsets of the test data distinguished by the age range of the speakers. A clear increase in WER can be observed with increasing speaker age. One-sided t-tests were performed to assess statistical significance of the WER differences. The alternate hypothesis was: 'the true difference in means is less than 0' between the WERs of the speakers of the first age range (60-65) and the WERs of the speakers of each of the larger age ranges. A p-value of 0.6252 indicated no significant difference with the closest 66-70 age range, but much slower p-values were obtained with the larger age-range (71 and above), with values about 0.03, validating the alternate hypothesis.

Table 3. Word error rates (WER) of the baseline system on the test set. Detailed WERs on age-range subsets are given in the bottom part of the table. M: Male, F: Female speakers.

Gender	WER(%)
all	35.3
M	33.5
F	36.0
Age range	WER(%)
60-65	29.1
66-70	28.1
71-75	36.1
76-80	45.1
81-85	41.0
86-90	54.9

4.2 Impact of Specific Age MLP Retraining

In order to further investigate the impact of age on ASR performance, basic adaptation of the acoustic models was tested by simply retraining the baseline MLPs with age-specific data from the train set. All the adapted MLPs shared the same MLP structure as the baseline MLP: 2 hidden layers with 2000 units each and an output layer with 500 units. All the remaining components were identical (the LM, the pronunciation lexicon and the decoding parameters).

Many prompts appear in both the adaptation ("train") and test sets. These prompts were removed from the train set used to adapt the AMs. Furthermore,

Table 4. WERs of the baseline and the six adapted systems on the test set. (AM for Acoustic Models).

System	WER(%)
Baseline	35.3
AM-60-65	31.5
AM-66-70	31.4
AM-71-75	31.1
AM-76-80	30.0
AM-81-85	30.0
AM-86-90	33.4

Table 5. P-values achieved with the MP test performed between the adapted systems

	AM-66-70	AM-71-75	AM-76-80	AM-81-85
AM-60-65	.582	.054	.001	.001
AM-66-70		.142	.001	.001
AM-71-75			.001	.001
AM-76-80				.741

the 86-90 age range was the one with the least data available: 2 hours (5 hours minus the common sentences with the test set). Experiments not reported here showed that this amount of data to retrain the MLPs led to limited improvements (the MLPs have about 5.7 million weights to re-estimate and the 500 output units need some representation in the adaptation corpus). Hence, we limited the adaptation data amount to 6 hours that was the amount of training data available for the 80-85 age range. Five sets of MLPs were adapted with 6 hours of data for the five age ranges from 60-65 to 80-85. The last one, 86-90, was adapted with the only 2h available. Each set is comprised of three MLPs for the three different feature streams (PLP, RASTA, and MSG), exactly as the baseline system.

Table 4 reports the WER of the baseline and the WERs of the six adapted systems achieved on the test set. 'AM-60-65' for example corresponds to the system where the AMs were adapted with data from 60-65 years old speakers. All the adapted MLPs showed improvement over the baseline, ranging from 10.7% to 15.0% relative. The smaller improvement observed for AM-86-90 may be explained by the smaller amount of adaptation data available for this age range (almost one-third less data).

Since all the systems were tested on the same test data, statistical significance can be assessed directly on the word outputs by a Matched Pairs Sentence-Segment Word Error (MAPSSWE or MP) test with the help of the NIST sc_stats tool. Each of the six adapted system's outputs was tested again the baseline output. All the one-to-one tests showed to be significant at the level of a 0.001 p-value. To determine whether the differences between the adapted systems were

significant, the same test was applied to each pair of adapted system word outputs. The p-values are given in Table 5. In general, the outputs of two systems adapted with data of close age ranges did not present significant differences, whereas outputs from disjoint age ranges did, with 0.001 values. This seems to confirm that using adaptation material that matches the speaker age of the test data lead to improvement.

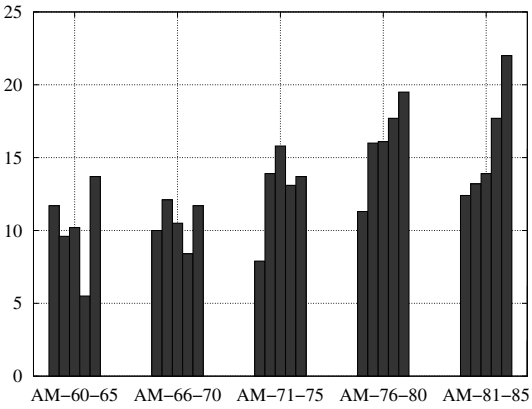


Fig. 1. Relative differences in WER between the baseline and each of the five systems with age-adapted AMs, for the five age-specific test subsets

Results are further illustrated in figure 1 where the Y-axis corresponds to the relative WER differences between the baseline and the WERs obtained with the adapted AMs. The higher the bar, the better the improvement. For each of the five age-specific adapted MLPs on the X-axis, five bars were plotted to give the

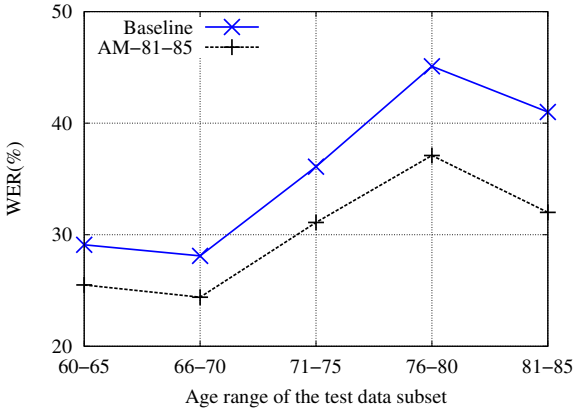


Fig. 2. Word error rates (WERs) of the baseline and one of the best adapted systems (AM-81-85) as a function of the age-range specific subsets of the test data

detail of the improvements according to the five age-specific test subsets. The results of the 86-90 range are not shown since the improvements are smaller due to less adaptation data. For each group of bars, the first one on the left corresponds to the 60-65 test subset, the first neighbor one to 66-70, etc, until the most right-handed bar that corresponds to the 81-85 test subset. As it can be observed, using adaptation data from older speakers gave better results on the test subsets with larger age ranges. For instance, AM-60-65 and AM-81-85 respectively showed 13.7% and 22.0% relative improvements over the baseline for the 81-85 test subset (5.6% and 9.0% absolute respectively). Figure 2 shows the WER points of one of the best adapted system, AM-81-85, with the baseline ones as a function of the age specific test subsets. The adapted curve globally follows the baseline one, with the largest relative gains obtained for the 66-70 and 81-85 age ranges.

5 Discussion and Future Work

In this paper, we presented ASR experiments that illustrate the impact of speaker age on ASR performance. Standard ASR systems use acoustic models typically trained with speech collected from young adult speakers. Hence, ASR performance is expected to decrease when recognizing elderly speech. The impact of aging on speech production and its consequences for ASR have already been well illustrated in the literature but this article reports results achieved on Portuguese, for which no similar study has been published to the best of our knowledge.

A large read speech corpus of European Portuguese elderly speech allowed us to measure statistically significant performance differences among different age groups with 60- to 90-year-old speakers. For instance, an increase of 41% relative (11.9% absolute) in the word error rate was observed between speakers in the 60-65 and 81-86 age groups.

To further illustrate the impact of ageing, preliminary retraining experiments showed that consistent gains in performance can be achieved by simply retraining the baseline MLPs with age-specific data. Differentiated impacts were observed according to the age range of the adaptation data. However, the limitation of these experiments lies in the fact that the adaptation data was very similar to the test data (similar prompts). Hence, additional experiments that use a completely different test set are needed to draw firmer conclusions on the impact of AM adaptation.

We plan to devise and test other adaptation techniques, for instance the adaptation of the MLP output layer alone may help in case of small amount of adaptation data. To be able to use age-specific ASR systems, one would need to detect the speaker age automatically if no *a-priori* information on it is available. Since chronological age is not a consistent indicator of ageing in speech production, other features (such as jitter and shimmer) will be investigated in order to build a classifier. Linguistic characterization of the errors observed in the ASR experiments will be performed with the objective of better understanding the special

needs of elderly speech recognition. Finally, in the long term, we plan to collect elderly speech in a Wizard-of-Oz framework in order to study the interaction of elderly people with dialog systems.

Acknowledgements. This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project PTDC/EEA-PLP/121111/2010 and under project PEst-OE/EEI/LA0021/2011.

References

1. Wilpon, J., Jacobsen, C.: A study of speech recognition for children and the elderly. In: Proc. ICASSP, Atlanta, pp. 349–352 (1996)
2. Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K.: Acoustic models of the elderly for large-vocabulary continuous speech recognition. Electronics and Communications in Japan 87(7), 49–57 (2004)
3. Vipperla, R., Renals, S., Frankel, J.: Longitudinal study of ASR performance on ageing voices. In: Proc. Interspeech, Brisbane, pp. 2550–2553 (2008)
4. Baeckman, L., Small, B., Wahlin, A.: Aging and memory: cognitive and biological perspectives. In: Handbook of the Psychology of Aging, pp. 349–377 (2001)
5. Fozard, J., Gordon-Salant, S.: Changes in vision and hearing with aging. In: Handbook of the Psychology of Aging, pp. 241–266 (2001)
6. Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R.: Recognition of elderly speech and voice-driven document retrieval. In: Proc. ICASSP, Phoenix, pp. 145–148 (1999)
7. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D.: Broadcast news subtitling system in portuguese. In: Proc. ICASSP 2008, Las Vegas, USA (2008)
8. Meinedo, H.: Audio pre-processing and speech recognition for broadcast news. Ph.D. dissertation, IST, Lisbon, Portugal (2008)
9. Meinedo, H., Caseiro, D.A., Neto, J.P., Trancoso, I.: AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.d.G.V. (eds.) PROPOR 2003. LNCS, vol. 2721, pp. 9–17. Springer, Heidelberg (2003)
10. Meinedo, H., Abad, A., Pellegrini, T., Neto, J., Trancoso, I.: The L2F Broadcast News Speech Recognition System. In: Proc. Fala, Vigo, pp. 93–96 (2010)
11. Abad, A., Neto, J.: Incorporating Acoustical Modelling of Phone Transitions in a Hybrid ANN/HMM Speech Recognizer. In: Proceedings of INTERSPEECH, Brisbane, pp. 2394–2397 (2008)

Mutual Information and Perplexity Based Clustering of Dialogue Information for Dynamic Adaptation of Language Models

Juan Manuel Lucas-Cuesta, Fernando Fernández-Martínez, Tirso Moreno, and Javier Ferreiros

Speech Technology Group, Department of Electronic Engineering, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid, Avenida Complutense 30, 28040, Madrid
{juanmak,ffm,tirsomoreno,jfl}@die.upm.es

Abstract. We present two approaches to cluster dialogue-based information obtained by the speech understanding module and the dialogue manager of a spoken dialogue system. The purpose is to estimate a language model related to each cluster, and use them to dynamically modify the model of the speech recognizer at each dialogue turn. In the first approach we build the cluster tree using local decisions based on a Maximum Normalized Mutual Information criterion. In the second one we take global decisions, based on the optimization of the global perplexity of the combination of the cluster-related LMs. Our experiments show a relative reduction of the word error rate of 15.17%, which helps to improve the performance of the understanding and the dialogue manager modules.

Keywords: Spoken Dialogue System, Language Models, Dialogue-based Information, Clustering.

1 Introduction

Statistical language model adaptation has become a current issue within the scope of Speech Technology. It aims at modifying the language model (LM) of which a speech recognition system (ASR) makes use, to improve the recognition performance. For instance we can modify a general LM to adapt it to a closed domain, trying to improve the overall response of a domain-dependent system in which the ASR is included.

There are several approaches to adapt LMs, depending on the sources of the adaptation models [3]. Perhaps the simplest one consists of a linear interpolation between LMs [6]. This approach tries to find out an accurate weight to combine a *background* LM, built with more general data, with one or several adaptation LM, usually built with more specific data.

The adaptation LMs could be estimated at each dialogue turn [9]. Dialogue systems that use dialogue-dependent LMs usually consider the semantic information of each utterance. We estimate the LMs using semantic information as well as the user intentions ellaborated by the dialogue manager [8].

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 148–157, 2012.

To learn more robust models, we group those information elements that share common features (such as the semantics or the word classes) prior to the LM estimation. To discover these relationships, several techniques such as the application of Latent Semantic Analysis (LSA) have been proposed [2].

In this work we propose two clustering techniques, using as clustering criteria two metrics derived from the Information Theory. On the one hand, the Normalized Mutual Information (NMI), previously used for the estimation of parameters of acoustic models for speech recognition [1], or for the adaptation of trigger-based LMs [5]. On the other hand, a minimization of the global perplexity of a LM obtained as the interpolation of all the clusters considered. Our aim is to reach a tradeoff between the *specificity* of having a large number of LMs related to single pieces of information, and the *robustness* of having few LMs, but trained with more data.

The rest of the paper is organized as follows. We first describe our dialogue system (Section 2), and our approaches to cluster dialogue elements (Section 3). The interpolation technique that we apply is shown in Section 4. Finally, the evaluation results are discussed in Section 5, and the conclusions of the work are drawn in Section 6.

2 Baseline Dialogue System

We have designed a user-independent, mixed-initiative dialogue system for controlling household devices. In this work we focus on the control of a Hi-Fi audio system using speech, instead of an infrared remote control.

The Dialogue Manager (DM) is based on a Bayesian Networks (BNs) solution [4] that exploits the causal relationships between the semantics of an utterance (i.e. the *dialogue concepts*), and the intention of the user (i.e. the *goals*). We will refer to both concepts and goals as *dialogue elements*. These elements have been defined by hand using expert knowledge of the application domain.

We have defined a set of 58 concepts that cover all the semantic categories in the application domain. These concepts could be classified into three sets: *actions* (22) to be executed (e.g. to play), *parameters* (16) that can be set up (e.g. the volume), and their corresponding *values* (20). We have also defined 15 goals, according to the available functionality of the Hi-Fi audio system. A concept or a goal is *present* only if it has been extracted from the recognized utterance (by the understanding module), or positively inferred (by the DM).

As an example of our definition of dialogue elements, let us consider the utterance *raise the volume to five*. The understanding module can extract the concepts PARAM_VOL ('volume'), VALUE_VOL ('five'), and ACTION_VOL ('raise'). The dialogue goal that should be inferred is MODIFY_VOLUME.

Once the ASR has recognized the input utterance, and the understanding module has extracted the concepts of that utterance, the DM has to identify the goals, using the information available (i.e. the concepts). This task is carried out by means of a *forward inference* procedure (FI), that estimates the posterior probability of each goal, given the available evidences (the presence or absence

of each concept in the history of the dialogue). By comparing the resulting probabilities with several predefined thresholds, the DM decides whether a goal is *present* or *absent*.

After the FI process, the DM estimates similar probabilities for the concepts, assuming the inferred goals as new evidences. This task is developed by means of a *backward inference* (BI) procedure. The decision of assuming whether a concept is needed or not is taken by comparing the probabilities against different thresholds. The result of this process is used to carry out the most suitable action (either performing the goals the user has addressed, if the system has the information needed to accomplish them, or asking the user for the wrong or the incomplete information otherwise).

3 Clustering of Dialogue Elements

This section presents the clustering approaches that we have developed to group dialogue elements, as well as the dynamic LM interpolation that we carry out.

Our proposal is a bottom-up, greedy algorithm that builds a hierarchy of clusters, each of which will have a LM associated. The hierarchy will be established from a starting point in which each cluster will be composed of a single dialogue element, to an ending cluster which contains all the dialogue elements (and therefore it could be assimilated to the general, background LM).

We have proposed two algorithms based on the estimation of the perplexity of LMs The first algorithm performs a method that exploits *local* information to decide which elements should be grouped (that is, the metric is obtained by using only those models directly related to the cluster that is potentially eligible). The second one estimates a *global* measure obtained as a contribution of all the models that are present at each step of the algorithm, and chooses the model that optimizes that measure.

3.1 Maximum Mutual Information Criterion

Let us suppose a set of labeled sentences with which we will train two different language models, A and B, each of which is related to a certain dialogue-specific content (for instance, a dialogue concept or a dialogue goal). We could assume that both LMs have a common subset of training sentences (i.e. they share some knowledge, either lexical, semantic, or intention). Let us further assume that we have obtained the perplexities of both models against an additional database.

The perplexity is related to the average number of words between which a model has to decide the most suitable one. We can estimate the perplexity of a model as $pp_A = 2^{H(A)}$, being $H(A)$ the entropy of that model. In other words, the entropy of the LM A can be obtained as $H(A) = log_2 pp_A$.

On the other hand, the *mutual information* shared between two random variables can be expressed as $I(A;B) = H(A) + H(B) - H(A,B)$. Instead of considering the Mutual Information between two LMs, we use the Normalized Mutual Information (NMI), that can be expressed as $NMI(A;B) = \frac{H(A)+H(B)}{H(A,B)}$.

According to this criterion, we will cluster the elements that maximize the NMI of their related LMs.

We can express the NMI between two models in terms of their perplexity:

$$NMI\,(A,B) = \frac{\log_2 pp_A\ pp_B}{\log_2 pp_{AB}} \tag{1}$$

where $pp_{A,B}$ stands for the perplexity of the joint LM, that is, the LM estimated when using the sentences that trained the models A and B (without repeating the common sentences).

This criterion tends to group elements that share common information (i.e. dialogue elements, or sentences that make reference to those elements). It also allows us to reach a tradeoff between low values of perplexity (that tends to lead to better LMs) and the complexity of the models (in terms of information used to estimate them). We use this criterion since we have several elements for which the number of training sentences is so reduced that their LMs give reduced perplexities, but only due to the lack of training data.

3.2 Minimum Perplexity Criterion

We could consider the NMI criterion as a local one, since the decision of which is the optimum group at each step of the algorithm is taken by considering only the mutual information between those elements that are to be merged, and the resulting cluster. We have also implemented a clustering strategy based on a global criterion, that is, in which the decision on which elements to cluster depends on a metric obtained from all the clusters considered at each step of the algorithm. This criterion is based on a linear interpolation between the LMs related to the clusters that are considered at each step of the algorithm. Then the system estimates the perplexity of the resulting LM. The cluster selected is the one that minimizes the perplexity of the global model.

We assign the same interpolation weight to each LM. That is, if at a certain step of the algorithm there are N_S clusters, the LM related to each model will have an interpolation weight of $1/N_S$.

Therefore, if we represent the probability of obtaining a word w given its history h with the LM related to cluster S_k as $p_{S_k}(w \mid h)$, the corresponding probability in the global, artificial model, p_G, at a certain iteration of the algorithm, can be obtained as

$$p_G\,(w \mid h) = \frac{1}{N_S}\left[p_{S_{ij}}\,(w \mid h) + \sum_{\substack{k=1,\\ k\neq i,j}}^{N_S} p_{S_k}\,(w \mid h)\right] \tag{2}$$

Once the system obtains the perplexity of p_G, the process is repeated for each available combination ij of elements to be grouped (i.e. for each potential cluster). As a result the algorithm obtains a set of global LMs related to all the potential clusters. The algorithm selects as the new cluster to be included in the hierarchy the one that obtains the lowest perplexity among all of them. The rest

of the potential clusters are disregarded in the current step of the algorithm. Nevertheless, they could be considered as potential clusters in further iterations.

The global perplexity minimization criterion is similar to the NMI-based one in the sense that both criteria allows us to obtain groups of elements that share common information. With the NMI metric the system groups those elements that share a high amount of common sentences (i.e. strongly related from the point of view of vocabulary and semantics). In the global perplexity one, the result is similar, but from the model robustness' perspective. That is, the elements that are clustered together are those ones that lead to a better estimated LM. The main difference between both criteria is related to the computing time. The global perplexity minimization one has a higher computational complexity since it has to estimate a higher number of models at each iteration (not only the LM related to the cluster that is included to the hierarchy, but also the specific models and the global one for each potential cluster).

3.3 Estimating a Correction Function

After carrying out some initial clustering experiments, we found that both the NMI and the global perplexity criteria have a main drawback. The cluster hierarchies that are obtained are unbalanced, in the sense that after the first grouping, a cluster with a high number of sentences is obtained. The rest of elements tend to join that cluster instead of building more specific groups. In order to reach a tradeoff between the perplexity of each LM and their complexity (in terms of the number of sentences that will train the corresponding LM, and the number of elements into each cluster), we propose to obtain a complexity correction function that will take a positive value.

The motivation of defining a correction function is to enable the clustering of those elements which have a strong lexical or semantic relationship, even though the related LMs are trained with a reduced number of sentences. This fact will avoid the generation of a too general model with which the rest of elements are progressively joined. In other words, the system can keep an important degree of specificity in the early steps of the clustering algorithm.

Taking into account that we want to optimize the criterion metric, the correction function is applied in two different ways, depending on the chosen criterion for the clistering. In the case of the the NMI measure (which is a maximization function), we will apply the function as a division factor prior to decide which elements to cluster. In a similar fashion, the global perplexity metric (minimization function) will be multiplied by the correction factor.

We will make the correction function dependent on the main features of each cluster, namely the number of dialogue elements that form each cluster, and the number of sentences with which the LM associated to the cluster will be estimated.

The number of elements joined in a given cluster S_i, which we denote as N_{S_i}, will model the complexity of the clusters. It is used to allow those clusters with few elements to be joined among them, avoiding thus the tendency to join a

cluster with more elements, which in turn leads to less specific LMs, especially in the initial steps of the clustering algorithm.

The correction criterion will also take into account the number of sentences n_A and n_B that have been used to train the LMs related to the clusters to be joined, as well as the number of sentences of the resulting cluster, n_{AB}. We use the number of sentences as a value that can measure both the complexity of the model and also its robustness (the larger the number of sentences to train a LM, the better it will be estimated).

The correction function will consider the number of sentences in the sense of favoring the union of those elements that share a large number of common sentences and a reduced number of different sentences.

The situation in which the correction function reaches its maximum value arises when there are not any sentence in common between both models. In other words, a lexical or semantic relationship between both clusters A and B is too weak or inexistent, and therefore both clusters should not be joined in the current step of the algorithm. This situation arises when $n_{AB} = n_A + n_B$.

A final restriction that we apply to the correction function is that the contribution of the number of sentences is measured on a logarithmic scale. We decide that since the number of sentences with which the LMs are trained is about two orders of magnitude over the entropy of the models (which is also a logarithmic magnitude).

Taking these conditions into account, the expression of the correction function CF for joining two clusters A and B into a single cluster AB is

$$CF = N_{S_i} \ \ln \left[\frac{\sqrt{(n_{AB} - n_B) \ (n_{AB} - n_A)}}{n_A + n_B - n_{AB}} + \mathcal{K}_0 \right] \tag{3}$$

where $\mathcal{K}_0$ is a constant that assures that the logarithm takes a positive value.

We finally apply a pruning process to the cluster hierarchies obtained. The idea is to keep these LMs that are trained with a sufficient number of sentences, and also assuring that each LM is related to a specific content (i.e. we try to reach a tradeoff between *robustness* and *specificity* of the LMs. The number of LMs to be considered are 10 (when using goal-based information), 23 (when considering concepts), and 25 (when grouping both dialogue elements).

4 Dynamic Language Model Generation

We have included a new module as a feedback loop between the ASR, the NLU, and the DM modules. This new element, the Dynamic LM Generator, will consider the information provided by the user in the current and the previous utterances to dynamically modify the LMs that the ASR makes use of.

We first estimate the LM related to each cluster. Instead of keeping a LM for each dialogue element, as we proposed in [7], we consider that keeping 73 LMs is a suboptimal approach, since several of these models are poorly estimated, due to the limited amount of sentences that make reference to those elements. Therefore, we proposed to group the dialogue elements in a hierarchical cluster

structure, according to the semantic relationships among them ([8]). Our aim is to reach a tradeoff between the *specificity* of having a large number of LMs related to single pieces of information, and the *robustness* of having few LMs, but trained with more data.

At each dialogue turn, once a sentence has been recognized, and the DM has developed both forward and backward inferences, the posterior probabilities of concepts and goals are used to decide which LMs will be interpolated. We base this decision on the comparison of the posterior probabilities of the dialogue elements against different *relevance thresholds*, Φ_C for concepts and Φ_G for goals. We find the optimal values for Φ_C and Φ_G at a validation stage. We perform the LM adaptation by means of a linear interpolation between a background LM, p_B, and a *context-dependent* LM, p_D. The probability of a word w given its preceding words (its history) h in the interpolated model will then be

$$p_I\,(w \mid h) = (1 - \lambda_D)\; p_B\,(w \mid h) + \lambda_D\; p_D\,(w \mid h) \tag{4}$$

being λ_D the interpolation weight between the background LM and the context-dependent LM, p_D. This model is also built by interpolating the LMs related to clusters to which the dialogue elements belong to. The interpolation weights are obtained as functions of the posterior probabilities of each dialogue element, and also as a function of the number of elements on each cluster.

By using the summation of posterior probabilities we can achieve a tradeoff between the contribution of the number of elements belonging to each cluster, and their posterior probabilities, giving more relevance to those clusters to which more dialogue elements belong to, or to those ones with the dialogue elements with greater posterior probabilities.

5 Experimental Setup

This section presents the database that we have used to assess the adapted system, and the evaluation results.

Our proprietary database comprises 1300 different sentences, uttered by 13 speakers (7 male, 6 female), giving a vocabulary of 391 words. Each sentence has been manually labeled with its appropriate concepts and goals. By means of a k-fold approach we have split the database into ten folds (each one with 130 sentences picked up randomly from the database), with which we build three sets: a *training* one, composed of eight folds (1040 sentences), and a *validation* and a *test* sets, each one with one fold (130 sentences). Using round-robin we develop ten experiments. On each one we use the training set to build the LMs, whereas the validation set is used to adjust the parameters of the system.

We have evaluated the word error rate (WER) of the speech recognizer, the concept error rate (CER) of the understanding module, and the goal error rate (GER, that is, the percentage of errors in the inference of goals).

Throughout the evaluation we have assessed the performance of the system when using the concepts and goals extracted from an utterance to dynamically adapt the LM, and use it to recognize again the same sentence. This way we can estimate an upper bound of the performance of our system.

5.1 Using the NMI Criterion

In our first experiment we consider the clustering strategy based on maximum normalized mutual information (NMI). Table 1 shows the results of the evaluation in terms of WER, CER and GER, when considering only concept-dependent information, only goal-dependent information, or when merging both elements for the clustering. We also include the performance of the baseline system (i.e. with the background, static LM).

Table 1. Performance of the NMI-based language modeling

Clustering approach	WER (%)	CER (%)	GER (%)
Baseline	5.33	13.37	26.20
Concepts	4.82	12.73	25.67
Goals	4.84	12.68	**25.53**
Both	4.70	12.66	25.71

The interpolation weight λ_D takes values of about 0.15. That is, it is enough to slightly modify the LM (keeping a 85% of the background LM) to achieve improvements in the three metrics considered. The improvements reach a maximum relative value (in terms of error reduction) of 11.80% WER and 5.34% CER (both when considering the clustering of both dialogue elements together). On the other hand, the maximum relative error reduction in Goal Error Rate (2.56%) is reached when considering only dialogue goals. The main reason for this behaviour is that using only goal-based information (that is, the more integrated source of information that the system considers) implies a reduction of the insertions of goals into the hypothesis, which are the most important source of errors. In any case, the size of our database makes that the improvements in GER are not statistically significant.

5.2 Using the Minimum Perplexity Criterion

We next evaluate the performance of the adapted system when using the Minimum Global Perplexity criterion. Table 2 shows the results of the evaluation of this strategy.

Table 2. Performance of the Minimum Perplexity-based language modeling

Clustering approach	WER (%)	CER (%)	GER (%)
Baseline	5.33	13.37	26.20
Concepts	**4.52**	**12.54**	25.60
Goals	4.60	12.59	25.64
Both	4.58	12.66	25.64

The interpolation weight λ_D between the background LM and the context-dependent one (i.e. the generated using the LMs associated to the clusters considered) takes a value of about 0.21. Using this clustering strategy, the relevance

of the context-dependent component is higher than with the NMI-based clustering approach. This fact implies that the LMs obtained with the Maximum Global Perplexity criterion tend to be better estimated. This leads to a slightly better performance of the system (with maximum relative error reduction of 15.17% for Word Error Rate, and 6.28% for Concept Error Rate, both when considering concept-based clustering). The improvement of the WER is statistically significant with confidence intervals of 90%. As regards the dialogue performance, the GER also tends to decrease (up to a maximum of 2.29% of relative reduction). However, this value is not statistically significant.

Merging both dialogue elements cannot outperform the strategies of using the elements separately. This could happen due to the fact that the goals are inferred using the concepts. Therefore, using both sources of information may cause the estimation of LMs with redundant information. This redundancy could cause the reduction of the performance observed. In any case, the differences between the performance of the clustering strategies are not significant.

6 Conclusions

We have presented two strategies to cluster dialogue-based information that is used to generate content-specific language models. The first approach is based on a local criterion that considers the Normalized Mutual Information (NMI) to decide which elements to cluster at each step of the algorithm. The second one is based on a global criterion that tries to minimize the perplexity of a model obtained as a linear interpolation of the LMs related to the clusters considered. The LMs obtained are interpolated at each turn with a background LM to dynamically adapt the model to be used by the recognizer.

Instead of training the most accurate interpolation weights, one of our main claims is that the system can estimate accurate interpolation weights dynamically using the posterior probabilities obtained by the DM. This way, the more confident the system is when inferring a given concept or goal, the more relevant the LM associated to that dialogue element will be in the dynamic LM estimated at that turn.

The evaluation results show that these clustering strategies lead to an estimation of LMs which can improve the recognition performance. More importantly, the improvement of these LMs (used by the speech recognizer) tends to improve the performance of other modules of the system (the speech understanding and the DM). We have also seen that the clustering based on the minimization of the perplexity tends to obtain better LMs (from both the specificity and the robustness points of view) than the NMI-based one.

We are aware that the databases that we have used are limited. We are now acquiring and preparing new data to train the LMs related to the different dialogue elements. This way we have to label this data at the three levels of information (lexical, semantic, and user intention).

We are now working on another interpolation strategy for the Minimum Perplexity approach. Instead of using the same weight for all the LMs, we will make them dependent on the complexity of each cluster.

We are also defining a strategy to adjust dynamically the weight λ_D between the background and the context-dependent LMs, instead of obtaining it at a validation stage.

We are also applying our adaptation paradigm to other information sources, such as the knowledge that the system has about the users, taking into account that each speaker may express their ideas in different ways. The system could take advantage of this information once it identifies the speaker, to adapt the LMs to the characteristics of each user.

Acknowledgements. This work has been partially supported by the Spanish Ministry of Science and Innovation, under contracts TIN2008-06856-C05-05 (SD-TEAM UPM) and DPI2010-21247-C02-02 (INAPRA), and by the Spanish Ministry of Education, under contract AP2007-00463 (FPU Grant).

References

1. Bahl, R.L., Brown, P.F., de Souza, P.V., Mercer, R.L.: Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In: Proc. ICASSP, pp. 49–52 (1986)
2. Bellegarda, J.R., Butzberger, J.W., Chow, Y.L., Coccaro, N.B., Naik, D.: A Novel Word Clustering Algorithm Based on Latent Semantic Analysis. In: Proc. ICASSP, vol. I, pp. 172–175 (1996)
3. Bellegarda, J.R.: Statistical language model adaptation: review and perspectives. Speech Comm. 42, 93–108 (2004)
4. Fernández, F., Ferreiros, J., Sama, V., Montero, J.M., San-Segundo, R., Macías-Guarasa, J.: Speech Interface for Controlling a Hi-Fi Audio System Based on a Bayesian Belief Networks Approach for Dialog Modeling. In: Proc. INTERSPEECH, pp. 3421–3424 (2005)
5. GuoDong, Z., KimTeng, L.: Interpolation of n-gram and mutual-information based trigger pair language models for Mandarin speech recognition. Comp. Speech & Lang. 13, 125–141 (1999)
6. Kneser, R., Steinbiss, V.: On the dynamic adaptation of stochastic language models. In: Proc. ICASSP, vol. II, pp. 586–589 (1993)
7. Lucas-Cuesta, J.M., Fernández, F., Ferreiros, J.: Using Dialogue-Based Dynamic Language Models for Improving Speech Recognition. In: INTERSPEECH, pp. 2471–2474 (2009)
8. Lucas-Cuesta, J.M., Fernández, F., López, V., Ferreiros, J., San-Segundo, R.: Clustering of syntactic and discursive information for the dynamic adaptation of Language Models. In: SEPLN, vol. 45, pp. 175–182 (2010)
9. Solsona, R.A., Fosler-Lussier, E., Kuo, H.K.J., Potamianos, A., Zitouni, I.: Adaptive Language Models for Spoken Dialogue Systems. In: Proc. ICASSP, vol. I, pp. 37–40 (2002)

A Multilingual SLU System Based on Semantic Decoding of Graphs of Words

Marcos Calvo, Lluís-F. Hurtado, Fernando García, and Emilio Sanchis

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{mcalvo,lhurtado,fgarcia,esanchis}@dsic.upv.es

Abstract. In this paper, we present a statistical approach to Language Understanding that allows to avoid the effort of obtaining new semantic models when changing the language. This way, it is not necessary to acquire and label new training corpora in the new language. Our approach consists of learning all the semantic models in a target language and to do the semantic decoding of the sentences pronounced in the source language after a translation process. In order to deal with the errors and the lack of coverage of the translations, a mechanism to generalize the result of several translators is proposed. The graph of words generated in this phase is the input to the semantic decoding algorithm specifically designed to combine statistical models and graphs of words. Some experiments that show the good behavior of the proposed approach are also presented.

Keywords: Multilingual Language Understanding, Graph of Words.

1 Introduction

In the last few years, different approaches have been developed for the problem of Spoken Language Understanding (SLU). There are many types of applications for SLU, and one of the most interesting is its use in limited-domain spoken dialog systems. Some characteristics of this kind of systems are that they have to deal with spontaneous speech, the size of the vocabulary is small or medium, and the semantic labels involved in the understanding process are strongly related to some specific words or segments of words present in the user turns. In the recent literature, a variety of approaches for automatic SLU have been proposed, like those explained in [1–3].

As in other speech areas, statistical modelization is one of the successfully approaches that have been used in SLU [4–7]. One of the advantages of these approaches is that the models can be automatically learned from labeled training samples and they can represent the variability of sequences of words and concepts. Due to the limited number of training samples, and the limitations of the learning methods, not all the variability of the speech messages can be correctly represented in the models, and some errors generated in previous phases can not be recovered in the following phases. This is the reason why the use of n-best or

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 158–167, 2012.

weighted graphs of linguistic units are interesting approaches to communicate information between the different modules [8, 9].

The use of some kind of graph of words as the input of the decoding module makes this task more difficult, as the search space becomes larger, and it is necessary to combine the different weights representing the accuracy of the words in the graph and the corresponding probabilities of the model of the decoding process (in our case the semantic models). On the other hand, the advantage of using graphs is that there is more information that could help to find the correct semantic interpretation, rather than just taking the best sentence given by the Automatic Speech Recognizer (ASR), or other module that provides the input sentence.

The methodology proposed in this paper is based on Stochastic Finite State Transducers (SFST). This is a generative approach that composes several transducers containing acoustic, lexical and semantic knowledge. Our method performs this composition obtaining as a result a graph of concepts, where semantic information is associated to segments of words. To carry out this step, we use a statistical modelization of the lexicalisation of concepts; that is, the sequences of words associated to the concepts, and also a statistical model of the sequences of concepts. All these probabilities are automatically learned from a training corpus segmented and labeled in terms of concepts.

One of the problems of the statistical modelization of semantics is that the training process needs a segmented and labeled corpus. In most cases it is necessary a very time-consuming work to label the training corpus. This is the reason why many works oriented to avoid this work have been developed, such as semi-supervised learning techniques, or active learning methods [10, 11]. These techniques are also used to facilitate the adaptation of the system to different tasks or new languages. In particular, when the problem is to translate a previously obtained SLU system into another language, some approaches can be used: to translate the corpus and to do a new labeling; to automatically obtain the translated system and labeling; or to process the sentences in the new language (after translating them) with the original models. The latter approach is the one that we have developed in this paper. That is, we obtained the semantic models for Spanish by using a labeled training corpus, and we used this system to interact with users of other language (French in this work) by translating their sentences into the target language and decoding these translated sentences. However, it must be considered that the quality of the general purpose translators is quite insufficient. This is the reason why it is necessary to supply the maximum information of the original sentences to the semantic decoding process, in order to better tackle the errors generated in the translation process.

In the proposed system, the sentence uttered in the source language is translated into a graph of words in the target language, by means of an adequate combination of the translations generated by several web translators. This way, we obtain a generalization of the translations that allows the semantic decoder to recover some of the errors generated in the translation phase.

This paper is organized as follows. In Section 2, the general framework of the system is presented. In Section 3, the process of generating the graph of words from the different sentences obtained by the translators is described. In Section 4, the algorithm of semantic decoding of the graph of words is presented. Section 5 shows some experimental results over the DIHANA task, and finally, in Section 6 some conclusions and future work are presented.

2 The SFST Approach for Multilingual SLU

The SLU problem can be expressed as stated in Equation 1, where C represents a sequence of concepts or semantic labels and A is the utterance that constitutes the input to the system.

$$\hat{C} = \underset{C}{\operatorname{argmax}} p(C|A) \tag{1}$$

The task we are addressing is multilingual SLU, which in our case means that the speaker utters a sentence in one language s, but our models are trained in another language t. Thus, a translation process between the source and target languages is needed. Taking into account the underlying sequence of words W_s uttered by the speaker in the source language and its translation into the target language W_t, this equation can be rewritten as follows.

$$\hat{C} = \underset{C}{\operatorname{argmax}} \max_{W_s, W_t} p(C, W_s, W_t|A) \tag{2}$$

Applying the Bayes' Rule and making some reasonable assumptions about the independence of these variables, the probability of this equation can be decomposed into several factors as shown in Equation 3.

$$\hat{C} = \underset{C}{\operatorname{argmax}} \max_{W_s, W_t} p(A|W_s) \cdot p(W_s|W_t) \cdot p(W_t|C) \cdot p(C) \tag{3}$$

This equation can be seen as the composition of 4 SFST, which are:

- A SFST λ_G generated by the ASR module where the acoustic probabilities $p(A|W_s)$ are represented.
- A SFST $\lambda_{W_s 2 W_t}$ that expresses the translation process between the source and target languages.
- A SFST $\lambda_{W_t 2 C}$ that represents the probability that a sequence of words in the language t corresponds to a concept C. Thus, it provides the probability distribution $p(W_t|C)$.
- A SFST λ_{SLM} which corresponds to a language model of the sequences of concepts. Thus, it modelizes the probability of a sequence of concepts $p(C)$.

It is possible to compose these four transducers as shown in Equation 4.

$$\lambda_{MSLU} = \lambda_G \circ \lambda_{W_s 2 W_t} \circ \lambda_{W_t 2 C} \circ \lambda_{SLM} \tag{4}$$

In consequence, finding the best path in the resulting transducer λ_{MSLU} provides as a result the best sequence of concepts $\hat{C}$, a translation $\tilde{W}_t$ of the transcription of the input utterance as well as a segmentation of $\tilde{W}_t$ according to $\hat{C}$.

In this work, our goal is to build and evaluate a multilingual understanding system that receives a sentence in one language and, passing this sentence through a translation process, is able to use understanding models trained in another language. If the input to the system were utterances, the recognition process would add some error to the output of the understanding module. Thus, in order to evaluate the performance of the understanding system without any other external factors, the input to our system will be correct written sentences, which is equivalent to assume that we have a "perfect" ASR. In terms of probabilities, this implies that $p(A|W_s) = 1$. For this reason, we will not use the λ_G transducer from Equation 4.

Moreover, $p(W_s|W_t)$ can be rewritten as $\frac{p(W_t|W_s)\cdot p(W_s)}{p(W_t)}$. Taking the written sentence as the input to the system, means that the whole sentence that is going to be translated is known[1]. From this known sentence in the source language, we will obtain a set of possible translations and represent them as a graph of words. If we consider that the probability $p(W_t)$ of any sentence of the set of possible translations is the same, then it is not necessary to take into account this probability in the maximization process. Considering these two simplifications, we can rewrite Equation 3 as:

$$\hat{C} = \underset{C}{\operatorname{argmax}} \max_{W_t} p(W_t|W_s) \cdot p(W_t|C) \cdot p(C) \tag{5}$$

Thus, the $\lambda_{W_s 2 W_t}$ transducer will represent the probability $p(W_t|W_s)$ that a sentence W_t in the target language is a translation of W_s.

3 Graph of Words Generation

In this section, the process of obtaining the word-graph in the target language from a sentence in the source language is be explained. This process is divided into three steps:

1. the source sentence (in French in this work) is translated to the target language using several free-available web translators. As a result, a set of sentences in the target language (Spanish in this work) that represent different possible translations of the source sentence is obtained.
2. this set of sentences are aligned using a multiple sequence alignment algorithm.
3. the aligned sentences are used to obtain the word-graph that will be the input to the graph-based understanding module.

A Multiple Sequence Alignment (MSA) is a sentence alignment process that allows the alignment of three or more sentences that minimize the number of substitutions, insertions and deletions among all the sentences. Although the original use of MSA is the alignment of biological sequences, MSA algorithms can align sequences of symbols of any kind. Within the frame work of Natural

[1] It would be the same if we took the 1-best from an ASR.

Language Processing (NLP), MSA has been mainly used in automatic translation tasks [12, 13]. All these approaches –including the one presented in this paper– coincide in the creation of a graph of words from the result of the MSA. However, they differ in how the graph is generated and what it is used for.

In this work, a modification of the well-known ClustalW [14] Multiple Sequence Alignment software has been used. These modifications consist basically of: i) it allows the alignment of sentences with any symbol, originally ClustalW only allows symbols representing protein, DNA, and RNA; ii) all weight matrices have been replaced by 0s and 1s (where 1 is the score for symbol matches and 0 is the score for symbol mismatches). That is, the same probability is assigned to all symbol substitutions.

3.1 From Alignment Matrix to Graph of Words

The result of the Multiple Sequence Alignment process is a MSA alignment matrix. Each row in the matrix represents a different aligned sentence and the columns are synchronization points. In other words, a column of the matrix indicates which symbols of the sentences are aligned at each point. Unless all the sentences are the same, there will be several non-alignment points. These non-alignment points are represented in the alignment matrix by the special symbol '-'.

From the MSA alignment matrix, a directed acyclic weighted graph of words is created. In order to build this graph of words, the following algorithm is used:

1. as many nodes as the number of columns in the alignment matrix plus one additional node to be used as the initial node are created.
2. for each matrix cell containing a symbol other than '-' –that is, a cell that represents a real word of an aligned sentence– an arc in the graph will be created. The destination node of the arc will be the one representing the column to which the cell belongs and the origin node will be the one representing the column of the previous word in the same sentence (or the initial node if the cell contains the first word of the sentence). The arc is labeled with the word in the cell and its weight is set to 1. If the arc already exists (because it has been previously added), its weight is incremented by 1.
3. the weights of the arcs are normalized to represent probabilities.

Figure 1 shows a real example –extracted from the test set– of the full process of obtaining the graph of words in the target language (Spanish) from a sentence in the source language (French). Firstly, the original sentence *pouvez vous répéter à quelle heure sort le premier* is translated using 6 different free-available web translators. Secondly, the 6 translations are aligned using a Multiple Sequence Alignment algorithm. Finally, the directed acyclic weighted graph of words is created from the MSA alignment matrix.

The obtained graph of words represents a language which is a generalization of the individual translations of the original sentence. A full path –from the initial node to the final node– over the graph may be seen as an alternative translation

(1) **source sentence** | *pouvez vous répéter à quelle heure sort le premier*

(1) **multiple translations**

poded a qué hora sale el premier
puede repetir a qué hora saca el primero
poded a qué hora sale el premier
puede repetirse a qué hora suerte el primero
pueda repetirse a qué hora suerte el primero
puede me repetir a qué hora saca el primero

(2)

(3) **alignment matrix**

-	-	poded	a	qué	hora	sale	el	premier
puede	-	repetir	a	qué	hora	saca	el	primero
-	-	poded	a	qué	hora	sale	el	premier
puede	-	repetirse	a	qué	hora	suerte	el	primero
pueda	-	repetirse	a	qué	hora	suerte	el	primero
puede	me	repetir	a	qué	hora	saca	el	primero

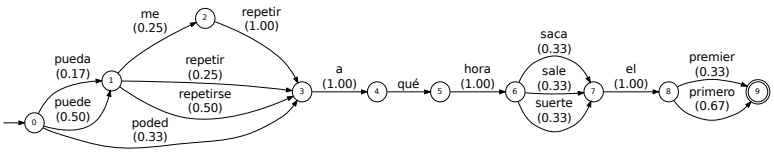


Fig. 1. Steps in the process of obtaining the graph of words from the original sentence *pouvez vous répéter à quelle heure sort le premier*

of the original sentence. In addition, the product of the weights of all the arcs in a full path may be considered as the probability of the string represented by the path W_t (in the target language) to be the translation of the original string W_s (in the source language); that is, $p(W_t|W_s)$.

4 Understanding the Translated Graphs of Words

As every arc in the graph of words is labeled with a word, any path between any pair of nodes represents a segment of words that can be related to one or more concepts. Consequently, it is possible to build a new graph with the same set of nodes but where each arc is labeled with a segment of words and a concept associated to it. Every arc of this new graph can be weighted with a combination of the original graph probability and the probability that the segment belongs to the concept.

To build this graph of concepts, for each pair of nodes i, j and each concept c, an arc that represents the most probable path associated to concept c between these nodes is created. We define the most probable path as the one that maximizes the expression $p(W_t^{i,j}|W_s)\cdot p(W_t^{i,j}|c)$ given a concept c, where $W_t^{i,j}$ denotes

a segment of words induced by a path from node i to node j. The resulting arc will be labeled with the concept c, and the sequence $W_t^{i,j}$ that maximizes the former probability. The arc will be weighted with the value of this expression for $W_t^{i,j}$ and c.

The last formula introduces the probability of a sequence of words given a concept. To estimate it, a model of the lexical structures associated to the concepts is needed. One way to estimate this is to train a language model for each concept, using the segments of words associated to each of them. Thus, the probability $p(W_t^{i,j}|W_s)$ is represented in the graph of words obtained from the translation and generalization process, and $p(W_t^{i,j}|c)$ is provided by the language model specific to each concept.

Finally, finding the best path in the graph of concepts between the initial and the final nodes, provides the best sequence of concepts and the sentence associated to it, as well as a segmentation of this sentence. To find this best path, a language model of the sequences of concepts may be used, in order to properly model their concatenation.

This way of obtaining the best sequence of concepts fulfills what was stated in Equation 4, as $\lambda_{W_d 2C}$ is composed by the set of the language models that provide the probability that a sequence of words belongs to a concept and λ_{SLM} is the language model of the sequence of concepts that helps to find the best path in the graph of concepts.

5 Experiments and Results

For the experimentation, we used the DIHANA corpus [15]. This is a corpus in Spanish composed by 900 dialogs acquired by 225 speakers using the Wizard of Oz technique, with a total of 6,229 user turns. The DIHANA task consists of acceding by phone to a spoken dialog system to ask for information about railway timetables and fares using spontaneous speech in Spanish. The experiments reported here were performed using the 5,369 user turns of the DIHANA corpus, splitting them into a set of 480 turns for test and the remaining 4,889 turns for training. Some interesting statistics about the DIHANA corpus are shown in Table 1.

Table 1. Characteristics of the DIHANA corpus

Number of words	47,222
Vocabulary size	811
Average number of words per user turn	7.6
Number of concepts	30
Average number of words per segment	2.5
Average number of segments per turn	3.0
Average number of samples per concept	599.6

In order to perform multilingual experiments using the DIHANA task, the 480 test sentences were correctly written in French. Then, we used 6

general-purpose free-available web translators to translate them into Spanish. The 6 translations of each sentence were combined using the algorithm explained in Section 3, obtaining as a result the graph of words which is the input for the decoding algorithm.

In order to learn all the semantic models, the Spanish training sentences were used. The transcriptions of the user sentences of the DIHANA training corpus were segmented and labeled in terms of concepts. This segmentation were used to learn a language model for each concept. In addition, a semantic model was also leaned using the sequences of concepts. All the language models involved in this experimentation were bigram models trained using Witten-Bell smoothing and linear interpolation.

For the evaluation, we have used three measures: the Translation Word Error Rate (TWER), the BLEU measure, and the Concept Error Rate (CER). The TWER represents the WER of the best path in the graph of words; that is, the path that has been generated by the semantic decoding process. The BLEU is a standard measure used to evaluate automatic translation systems. The CER is the rate of incorrectly understood concepts, considering that the reference sequence of concepts is the same in both languages (which means that, in some cases, some correct sequences in French can be counted as errors). The TWER and BLEU measures represent not only the quality of the composition of transducers but also the behavior of the search algorithm guided by the semantics.

Table 2 shows the results obtained in the experiments, both for the combination of translators and for each one of them, individually. It also shows the results considering the correct sentences in Spanish as input. This result gives an idea of the best CER that could be achieved with our semantic modelization if no error were introduced in the translation and generalization processes.

Table 2. TWER, BLEU, and CER obtained from the combination of translators and each of them individually, as well as for the reference sentences in Spanish

Input graphs of words	TWER	BLEU	CER
Reference sentences	–	–	9.09
Translator 1	30.74	50.37	15.77
Translator 2	27.49	52.00	16.67
Translator 3	30.50	50.71	15.22
Translator 4	24.04	61.35	13.09
Translator 5	23.85	59.79	14.60
Translator 6	27.38	50.82	19.35
Combination of all the translators	**18.68**	**67.40**	**11.78**

These results show that the combination of the translators obtains better results that considering them individually. That is, the increasing of the coverage given by the use of several translators, and the adequate combination in the graph of words outperforms the behavior of each isolated translator. These better performances are observed in terms of TWER, BLEU, and CER. In addition,

the CER obtained for the combination of the translators is less than 2.7 points higher than the one achieved using the reference Spanish sentences. This means that, although the translation process introduces some syntactic errors (which can be seen in the TWER score), most of the semantic meaning is kept.

6 Conclusions and Future Work

We have presented an approach to multilingual language understanding. One of the advantages of this approach is that it is not necessary to estimate different models depending on the language. The modelization of the semantics of the task is done by statistical models. The way to represent the variability of the translation process is done by the construction of graphs of words. We have developed a search algorithm to generate graphs of concepts form the graphs of words and the semantic models. Experiments show that the proposed approach achieves good results. It would be interesting, as future work, to adapt the system to other languages –like English– that have syntactic structures different from Latin languages.

Acknowledgments. This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, by the Vic. d'Investigació of the UPV under contract 20110897, and by the Spanish MICINN under FPU Grant AP2010-4193.

References

1. Hahn, S., Dinarelli, M., Raymond, C., Lefèvre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., Riccardi, G.: Comparing stochastic approaches to spoken language understanding in multiple languages. IEEE Transactions on Audio, Speech, and Language Processing 6(99), 1569–1583 (2010)
2. Raymond, C., Riccardi, G.: Generative and discriminative algorithms for spoken language understanding. In: Proceedings of Interspeech 2007, pp. 1605–1608 (2007)
3. Tur, G., Mori, R.D.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, 1st edn. Wiley (2011)
4. Maynard, H.B., Lefèvre, F.: Investigating Stochastic Speech Understanding. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop, ASRU (2001)
5. Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L.: Extracting Semantic Information Through Automatic Learning Techniques. IJPRAI 16(3), 301–307 (2002)
6. He, Y., Young, S.: Spoken language understanding using the hidden vector state model. Speech Communication 48, 262–275 (2006)
7. De Mori, R., Bechet, F., Hakkani-Tur, D., McTear, M., Riccardi, G., Tur, G.: Spoken language understanding: A survey. IEEE Signal Processing Magazine 25(3), 50–58 (2008)
8. Hakkani-Tür, D., Béchet, F., Riccardi, G., Tur, G.: Beyond ASR 1-best: Using word confusion networks in spoken language understanding. Computer Speech & Language 20(4), 495–514 (2006)

9. Tur, G., Wright, J., Gorin, A., Riccardi, G., Hakkani-Tür, D.: Improving spoken language understanding using word confusion networks. In: Proceedings of the ICSLP. Citeseer (2002)
10. Tur, G., Hakkani-Tür, D., Schapire, R.E.: Combining active and semi-supervised learning for spoken language understanding. Speech Communication 45, 171–186 (2005)
11. Ortega, L., Galiano, I., Hurtado, L.F., Sanchis, E., Segarra, E.: A statistical segment-based approach for spoken language understanding. In: Proc. of InterSpeech 2010, Makuhari, Chiba, Japan, pp. 1836–1839 (2010)
12. Sim, K.C., Byrne, W.J., Gales, M.J.F., Sahbi, H., Woodland, P.C.: Consensus network decoding for statistical machine translation system combination. In: IEEE Int. Conference on Acoustics, Speech, and Signal Processing (2007)
13. Bangalore, S., Bordel, G., Riccardi, G.: Computing Consensus Translation from Multiple Machine Translation Systems. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2001, pp. 351–354 (2001)
14. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: ClustalW and ClustalX version 2.0. Bioinformatics 23(21), 2947–2948 (2007)
15. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López de Letona, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proceedings of LREC 2006, Genoa, Italy, pp. 1636–1639 (May 2006)

Merging Intention and Emotion to Develop Adaptive Dialogue Systems

Zoraida Callejas[1], David Griol[2], and Ramón López-Cózar Delgado[1]

[1] Dep. Languages and Computer Systems,
University of Granada, CITIC-UGR, 18071 - Granada, Spain
{zoraida,rlopezc}@ugr.es
[2] Computer Science Department,
Universidad Carlos III de Madrid, 28911 - Leganés, Spain
dgriol@inf.uc3m.es

Abstract. In this paper we propose a method for merging intentional and emotional information in spoken dialogue systems in order to make dialogue managers more efficient and adaptive. The prediction of the user intention and emotion is carried out for each user turn in the dialogue by means of a module conceived as an intermediate phase between natural language understanding and dialogue management in the architecture of these systems. We have applied and evaluated our method in the UAH system, for which the evaluation results show that merging both sources of information improves system performance as well as its perceived quality.

Keywords: Spoken Dialogue Systems, Emotion Processing.

1 Introduction and Related Work

With the aim of developing systems capable of maintaining a conversation as natural and rich as a human conversation, emotion is gaining increasing attention from the dialogue systems community as it affects the actions that the user chooses to communicate with the system.

In [1] the authors found that emotional information can be useful to improve dialogue strategies and predict system errors, but it was not employed in their system to adapt dialogue management. Boril et al. [2] discuss that cognitive load and emotional states affect the number of query repetitions required for drivers to obtain information from commercial dialogue systems. In [3], the authors implemented an adapted strategy for providing support to users depending on their emotional state while solving a puzzle. Although the help policy was adapted to emotion, the rest of the decisions made by the dialogue manager did not take into account emotional information.

Our proposal merges the traditional view of the dialogue act theory in which communicative acts are defined as intentions or goals, with the recent trends that consider emotional states in order to carry out enhanced dialogue management. To do so, we propose a user state prediction module which can be integrated in

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 168–177, 2012.

the architecture of a spoken dialogue system to adapt the dialogue management, as will be explained in Section 2.

Other authors have developed alternative affective dialogue models. For example, the model proposed in [4] derived the next dialogue state from a combination of a plain dialogue model and a combined model including the dependencies between dialogue and emotional states. This dialogue manager was developed in VoiceXML and ECMAScript. In our proposal, we employ statistical techniques to infer user acts, which makes it easier porting it to different application domains. Moreover, our proposed architecture is modular, which makes it possible to employ different emotion and intention recognizers.

In [5] the dialogue model was based on POMDPs [6] to adapt the dialogue strategy to the user actions and emotional states. To reduce the computational cost when many emotions and dialogue acts are considered, the authors complemented POMDPs with decision networks. We propose an alternative to this statistical modeling that can be used in realistic dialogue systems, and evaluate it in a less emotional application domain in which emotions are produced more subtly.

2 Our Proposal

The proposed method predicts user states in terms of intention and emotion, and can be integrated in the architecture of a spoken dialogue system as a module placed between the natural language understanding and the dialogue management phases, as shown in Figure 1. The module is comprised of an emotion recognizer, an intention recognizer, and a user state composer. The emotion recognizer detects the user emotional state by extracting an emotion category from the voice signal and the dialogue history. The intention recognizer takes the semantic representation of the user input and predicts the next user action. Then, in the user state composition phase, a data structure is built from the emotion and intention recognized, which is passed on to the dialogue manager.

2.1 The Emotion Recognizer

As the architecture shown in Figure 1 has been designed to be highly modular, different emotion recognizers can be employed within it. We have employed a recognition method based on our previous work [7].

We are interested in recognizing negative emotions that might discourage users from employing the system again or lead them to abort an ongoing dialogue. Concretely, we have considered three negative emotions: anger, boredom, and doubtfulness, where the latter refers to a situation in which the user is uncertain about what to do next. The recognizer employs acoustic information to distinguish anger from doubtfulness or boredom, and dialogue information to discriminate between doubtfulness and boredom, which are more difficult to discriminate only by using phonetic cues.

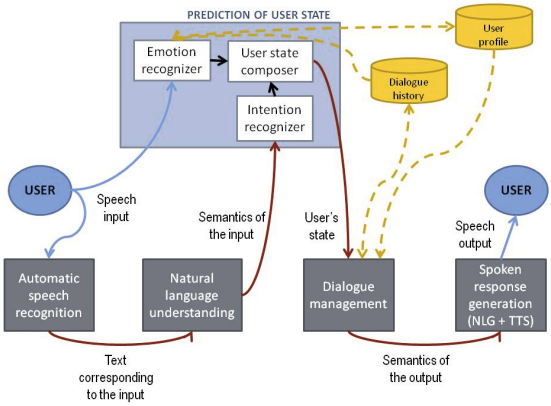


Fig. 1. Integration of the user state prediction into the architecture of a spoken dialogue system

In most information providing spoken dialogue systems, the application domain is not highly affective, thus a baseline algorithm which always chooses 'neutral' provides a very high accuracy (in our case 85%). This rate is difficult to improve by classifying the rest of emotions, which are very subtlety produced. Instead of considering neutral as another emotional class, we calculate the most likely non-neutral category. The dialogue manager employs the intention information together with this category to decide whether to treat the user input as emotional or neutral, as will be explained in Section 3.

2.2 The Intention Recognizer

The methodology that we have developed for modelling the user intention extends our previous work in statistical models for dialogue management [8]. We consider user intention as the predicted next user action to fulfill their objective in the dialogue. It is computed taking into account the information provided by the user during the dialogue and the last system turn.

The formal description of the proposed model is as follows. Let A_i be the output of the dialogue system (the system answer) at time i, expressed in terms of dialogue acts. Let U_i be the semantic representation of the user intention. We represent a dialogue as a sequence of pairs (A_i, U_i), where A_1 is the system greeting (the first dialogue turn), and U_n is the last user turn.

The objective of the user intention recognizer at time i is to select an appropriate user answer U_i. This selection is a local process for each time i, which takes into account the sequence of pairs that precede time i and the system answer at time i. The selection of the most likely user intention $\widehat{U}_i$ at each time i, is made using the following maximization rule: $\widehat{U}_i = argmax_{U_i \in U} P(U_i | UR_{i-1}, A_i)$, where the set U contains all the possible user answers, and UR_i is what we call the *user register* at time i.

The user register is a data structure that, on the one hand, contains information about concepts and attribute values provided by the user throughout the previous dialogue history. On the other hand, it contains information regarding the user profile: id, gender, experience, skill level, most frequent objective of the user, a reference to the location of all the information regarding the previous interactions and the corresponding objective and subjective parameters for that user, and the parameters of the user neutral voice.

To recognize the user intention, we assume that two different sequences of states are equivalent if they lead to the same UR and that the exact values for the attributes provided by the user are not significant to determine the user intention. Therefore, the values of the attributes in the UR are coded in terms of three values: 0 (not provided), 1 (provided with high confidence), and 2 (provided with low confidence).

3 The Enhanced UAH Dialogue System

Universidad Al Habla (UAH - University on the Line) is a spoken dialogue system that provides academic information about the Dept. of Languages and Computer Systems at the University of Granada, Spain. The information that the system provides can be classified in four main groups: subjects, professors, PhD courses and student registration [9].

A corpus of 100 dialogues was acquired with this system from student telephone calls. The total number of user turns was 422 and the recorded speech has a duration of 150 minutes. In order to develop an enhanced version of the system that includes the module shown in Figure 1, we carried out two types of corpus annotation: intentional and emotional.

On the one hand, we estimated the user intention for each user utterance by using concepts and attribute-value pairs. One or more concepts represent the intention of the utterance, and a sequence of attribute-value pairs contains the information about the values provided by the user. We defined four concepts to represent the different queries that the user can perform (*Subject*, *Lecturers*, *Doctoral studies*, and *Registration*), three task-independent concepts (*Affirmation*, *Negation*, and *Not-Understood*), and eight attributes (*Subject-Name*, *Degree*, *Group-Name*, *Subject-Type*, *Lecturer-Name*, *Program-Name*, *Semester*, and *Deadline*).

The labelling of the system turns was similar to that for user turns. To do so, 30 concepts were defined and grouped as task-independent concepts (e.g. *Affirmation* and *Negation*), concepts used to inform the user about the result of a specific query (e.g. *Subject* or *Lecturers*), concepts defined to require the user the attributes that are necessary for a specific query (e.g. *Subject-Name*), and concepts used for the confirmation of concepts and attributes. As shown in Figure 2, the UR defined for the task is a sequence of 16 fields corresponding to the concepts and attributes defined for the task and the user profile.

On the other hand, we assigned an emotion category (neutral, doubtful, angry, or bored) to each user utterance. Nine annotators tagged the corpus twice and

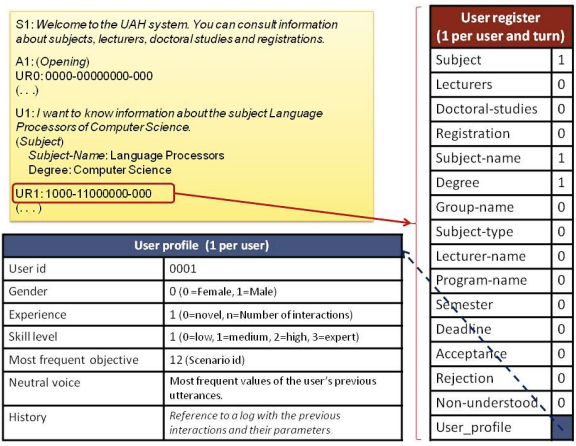


Fig. 2. User Register of the UAH system

the final emotion for each utterance was assigned by majority voting. A detailed description of the annotation procedure and the intricacies of the calculation of inter-annotator reliability can be found in a previous study [7].

Additionally, we have modified the dialogue manager to process the user state information in order to reduce the impact of the user negative states and the user experience on the communication, by adapting the system responses considering user states. The dialogue manager tailors the next system answer to the user state by changing the help providing mechanisms, the confirmation strategy and the interaction flexibility. The conciliation strategies adopted are, following the constraints defined in [10], straightforward and well delimited in order not to make the user loose the focus on the task.

If the recognized emotion is doubtful and the user has changed his behaviour several times during the dialogue, the dialogue manager changes to a system-directed initiative and generates a help message describing the available options. This approach is also selected when the user profile indicates that the user is non-expert (or if there is no profile for the current user), and when their first utterances are classified as doubtful.

In the case of anger, if the dialogue history shows that there have been many errors during the interaction, the system apologizes and switches to DTMF (Dual-Tone Multi-Frequency) mode. If the user is assumed to be angry but the system is not aware of any error, the system's prompt is rephrased with more agreeable phrases and the user is advised that they can ask for help at any time.

In the case of boredom, if there is information available from other interactions of the same user, the system tries to infer from those dialogues what the most likely objective of the user might be. If the detected objective matches the predicted intention, the system takes the information for granted and uses implicit confirmations. For example, if a student always asks for subjects of the same degree, the system can directly disambiguate a subject if it is in several degrees.

In any other case, the emotion is assumed to be neutral, and the next system prompt is decided only on the basis of the user intention and the user profile (i.e., considering user preferences, previous interactions, and expertise level).

4 Experiments

In order to evaluate our proposal, we have recorded the interactions of 6 recruited users. Four of them recorded 30 dialogues (15 scenarios with the baseline system and 15 with the enhanced system), and two of them recorded 15 dialogues (15 dialogues with the baseline or the enhanced system only). Thus, a total of 150 dialogues were recorded in such a way that there were two dialogues recorded per scenario, three in the case of the five most frequent scenarios of the initial UAH corpus.

Table 1. Results of the objective evaluation of the systems

Evaluation metrics	**Baseline**	**Enhanced**
Dialogue success rate	85.0	96.0
Error correction rate	81.0	91.5
Average number of turns per dialogue	12.1	8.1
Average number of actions per turn	1.8	1.5
% of different dialogues (intention only)	85.0	83.5
% of different dialogues (intention and emotion)	85.0	88.0
Number of repetitions of the most seen dialogue	3.5	6
Number of turns of the most seen dialogue	5.5	4.5
Number of turns of the shortest dialogue	4.5	4.5
Number of turns of the longest dialogue	14.5	12.0

As observed in Table 1, on the one hand the success rate for the enhanced system is higher than the baseline. This difference showed a significance of 0.03 in a two-tailed t-test. On the other hand, although the error correction rate is also higher in absolute values in the enhanced system, this improvement is not significant. Both results are explained by the fact that we have not designed a specific strategy to improve the recognition or understanding processes and decrease the error rate. Instead, our proposal for adaptation to the user state overcomes these problems during the dialogue once they are produced.

Regarding the number of dialogue turns, the enhanced system produced shorter dialogues (with a 0.00 significance value in a two-tailed t-test when compared to the number of turns of the baseline system). As shown in Table 1, this general reduction appears also in the case of the longest, shortest and most seen dialogues for the enhanced system. There is also a slight reduction in the number of actions per turn for the dialogues of the enhanced system (with a 0.00 significance value in the t-test). This might be because users have to explicitly provide and confirm more information using the baseline system, whereas the enhanced system automatically adapted the dialogue to the user and the dialogue history.

Regarding the percentage of different dialogues obtained, the rate was lower using the enhanced system, due to an increment in the variability of ways in which users can provide the different data required to the enhanced system. This result was significant when the dialogues were considered different only when they differed in the sequence of observed user intentions, and also when even with the same sequence of intentions, two dialogues were considered different if the emotions observed were different. This is consistent with the fact that the number of repetitions of the most observed dialogues is higher for the baseline system.

With respect to the dialogue participant activity, Figure 3 shows the ratio of user versus system actions. The dialogues of the enhanced system have a higher proportion of system actions due to a reduction of the confirmation turns.

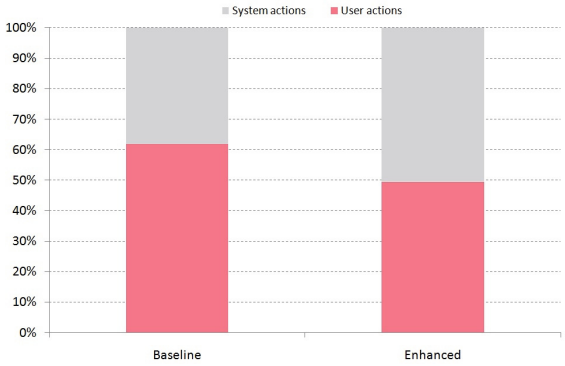


Fig. 3. Ratio of user vs. system actions in the enhanced and baseline systems

Regarding dialogue style and cooperativeness, Figures 4 and 5 respectively show the frequency of the most dominant user and system dialogue acts in the dialogues collected with the enhanced and baseline systems. On the one hand, Figure 4 shows that users need to provide less information explicitly using the enhanced system, which explains the higher proportion of queries (significant over 98%). On the other hand, Figure 5 shows that there is a reduction in the system requests when the enhanced system is used. This explains a higher proportion of system turns to provide information in the enhanced system.

Table 2 shows the average results obtained with respect to the subjective evaluation. As can be observed, both systems correctly understand the different user queries and obtain a similar evaluation regarding the user observed easiness in correcting errors made by the ASR module. However, the enhanced system is judged to be better regarding the user observed easiness in obtaining the data required to fulfill the complete set of objectives defined in the scenario, as well as the suitability of the interaction rate during the dialogue.

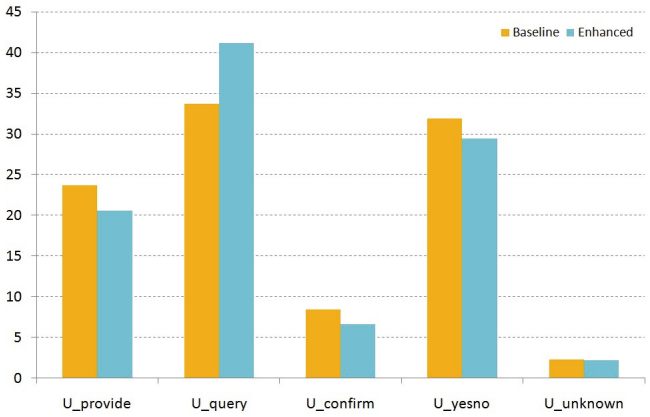


Fig. 4. Histogram of user dialogue acts in the enhanced and baseline systems

Table 2. Results of the subjective evaluation of the systems

Questions (1 to 5 scale)	Baseline	Enhanced
How well did the system understand you?	4.6	4.8
How well did you understand the system messages?	3.6	3.9
Was it easy to obtain the requested information?	3.8	4.3
Was the interaction rate adequate?	3.4	4.2
If the system made errors, was it easy for you to correct them?	3.2	3.3

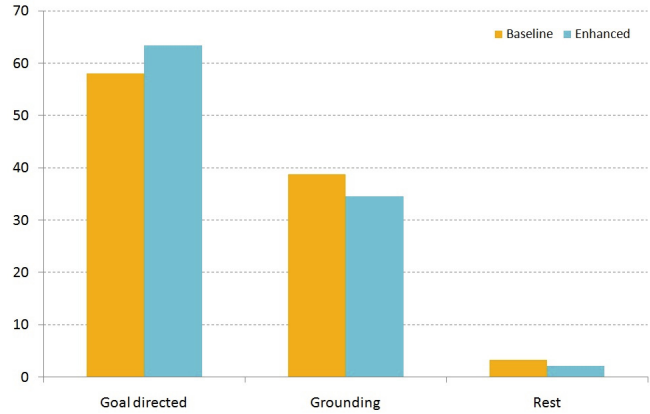


Fig. 5. Histogram of system dialogue acts in the enhanced and baseline systems

5 Conclusions and Future Work

In this paper we have presented a method for predicting user states considering their emotions and intentions, which can be employed to develop more adaptive spoken dialogue systems. We have proposed an architecture in which our method is implemented as a module comprised of an emotion recognizer and an intention recognizer. The former deduces user emotional states from the acoustics of their utterances as well as the dialogue history. The latter decides the next user action and their dialogue goal using a statistical approach that relies on the previous user input and system prompt.

We have evaluated the method with the UAH spoken dialogue system, implementing the prediction module between the system's natural language understanding module and dialogue manager. Additionally, we have improved the dialogue manager to take this information into account in order to compute and adapt the system responses.

The evaluation was carried out using a corpus of interactions of recruited users with the enhanced version of the system. The results show that this version of the system performs better in terms of duration of the dialogues, number of turns needed for successful dialogues, and number of confirmations and repetitions needed. Additionally, the test users judged the system to be better when it could adapt its behaviour to their intentions and emotions.

As a future work we plan to annotate the emotions of the collected corpus in order to refine the adaptation strategies of the dialogue manager.

Acknowledgments. This research has been funded by the Spanish Ministry of Science and Innovation under the Project ASIES TIN2010-17344.

References

1. Riccardi, G., Hakkani-Tür, D.: Grounding Emotions in Human-Machine Conversational Systems. In: Maybury, M., Stock, O., Wahlster, W. (eds.) INTETAIN 2005. LNCS (LNAI), vol. 3814, pp. 144–154. Springer, Heidelberg (2005)
2. Boril, H., Sadjadi, O., Kleinschmidt, T., Hansen, J.: Analysis and detection of cognitive load and frustration in drivers' speech. In: Proc. of Interspeech 2010, Makuhari, Chiba, Japan, pp. 502–505 (2010)
3. Gnjatović, M., Rösner, D.: Adaptive Dialogue Management in the NIMITEK Prototype System. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) PIT 2008. LNCS (LNAI), vol. 5078, pp. 14–25. Springer, Heidelberg (2008)
4. Pittermann, J., Pittermann, A., Minker, W.: Emotion recognition and adaptation in spoken dialogue systems. Int. Journal of Speech Technology 13, 49–60 (2010)
5. Bui, T.H., Poel, M., Nijholt, A., Zwiers, J.: A tractable hybrid DDN-POMDP approach to affective dialogue modeling for probabilistic frame-based dialogue systems. Natural Language Engineering 15, 273–307 (2009)
6. Williams, J.D., Young, S.: Partially observable Markov decision processes for spoken dialog systems. Computer Speech & Language 21, 393–422 (2007)

7. Callejas, Z., López-Cózar, R.: Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication 50, 416–433 (2008)
8. Griol, D., Hurtado, L.F., Segarra, E., Sanchis, E.: A statistical approach to spoken dialog systems design and evaluation. Speech Communication 50, 666–682 (2008)
9. Callejas, Z., López-Cózar, R.: Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication 50, 646–665 (2008)
10. Burkhardt, F., van Ballegooy, M., Engelbrecht, K., Polzehl, T., Stegmann, J.: Emotion detection in dialog systems - Usecases, strategies and challenges. In: Proc. of ACII 2009, Amsterdam, The Netherlands (2009)

Language Technology for Handwritten Text Recognition

Alejandro H. Toselli, Nicolás Serrano, Adrià Giménez-Pastor, Ihab Khoury, Alfons Juan, and Enrique Vidal

DSIC/PRHLT, Universitat Politècnica de València
Camí de Vera, s/n, 46022 València, Spain
{ahector,nserrano,evidal}@iti.upv.es,
{agimenez,ialkhoury,ajuan}@dsic.upv.es

Abstract. This paper shows how the nowadays prevalent technology used in HTR borrows concepts and methods from the field of ASR; i.e. those based on Hidden Markov Models (HMMs). Additionally, it will be described a HTR approach based on employing Bernoulli distributions rather than Gaussian-Mixture distributions for the HMM-state emission probability of observations. Finally, handwritten text recognition evaluation results are reported for several corpora involving different characteristics and languages.

Index Terms: Off-Line Continuous Handwritten Text Recognition, Mixture of Gaussian Densities, Mixture of Bernoulli Distributions. Hidden Markov Model Emission Probability.

1 Introduction

Analogously to Automatic Speech Recognition (ASR), handwritten text image transcription (or "Off-Line" HTR) can be defined as the task of converting handwritten text images into an electronic text format such as ASCII or PDF, which allows taking advantage of the modern text-based storing, typesetting, searching and retrieval technologies.

For some time in the past decades, the interest in Off-line HTR was diminishing, under the assumption that modern computer technologies will soon make paper-based documents useless. However, in the last years, HTR has become an important research topic, specially because of the increasing number of on-line digital libraries publishing large quantities of digitized legacy documents. The vast majority of these documents, hundreds of terabytes worth of digital image data, remain waiting to be transcribed into a textual electronic format that would provide historians and other researchers new ways of indexing, consulting and querying these documents.

HTR should not be confused with OCR (Optical Character Recognition), because in HTR it is generally impossible to reliably isolate the characters or even the words that compose a handwritten text. HTR, specially for historical documents, is a very difficult task. To some extent HTR is comparable with

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 178–186, 2012.

the task of recognizing continuous speech in a significantly degraded audio file. And, in fact, the nowadays prevalent technology for HTR borrows concepts and methods from the field of Automatic Speech Recognition.

In this sense, the handwritten text recognition (HTR) approaches considered here are based on Hidden Markov Models (HMMs). Specifically in this work, two different HTR versions are presented according to the observation emission probability law associated to each HMM-state: one using a mixture of Gaussian densities and another employing a mixture of Bernoulli probability distributions.

Finally, to show that the ASR technology applied to the HTR field is producing promising results, in section 4 are carried out several HTR experiments on several hadwritten document corpora involving different characteristics and languages.

2 Previous Required Preprocessing Steps

Text line images constitute here the basic input of the HTR approaches described in this paper. Therefore, given document page images, first it is necessary to detect their text blocks and after that in turn, to proceed to detect and extract their constituent text line images. In this way, document layout analysis comes to play an important role in this task.

Thus, a preprocessing step entailing background removal, noise reduction, and page skew correction is applied on each document page image before performing on it the text line detection procedure. Actually, this detection process is fully automatically carried out using standard preprocessing techniques based on horizontal and vertical projection profiles [1], and on the run-length smoothing algorithm (RLSA) [2].

3 General Statistical Framework for HTR

As ASR, the continuous handwritten text recognition problem can be also formulated as the problem of finding the most likely word sequence, $\hat{\mathbf{w}} = \langle w_1, w_2, \ldots, w_n \rangle$, for a given handwritten sentence image represented by an observation sequence $\mathbf{o} = \langle o_1, o_2, \ldots, o_m \rangle$, i.e., $\hat{\mathbf{w}} = \arg\max_w \Pr(\mathbf{w} \mid \mathbf{o})$. Using the Bayes' rule we can decompose the probability $\Pr(\mathbf{w} \mid \mathbf{o})$ into two probabilities, $\Pr(\mathbf{x} \mid \mathbf{w})$ and $\Pr(\mathbf{w})$, representing morphological-lexical and syntactical knowledges, respectively:

$$\hat{\mathbf{w}} = \arg\max_w \Pr(\mathbf{w} \mid \mathbf{o}) = \arg\max_w \Pr(\mathbf{o} \mid \mathbf{w}) \cdot \Pr(\mathbf{w})$$

$\Pr(\mathbf{o} \mid \mathbf{w})$ is typically approximated by concatenated character models and $\Pr(\mathbf{w})$ is approximated by a word language model (usually n-grams [3]).

Characters (or graphemes) are considered here as the basic recognition units in the same way as phonemes in ASR, and hence, they are modeled by left-to-right HMMs, with continuous or discrete observation emission probability distribution on each HMM-state. Thereby, the total number of parameters to be estimated depends mainly on the number of HMM-states and their associated

emission probability distributions, which need to be tuned empirically to optimize the overall performance on a given amount of available training samples. As phonemes HMMs are trained from the acoustic data in ASR, character HMMs are trained from images of continuously handwritten text (without any kind of segmentation) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called forward-backward or Baum-Welch re-estimation [3].

Each *lexical entry (word)* is modelled by a stochastic finite-state automaton which represents all possible concatenations of individual characters that may compose the word. By embedding the character HMMs into the edges of this automaton, a *lexical HMM* is obtained. Finally, the concatenation of words into text lines or sentences is modelled by a bi-gram *language model*, with Kneser-Ney back-off smoothing [4,5], estimated from the given transcriptions of the trained set.

Once all the *character*, *word* and *language* models are available, recognition of new test sentences can be performed. Thanks to the homogeneous finite-state (FS) nature of all these models, they can be easily *integrated* into a single *global* (huge) FS model. Given an input sequence of feature vectors, the output word sequence hypothesis corresponds to a path in the integrated network that, with highest probability, produces the input sequence. This optimal path search is very efficiently carried out by the well known (*beam-search*-accelerated) Viterbi algorithm [3]. This technique allows integration to be performed "on the fly" during the decoding process.

On what follows, we will explain in some detail two HTR implemented approaches employing different HMM-state emission probability functions: mixture of Gaussians densities (GHMM) and mixture of Bernoulli distributions (BHMMs) respectively, which model observation sequences of different nature.

3.1 Mixture of Gaussians as State Emission Probabilities

In this case, we employ a HMM-state emission probability distribution given by a mixture of Gaussian densities, which implies working directly with observation sequences in the form of sequences of real-value D-dimensional feature vectors $\langle \boldsymbol{o_1}, \boldsymbol{o_2}, \ldots, \boldsymbol{o_T} \rangle$, $\quad \boldsymbol{o_i} \in \mathcal{R}^D$. These sequences represent (and are extracted from) whole preprocessed handwritten text line images. The probability density function in this case is defined as a weighted sum of K Gaussian distributions as follows:

$$b_j(\boldsymbol{o}) = \sum_{k=1}^{K} c_{jk} \frac{1}{\sqrt{(2\pi)^d |\Sigma_{jk}|}} e^{\left(-\frac{1}{2}(\boldsymbol{o}-\mu'_{jk})\Sigma_{jk}^{-1}(\boldsymbol{o}-\mu_{jk})\right)} \tag{1}$$

where, c_{jk}, μ_{jk} and Σ_{jk} are respectively the weighting coefficient, the mean vector and the covariance matrix for the mixture component k of state j.

The feature extraction process approach used to obtain the feature vectors sequence follows similar ideas described in [6]. First, a grid is applied to divide the text line image into $M \times N$ squared cells. M is chosen empirically and N is such that N/M equals the original line image aspect ratio. Each cell is characterized

by the following features: *average gray level*, *horizontal gray level derivative* and *vertical gray level derivative*. To obtain smoothed values of these features, a $s \times s$ cells analysis window, centered at the current cell, is used in the computations [7]. The smoothed cell-averaged gray level is computed through convolution with two 1-d Gaussian filters. The smoothed horizontal derivative is calculated as the slope of the line which best fits the horizontal function of column-average gray level in the analysis window. The fitting criterion is the sum of squared errors weighted by a 1-d Gaussian filter which enhances the role of central pixels of the window under analysis. The vertical derivative is computed in a similar way.

Columns of cells (also called *frames*) are processed from left to right and a feature vector is constructed for each *frame* by stacking the three features computed in their constituent cells. Hence, at the end of this process, a sequence of N $3 \times M$-dimensional feature vectors is obtained (M normalized gray-level components and M horizontal and vertical derivatives components). Figure 1 shows a representative visual example of the feature vector sequence for the Spanish word "cuarenta" ("forty") and how a continuous density HMM models two feature vector subsequences corresponding to the character "a".

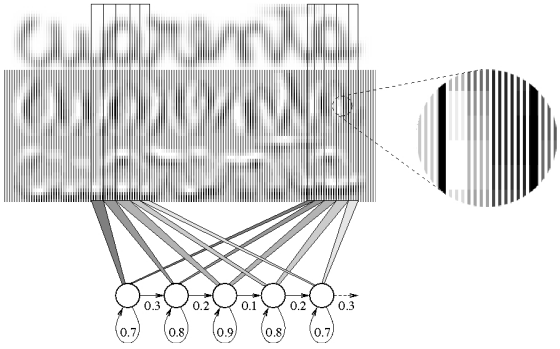


Fig. 1. Example of feature-vector sequence and HMM modeling of instances of the character "a" within the Spanish word "cuarenta" ("forty"). The model is shared among all instances of characters of the same class. The zones modelled by each state show graphically subsequences of feature vectors (see details in the magnifying-glass view) compounded by stacking the normalized grey level and its both derivatives features.

3.2 Mixture of Bernoullis as State Emission Probability

In speech recognition, the use of certain real-valued speech features and embedded Gaussian mixture HMMs is a de-facto standard [8]. However, in the case of HTR, there is no such a standard and, indeed, very different sets of features are in use today. In [9] has been proposed to by-pass feature extraction and to directly feed windows of raw, binary pixels into embedded Bernoulli mixture HMMs. The basic idea is to ensure here that no discriminative information is filtered out during feature extraction, which in some sense is integrated into the recognition model.

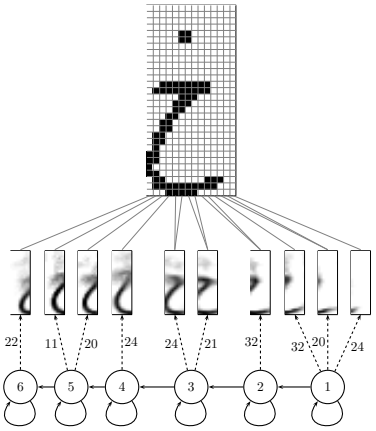


Fig. 2. Example of HMM, using a mixture of 32 Bernoulli components as observation emission probability, for an specific Arabic character extracted from the IfN/ENIT database. Here, each state models an extracted sequence of pixels windows (9 columns size).

A Bernoulli mixture HMM (BHMM) is an HMM in which the probability of observing $\boldsymbol{o}$, when in the state j, is given by a Bernoulli mixture probability function for the state j:

$$b_j(\boldsymbol{o}) = \sum_{k=1}^{K} \pi_{jk} \prod_{d=1}^{D} p_{jkd}^{o_{td}} \left(1 - p_{jkd}\right)^{1-o_{td}} \tag{2}$$

where $\boldsymbol{\pi}_j$ are the priors of the jth state mixture components, and $\boldsymbol{p}_{jk}$ is the kth Bernoulli component prototype in state j. As usual, the probability function (2) can be seen as an emission model which first selects the kth component with probability π_{jk}, and then emits $\boldsymbol{o}$ in accordance with a Bernoulli prototype $\boldsymbol{p}_{jk} \in [0,1]^D$; i.e. with probability p_{jkd} for bit o_d to be 1, for all d.

Given a binary image normalized in height, a feature vector $\boldsymbol{o}_t$ can be seen as a concatenation of columns in a window of W columns in width, centered at position t. Each window is then repositioned by vertically realigning the center with its center of mass. As an example, Figure 2 shows a binary image of an specific Arabic character extracted from the IfN/ENIT database, of 14 columns and 30 rows, which is transformed into a sequence of 10 270-dimensional feature vectors by application of a sliding window of width 9. For clarity, feature vectors are depicted as 9×30 sub-images instead of 270-dimensional column vectors. In addition, to get some insight into the behavior of the Bernoulli HMMs, the model for this specific Arabic character, is (partially) shown in Figure 2 (bottom) together with its Viterbi alignment. Here, the Bernoulli prototypes are represented as grey images where the grey level of each pixel measures the probability of its corresponding pixel to be black (white = 0 and black = 1).

4 Experimental Results

4.1 Corpora for Transcription Tasks

In order to asses the effectiveness of the above-presented off-line HTR systems, six corpora with more less similar HTR difficulty were employed in the experiments. The first two, ODEC-M3 [10] and IAMDB [11], contain handwritten text in modern Spanish and English, respectively. IAMDB is publicly available, thereby serving as a reference to compare the obtained results. The following three corpora: CS [12], Germana [13] and Rodrigo [14], consist of cursive handwritten page images in old Spanish from 19th and 16th century. The last corpus: IfN/ENIT [15], database of Arabic handwritten Tunisian town names, which comprises more than 32K Arabic words written by more than one thousand different writers, from a lexicon of one thousand Tunisian town/village names. Figure 3 shows examples of each of them.

Fig. 3. From top-to-bottom and left-to-right: Handwritten Text from the Lancaster-Oslo/Bergen Corpus (IAMDB), Answers extracted from Survey Forms made for a Telecommunication Company (ODEC-M3), Single-Writer Manuscripts from the XIX Century (CS and Germana), Single-Writer Spanish manuscript from XVI century (Rodrigo) and multi-writer Arabic manuscript forms of Tunisian town/village names (IfN/ENIT)

4.2 Results

The quality of the transcriptions obtained with the off-line HTR system is given by the word error rate (WER). The corresponding morphological (HMMs) and language (bi-gram) models associated with each corpus were trained from their respective training images and transcriptions. Besides, all results reported in Table 1 have been obtained after optimizing the parameters values corresponding to the preprocessing, feature extraction and modelling processes for each of the tasks.

Concerning to the HTR using Gaussians mixture as HMM-state emission probabilities, on the ODEC-M3, IAMDB and CS corpora, the obtained WER(%) results were 22.9%, 35.5% and 28.5% respectively, using for all these cases a closed-vocavulary. For the GERMANA corpus, the best WER achieved is around

Table 1. Basic statistics information from each corpus along with the WER(%) obtained using the off-line HTR with Gaussians and Bernoulli mixtures as HMM-state emission probability functions (GHMMs and BHMMs)

Corpus		IAMDB	ODEC	CS	GERMANA	RODRIGO	IFN/ENIT
Language		English	Spanish	19th C Sp.	19th C Sp.	16th C Sp.	Arabic
Writters		many	many	1	1	1	many
Lan. Model	Lexicon	8 017	2 790	2 277	7 477	17.3K	937
	Train. Ratio	128	4.4	2.8	4.5	12.5	25
HMMs	Characters	78	80	78	82	115	120
	Train. Ratio	2 779	808	470	2 309	7 930	1 410
Open Vocabulary		N	N	N	N **Y**	**Y**	N
WER (%)	(GHMMs)	**35.5**	**22.9**	**28.5**	**8.9** **26.9**	**21.2**	–
WER (%)	(BHMMs)	**34.3**	–	–	–	–	**6.2**

8.9% and 26.9% using close- and open-vocabulary respectively. Regarding the OOV words, it becomes clear that a considerable fraction of transcription errors is due to the occurrence of unseen words in the test partition. More precisely, unseen words account here for approximately 50% of transcription errors. Although comparable in size to GERMANA, RODRIGO comes from a much older manuscript (from 1545), where the typical difficult characteristics of historical documents are more evident. The best WER figure achieved in this corpus until the moment is around 21.2%, where most of the errors are also caused by the occurrence of out-of-vocabulary words.

Concerning to the HTR approaches using Bernoulli mixture as HMM-state emission probability functions, a WER of 34.3% was attained, which is slightly better than the 35.5% obtained with a similar system based on Gaussian HMMs. However, for the IFN/ENIT database, it can be seen that the achieved 6.2% of

WER with this BHMM system, outperforming by far the 14.6% obtained in the ICDAR 2009 competition using this database.

5 Conclusions

Two off-line HTR systems based on Hidden Markov Models using Gaussians and Bernoullis Mixture as HMM state emission probabilty functions have been presented. The HMM-based HTRs has a hierarchical structure with character models at the lowest level. These models are concatenated to words and to whole sentences. The HMM used in this work was furthermore enhanced by a 2-gram incorporating linguistic information beyond the word level.

Several tasks have been considered to assess these both HTR approaches. In spite of the extreme difficulty that entails the corpora used in the experiments, the results achieved are really encouraging.

Acknowledgements. Work supported by the EC (FEDER), the Spanish MEC under the MIPRCV "Consolider Ingenio 2010" research programme (CSD2007-00018) and the Spanisg Government (MICINN and "Plan E") under the MIT-TRAL (TIN2009-14633-C03-01) research project.

References

1. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. International Journal on Document Analysis and Recognition 9, 123–138 (2007)
2. Wong, K.Y., Wahl, F.M.: Document analysis system. IBM Journal of Research and Development 26, 647–656 (1982)
3. Jelinek, F.: Statistical methods for speech recognition. MIT Press (1998)
4. Katz, S.M.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. In: Proceedings of the IEEE Transactions on Acoustics, Speech and Signal Processing (ICASSP 1987), vol. ASSP-35, pp. 400–401(March 1987)
5. Kneser, R., Ney, H.: Improved backing-off for n-gram language modeling. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 181–184 (1995)
6. Bazzi, I., Schwartz, R., Makhoul, J.: An omnifont open-vocabulary OCR system for English and Arabic. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(6), 495–504 (1999)
7. Toselli, A.H., Juan, A., Keysers, D., González, J., Salvador, I., Ney, H., Vidal, E., Casacuberta, F.: Integrated handwriting recognition and interpretation using finite-state models. International Journal of Pattern Recognition and Artificial Intelligence 18(4), 519–539 (2004)
8. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice-Hall, Englewood Cliffs (1993)
9. Giménez, A., Juan, A.: Embedded bernoulli mixture hmms for handwritten word recognition. In: Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, pp. 896–900. IEEE Computer Society (July 2009)

10. Toselli, A., Juan, A., Vidal, E.: Spontaneous handwriting recognition and classification. In: Proceedings of the International Conference on Pattern Recognition (ICPR 2004), Cambridge, United Kingdom, vol. 1, pp. 433–436 (August 2004)
11. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition (IJDAR) 5(1), 39–46 (2002)
12. Romero, V., Toselli, A.H., Rodríguez, L., Vidal, E.: Computer Assisted Transcription for Ancient Text Images. In: Kamel, M.S., Campilho, A. (eds.) ICIAR 2007. LNCS, vol. 4633, pp. 1182–1193. Springer, Heidelberg (2007)
13. Pérez, D., Tarazón, L., Serrano, N., Castro, F.M., Ramos-Terrades, O., Juan, A.: The germana database. In: Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, pp. 301–305. IEEE Computer Society (July 2009)
14. Serrano, N., Juan, A.: The rodrigo database. In: Proceedings of the The Seventh International Conference on Language Resources and Evaluation (LREC 2010), Malta, May 19-21 (2010)
15. Pechwitz, M., Maddouri, S.S., Magïer, V., Ellouze, N., Amiri, H.: IFN/ENIT-database of handwritten Arabic words. In: Proc. of the Colloque International Francophone sur l'Ecrit et le Document (CIFED), Hammmamet, Tunisia (October 2002)

Character-Based Handwritten Text Recognition of Multilingual Documents

Miguel A. del Agua, Nicolás Serrano, Jorge Civera, and Alfons Juan

DSIC/ITI, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{mdelagua,nserrano,jcivera,ajuan}@dsic.upv.es

Abstract. An effective approach to transcribe handwritten text documents is to follow a sequential interactive approach. During the supervision phase, user corrections are incorporated into the system through an ongoing retraining process. In the case of multilingual documents with a high percentage of out-of-vocabulary (OOV) words, two principal issues arise. On the one hand, a minor yet important matter for this interactive approach is to identify the language of the current text line image to be transcribed, as a language dependent recognisers typically performs better than a monolingual recogniser. On the other hand, word-based language models suffer from data scarcity in the presence of a large number of OOV words, degrading their estimation and affecting the performance of the transcription system. In this paper, we successfully tackle both issues deploying character-based language models combined with language identification techniques on an entire 764-page multilingual document. The results obtained significantly reduce previously reported results in terms of transcription error on the same task, but showed that a language dependent approach is not effective on top of character-based recognition of similar languages.

1 Introduction

Have not been until recently when large volumes of old handwritten documents have undergone an image digitalisation process in order to give general public access to this new source of information. However, digitalised handwritten documents cannot be fully exploited by natural language processing (NLP) tools, if texts are not available in electronic format. For this reason, a continuous time-consuming transcription effort is nowadays being carried out by digital libraries.

To alleviate this effort, automatic handwriting transcription techniques based on speech recognition technology have flourished over the last years, although the quality of the transcriptions provided by these techniques is still far from not being in need of supervision [1]. An effective approach to supervision is to integrate an ongoing retraining system that interactively incorporates user corrections once a line has been reviewed. Such a system, along with layout analysis and line detection features, has been implemented in an open source tool called *Gimp-based Interactive transcription of old text DOCuments* (GIDOC) [2].

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 187–196, 2012.

GIDOC has been used as a platform to develop techniques aimed at reducing user effort and maximise its usability. These techniques range from adapting models from partially supervised transcriptions [3], over an adequate trade-off between error and supervision effort [4], to a variety of active learning strategies to improve the interaction with the user on each new system hypothesis [5].

A specially appealing case in automatic handwritten text recognition is the transcription of multilingual documents. A good example of multilingual document is the GERMANA database [6]. GERMANA is the result of digitizing and annotating a 764-page, single-author manuscript from 1891, written in Spanish up to page 180, but then also written in five other languages, mainly in Catalan and Latin. Another distinctive feature of GERMANA is the large number of out-of-vocabulary (OOV) words accentuated by its multilingual nature. This feature has been the main reason for the relatively poor results obtained so far on the GERMANA database [7].

The work presented in this paper targets both characteristic features of the GERMANA database: Multilinguality and OOV words. Multilinguality is captured by language identification models already discussed in [7]. The problem of OOV words is tackled by deploying character-based n-gram language models. As a consequence, the reported results are the best ever achieved on the GERMANA database.

The rest of this paper is structured as follows. Previous work related to multilinguality and character-based language modelling in speech and handwriting recognition is reviewed in Section 2. In Section 3, the probabilistic framework for language identification on a character-based handwriting recognition approach is presented. Section 4 is devoted to empirical results on the whole GERMANA manuscript. Finally, conclusions and future work are discussed in Section 5.

2 Previous Work

Multilinguality in handwritten text recognition arises the challenge of taking advantage of language identification in order to interactively adapt the underlying models of the system and to minimise transcription errors. However, conventional (non-interactive) script and language identification are still in its early stage of research [8], and have remained unexplored until very recently [7].

Preliminary results exploiting multilinguality on the GERMANA database proved the benefits of explicitly modelling language identification at the line level in a interactive transcripcion scenario [7]. However, these results are far from allowing an effective interactive transcription. In that work, the supervision effort would be excesively high, and the user might prefer to ignore the automatically generated output and transcribe the manuscript from scratch. An error analysis revealed that most of these errors were due to out-of-vocabulary (OOV) words. In fact, 53% to 71% of the words in the GERMANA database are singletons, words ocurring only once in the lexicon of each language. Another important problem was the scarse resources available for some languages in the GERMANA database, so as to train their corresponding word-based language models.

The treatment of OOV words is an open problem in different areas of NLP. In speech recognition, which is closely related to handwritten text recognition as far as modelisation is concerned, notable efforts has been deployed over the last decades to deal with OOV words. In [9], the original lexicon is extended with words from external resources that are represented as a sequence of characters (graphemes, to be more precise) converted into phonemes. In [10], several subword based methods for spoken term detection task and phone recognition are presented to search OOV words. Phone and multigram-based systems provide similar performance on the phone recognition task, superseding the standard word-based system.

Regarding handwriting text recognition, the authors in [11] compared the performance of a conventional word-based language model to that of a character-based language model in the context of a German offline handwritten text recognition task. However, character-based language models were not superior to their word-based counterparts. A hybrid approach between a standard character-based n-gram language model and a character-based connectionist language model is proposed in [12], which obtain similar results to word-based systems on the IAM corpus [13].

To the best of our knowledge, character-based language models has not been able so far to supersede word-based language models in handwritten text recognition. Our hypothesis is that tasks tackled in previous work did not contain a significant number of OOV words compared to the figures of the GERMANA database[1]. In GERMANA, the problem of OOV words is aggravated by its multilingual nature, since the presence of languages such as Latin, French, German and Italian is less than 4% of the total number of words. Therefore, the estimation of word-based language models is notably poor, and it is necessary to fall back to adequate character-based language models.

3 Probabilistic Framework

Let t be the number of the current text line image to be transcribed, and let x_t be its corresponding sequence of feature vectors. The task of our system is to predict for each text line image first its language label, l_t, and then its transcription, c_t. We assume that all preceding lines have been already annotated in terms of language labels, l_1^{t-1}, and transcriptions, c_1^{t-1}. By application of the Bayes decision rule, the minimum-error system prediction for l_t is:

$$\begin{aligned} l_t^*(x_t, l_1^{t-1}) &= \underset{\tilde{l}_t}{\operatorname{argmax}}\ p(\tilde{l}_t \mid x_t, l_1^{t-1}) \\ &= \underset{\tilde{l}_t}{\operatorname{argmax}}\ p(\tilde{l}_t \mid l_1^{t-1})\, p(x_t \mid \tilde{l}_t) \end{aligned} \tag{1}$$

where in Eq. (1), it is assumed that x_t is conditionally independent of all preceding language labels, l_1^{t-1}, given the current line language label, $\tilde{l}_t$. For the

[1] For example, the IAM corpus only contains about 7% of OOV words.

term $p(x_t \mid \tilde{l}_t)$, we marginalise over all possible character-based transcriptions for language l_t, that is, $C(\tilde{l}_t)$

$$p(x_t \mid \tilde{l}_t) = \sum_{\tilde{c}_t \in C(\tilde{l}_t)} p(\tilde{c}_t \mid \tilde{l}_t)\, p(x_t \mid \tilde{l}_t, \tilde{c}_t) \tag{2}$$

$$\approx \max_{\tilde{c}_t \in C(\tilde{l}_t)} p(\tilde{c}_t \mid \tilde{l}_t)\, p(x_t \mid \tilde{l}_t, \tilde{c}_t). \tag{3}$$

Eq. (3), the Viterbi (maximum) approximation to the sum in Eq. (2), is applied to only consider the most likely transcription. It must be noted that, this language identification technique is one of the most effectives in Automatic Speech Recognition (ASR) [14].

The decision rule (1) requires a *language identification model* for $p(\tilde{l}_t \mid l_1^{t-1})$ and, for each possible language $\tilde{l}_t$, a $\tilde{l}_t$-dependent *character-based language model* for $p(\tilde{c}_t \mid \tilde{l}_t)$ and a $\tilde{l}_t$-dependent *image model* for $p(x_l \mid \tilde{l}_t, \tilde{c}_t)$.

A series of n-gram language identification models were proposed in [7]. In this work, we applied the best performing models, the unigram model

$$\hat{p}(\tilde{l}_t \mid l_{t-1}) = \frac{N(\tilde{l}_t)}{t-1} \tag{4}$$

and the bigram model

$$\hat{p}(\tilde{l}_t \mid l_{t-1}) = \frac{N(l_{t-1}\, \tilde{l}_t)}{N(l_{t-1})}, \tag{5}$$

both estimated by relative frecuency counts, where $N(\cdot)$ denotes the number of occurrences of a given event in the preceding lines, such as the bigram $l_{t-1}\, \tilde{l}_t$ or the unigram $\tilde{l}_t$. It should be noticed that the bigram model makes use of prior knowledge about the GERMANA database, assuming that consecutive lines are usually written in the same language.

A character-based language model for each language $p(\tilde{c}_t \mid \tilde{l}_t)$ is implemented as a conventional *n-gram language model* [15], but considering characters instead of words. Each $\tilde{l}_t$-dependent language model is trained only from those transcriptions labeled with $\tilde{l}_t$. In the case of character-based n-gram language models, the order of the n-gram is normally higher than that employed in word-based models. The aim is to capture information not only regarding intra-word character sequence, but also inter-word relationship, and word tokenisation and segmentation. This information is specially useful in the transcription of OOV words.

Image models for the different languages are implemented in terms of *character HMMs* [2]. Taking advantage that only a single script is used for all the languages considered in the GERMANA database (e.g. Latin), a unique, shared image model is estimated.

Finally, it is often useful in practice to introduce scaling parameters in the decision rule so as to empirically adjust the contribution of the different models involved. In our case, the decision rule given in Eq. (3) can be rewritten as

$$l_t^*(x_t, l_1^{t-1}) \approx \operatorname*{argmax}_{\tilde{l}_t} p(\tilde{l}_t \mid l_1^{t-1})^{\beta} \max_{\tilde{c}_l \in C(\tilde{l}_t)} p(x_t \mid \tilde{l}_t)^{\alpha_{\tilde{l}_t}} \tag{6}$$

being

$$p(x_t \mid \tilde{l}_t)^{\alpha_{\tilde{l}_t}} = p(\tilde{c}_t \mid \tilde{l}_t)^{\alpha_{\tilde{l}_t}}\, p(x_t \mid \tilde{l}_t, \tilde{c}_t) \tag{7}$$

where we have introduced an *Identification Scale Factor (ISF)* β and, for each language $\tilde{l}_t$, a language-dependent *Grammar Scale Factor (GSF)* $\alpha_{\tilde{l}_t}$. In the experiments reported below, these parameters are tuned on a validation set.

4 Experiments

Experiments were performed in the GERMANA database [6]. GERMANA is a single-author manuscript from 1891, which contains 764 pages written in up to six different languages. Our main objective is to study the use of character-based models in an interactive transcription task. As it has been said, the utilization of character-based models is motivated by two main features of GERMANA: the high number of OOVs, and the resource scarcity to train robust word language models. In addition, we analyze the performance of the language identification techniques presented in previous section.

Some basic yet precise statistics of GERMANA are given in Table 1. In terms of running words, Spanish comprises about 81% of the document, followed by Catalan (12%) and Latin (4%), while the other three languages only account for less than a 3%. Similar percentages also apply for the number of lines. In terms of lexicons, it is worth noting that Spanish and, to a lesser extent, Catalan and Latin, have lexicons comparable in size to standard databases, such as IAM [13]. Also note that the sum of individual lexicon sizes (29.9K) is larger than the size of the global lexicon (27.1K). This is due to presence of words common to different languages, such as Spanish and Catalan. On the other hand, singletons, that is, words occurring only once, account for most words in each lexicon ($55\% - 71\%$). It goes without saying that, as usual, language modelling is a difficult task. To be more precise, in Table 1 we have included the global perplexity and the perplexity of each language, as given by an optimised language model on a 10-fold cross-validation experiment.

Table 1. Basic statistics of GERMANA

Language	Lines	Running	Chars	Lexicon	Perplexity
All	20151	1.08M		121	13.1 ± 0.61
Spanish	80.9%	81.2%		114	12.24± 0.15
Catalan	11.8%	11.7%		93	10.39± 0.34
Latin	4.6%	5.2%		91	10.44± 0.36
French	1.3%	1.3%		79	10.96± 0.81
German	1.1%	0.4%		61	10.17± 0.20
Italian	0.3%	0.3%		61	9.44± 0.24

In our experiments, we followed an interactive transcription framework, where the user supervises the output of a system, which is continuosly retrained. To

this purpose, we divided GERMANA in blocks of 500 lines, numbered from 1 to 40. First, blocks number 1 and 2 were manually transcribed and used to build an initial system and tune the training and recognition parameters. Training parameters, such as number of mixture components and states per HMM, remains unchanged in all experiments. Then, starting from block number 3 to the last. First, the language of each line is identified (if needed) and its transcriptions is recognised by the corresponding language dependent system. Next, its transcription and language label is supervised. Finally, after a full new block is supervised, the system is re-trained from all supervised blocks and adapted on the last supervised block. It must be noted that, HMMs image modeling is carried out by the RWTH ASR toolkit [16] and language modeling by SRILM toolkit [15]. We performed two different sets of experiments on the described framework. The objective of the first set was to study the performance of the language identification methods proposed. On other hand, the objective of the second set was to study the transcription accuracy of the system when using each different language identification method.

In the first set of experiments, we compared three different approaches for language identification: *CPL* (simply assigns to a given line the language of the previous one), *unigram* (uses Eq. 4) and *bigram* (uses Eq. 5). We performed the interactive transcription of GERMANA using described framework for each of the approaches. Each time a block is recognised, we measured the number of errors committed by the language identification method used. It must be noted that, in this set of experiments, parameters were tuned to minimise the number of language identification errors. Table 2 shows the results in terms of language identification error-rate (IER) for the whole document. We also included the results on the same framework of the word-based approach presented in [7].

Table 2. Language identification results on GERMANA

System	CPL	Unigram	Bigram
Character-based	2.5	14.2	4.0
Word-based		15.9	5.0

From the results in Table 2, it can be observed that CPL achieved the best performance. CPL took fully advantage of document sequentiality and it only committed errors when the language changed from line to line, which only occurs a few times in GERMANA. In both, character and word based systems, the bigram approach tunned its parameters to ignore the language dependent recogniser probability in Eq. 7 and it forces the system to only relay on the language model probability of language labels. In this case, the bigram approach identifies the language only using the bigram probability. However, the bigram approach only adapts its parameters each time a block is supervised, and thus, it fails to identify all lines of a language when it appears the first time in the transcription process. On the other hand, the character-based unigram approach achieved slightly better results than its word-based version.

In the second set of experiments, we compared five different approaches in terms of Word Error Rate (WER) on recognised transcriptions. WER is defined as the ratio between the minimum number of editing operations to convert the recognised words into the reference, and the number of reference words. In the first approach, we built a *monolingual* system, where we assume all lines to belong to the same language. This approach is considered the baseline, as language identification step is not needed and it is the simplest aproximation to the problem. Next, motivated from the results in [7], we also built four different language dependent systems, which differ on which language identification method is used to switch on the proper language dependent recogniser. All the language dependent systems shared the same HMM image models but differ on their language models, which are only trained from the transcriptions of their corresponding languages. These multilingual systems are named as: *supervised* (language label is manually given), *CPL* (copy previous label), *bigram* (using Eq. (5)), and *unigram* (using Eq. (4)). It must be noted that, in this case, all approaches adapted their parameters to optimize the WER on last block. As the unigram and bigram approaches can be optimized for WER or IER, we also compared the results of both optimizations when transcribing, as the transcriptions produces are different. The results are represented in Fig. 1, in terms of WER of the recognized text up to the current line.

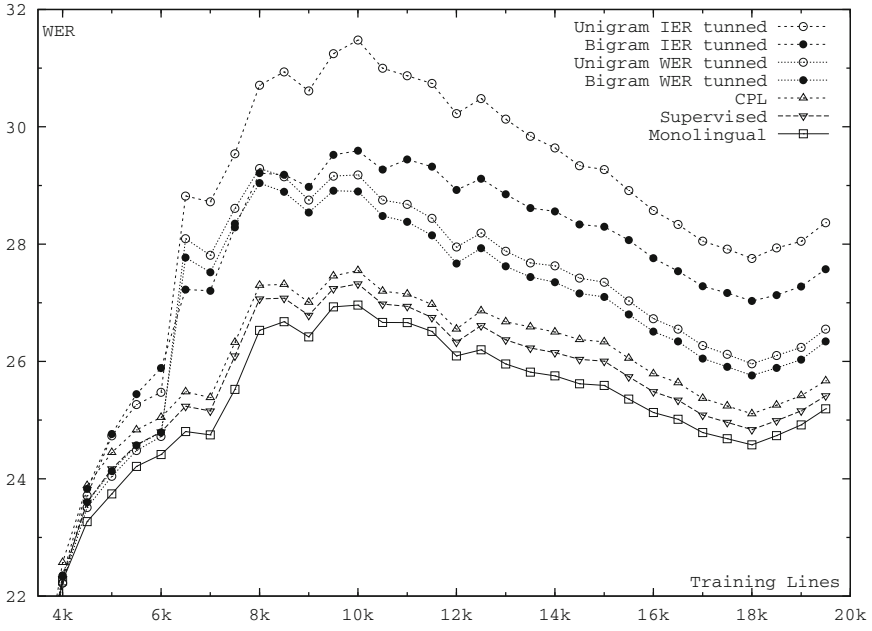


Fig. 1. WER in GERMANA as a function of the number of recognized lines for the monolingual and language-dependent approaches. Results are presented from line 3500, in which a different language apart from Spanish appears.

On the contrary, as it happened in [7], all multilingual systems achieved worse results than the monolingual system. However, even though there is not significant difference between the three best approaches, as corroborated by a bootstrap evaluation [17]; the monolingual approach is considered the best as it is easier to build and it does not need a language identification step in recognition. In error mean terms, even in the supervised approach, where the language is given, the use of language dependent recognizers could not outmatch the monolingual approach. The main cause of the monolingual performance is produced by the origin of all languages but German in GERMANA. Most languages in this document are *Romance* languages, which come from the same original language, sharing a common underlying language structure. For instance, the lexeme of many words can be correctly estimated from the Spanish part in order to recognise other similar romance languages, such as Catalan. In fact, the main responsible of the monolingual result is the high order (9-grams) character-based language model, which was able to estimate the common lexeme structure of all romance languages.

In language dependent approaches, it can be observed that, even though both supervised and CPL approaches achieved the best transcription results, the system performance did not always depend on the language identification performance. On one hand, there is not always a direct relationship between IER and WER. For instance, the unigram and bigram IER optimised approaches achieved a IER of 14.2 and 4.0, respectively, while the WER results were 28.36 and 27.57. On the other hand, as observed from the difference between the different optimizations of unigram and bigram approaches, a system with a worse IER can obtain a better WER results. For example, the bigram WER optimised approach obtained 26.34 of WER from a IER of 8.5, while optimising the IER on the same approach achieved 27.57 of WER from a IER of 4. These results corroborate our conclusions in [7], in which we observed that a language is better recognised using a different language dependent recogniser. However, as said, the monolingual approach achieved better recognition results because the improvement from better estimated languages is already included in the character-based language model.

In terms of transcription performance, in our previous work [7], we also dealt with the complete transcription of GERMANA, but using word-based models. In that case, the monolingual approach obtained 44.39% of WER, however, in this work the same approach obtains 25.19%. These improvement is caused by two factors. On one hand, the RWTH recogniser improved the results due to a new feature extraction method. On the other hand, further error analysis revealed that, as expected, most of this improvement is due to the correct recognition of OOVs words, and punctuation signs. In Figure 2, we can observe the performance of both models in the recognition of a line, concretely, in this example, word-based errors ("estado", "Viuda", and "reflejasen") occured due to OOVs words ("citado", "Vidal", and "refleja"). On the other hand, punctuation signs ("," after "Vidal" and "Reina"), are successfully recognized in the character-based approach, whereas, the word-based approach failed to recognize this signs due

to its scarcity in the training dataset. In past works [6], we only dealt with GERMANA first part, where we reported a performance of 34.51% of WER, in this same partition, the character-based system obtained a performance of 12.12% WER.

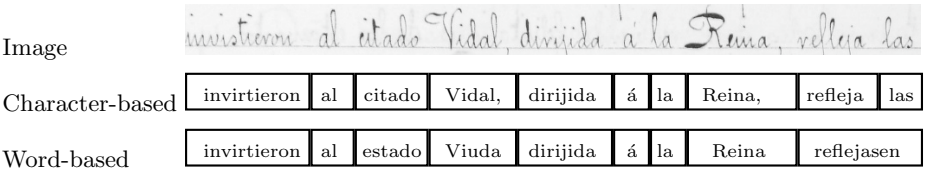


Fig. 2. Comparison of word-based and character-based recognition

5 Conclusions and Future Work

We have proposed a character-based approach for interactive transcription of multilingual documents. This approach is motivated by the high number of OOV words in these handwritten text documents. In addition, we have adapted our previous probabilistic framework for language identification in interactive transcription of multilingual documents to be use in a character-based system. Empirical results are presented on the whole GERMANA database, a 764-page, single-author manuscript from 1891 written in up to six different languages. Two different sets of experiments were performed: language identification and automatic recognition experiments. According to the empirical results, in terms of language identification, the simplest technique, that is, the "copy the preceding label" (CPL) bigram model is also the most accurate. On the other hand, in terms of transcription performance, the monolingual approach achieved the best results. This is mainly caused by the use of character-based language models, which successfully estimates the underlying structure of similar languages. We also observed that language identification results did not always correlate with transcription results, and that the use of a language dependent recogniser was not needed in the transcription task proposed. However, a language dependent approach can be useful when dealing with very different languages, which structure do not share any similarities. In addition, the monolingual language model was build from the concatenation of all transcription. A more adequate approach would be to create a mixture of language dependent models, which could improve the monolingual results. Transcription of other multilingual documents remains as future work to better generalise the effectiveness of the presented approach.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287755. Also supported by the Spanish Government (MIPRCV "Consolider Ingenio 2010", iTrans2 TIN2009-14511, MITTRAL TIN2009-14633-C03-01 and FPU AP2007-0286) and the Generalitat Valenciana (Prometeo/2009/014).

References

1. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(5), 855–868 (2009)
2. Serrano, N., Tarazón, L., Pérez, D., Ramos-Terrades, O., Juan, A.: The GIDOC prototype. In: Proc. of the 10th Int. Workshop on Pattern Recognition in Information Systems (PRIS 2010), Funchal, Portugal, pp. 82–89 (2010)
3. Serrano, N., Pérez, D., Sanchis, A., Juan, A.: Adaptation from Partially Supervised Handwritten Text Transcriptions. In: Proc. of the 11th Int. Conf. on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009), Cambridge, MA, USA, pp. 289–292 (2009)
4. Serrano, N., Sanchis, A., Juan, A.: Balancing error and supervision effort in interactive-predictive handwriting recognition. In: Proc. of the Int. Conf. on Intelligent User Interfaces (IUI 2010), Hong Kong, China, pp. 373–376 (2010)
5. Serrano, N., Giménez, A., Sanchis, A., Juan, A.: Active learning strategies in handwritten text recognition. In: Proc. of the 12th Int. Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), Beijing, China, vol. (86) (November 2010)
6. Pérez, D., Tarazón, L., Serrano, N., Castro, F., Ramos-Terrades, O., Juan, A.: The GERMANA database. In: Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain, pp. 301–305 (2009)
7. del Agua, M.A., Serrano, N., Juan, A.: Language Identification for Interactive Handwriting Transcription of Multilingual Documents. In: Vitrià, J., Sanches, J.M., Hernández, M. (eds.) IbPRIA 2011. LNCS, vol. 6669, pp. 596–603. Springer, Heidelberg (2011)
8. Ghosh, D., Dube, T., Shivaprasad, P.: Script Recognition: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 32(12), 2142–2161 (2010)
9. Bisani, M., Ney, H.: Open vocabulary speech recognition with flat hybrid models. In: Proc. of the European Conf. on Speech Communication and Technology, pp. 725–728 (2005)
10. Szoke, I., Burget, L., Cernocky, J., Fapso, M.: Sub-word modeling of out of vocabulary words in spoken term detection. In: IEEE Spoken Language Technology Workshop, SLT 2008, pp. 273–276 (December 2008)
11. Brakensiek, A., Rottl, J., Kosmala, A., Rigoll, G.: Off-Line handwriting recognition using various hybrid modeling techniques and character N-Grams. In: 7th International Workshop on Frontiers in Handwritten Recognition, pp. 343–352 (2000)
12. Zamora, F., Castro, M.J., España, S., Gorbe, J.: Unconstrained offline handwriting recognition using connectionist character n-grams. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (July 2010)
13. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. IJDAR, 39–46 (2002)
14. Schultz, T., Kirchhoff, K.: Multilingual Speech Processing (2006)
15. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. of ICSLP 2002, pp. 901–904 (September 2002)
16. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., Ney, H.: The RWTH aachen university open source speech recognition system. In: Interspeech, Brighton, U.K., pp. 2111–2114 (September 2009)
17. Efron, B., Tibshirani, R.J.: An Introduction to Bootstrap. Chapman & Hall/CRC (1994)

A Robust Pitch Extractor Based on DTW Lines and CASA with Application in Noisy Speech Recognition

Juan A. Morales-Cordovilla[1], Pablo Cabañas-Molero[2], Antonio M. Peinado[1], and Victoria Sánchez[1,⋆]

[1] Dept. of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Spain
[2] Dept. of Engeniería de la Telecomunicación, Universidad de Jaén, Spain
{jamc,amp,victoria}@ugr.es,cabanas@ujaen.es

Abstract. This paper proposes a robust pitch extractor with application in Automatic Speech Recognition and based on selecting pitch lines of a tonegram (a representation of the different pitch energies at each frame time). First, the tonegram and its maximum energy regions are extracted and a Dynamic Time Warping algorithm finds the most energetic trajectories or pitch lines from these regions. A second stage estimates the tonegram of the most energetic lines by applying Computational Auditory Scene Analysis rules which reject and group octave-related lines. The mean pitch of the speaker is estimated and the final pitch is estimated by rejecting lines which are outside from the mean pitch. The proposed pitch extractor is evaluated in a novel way - by means of the word accuracy of a Missing Data recognizer on Aurora-2 database.

Keywords: pitch extractor, pitch line, CASA, DTW, noise, robust speech recognition.

1 Introduction

Acoustic noise represents one of the major challenges for Automatic Speech Recognition (ASR) systems. Many different approaches have been proposed to deal with this problem [10,1,3] but if we consider voiced speech (i.e. not whispering speech) and the manner in which the auditory system works, pitch information can be a very useful cue to separate noise from speech and to obtain high performance in ASR [5,7,8].

One of the main challenges for pitch-based ASR techniques is that they need a robust pitch extractor. We can distinguish two stages in pitch extractors: a frame stage that obtains the pitch (or pitches) at each frame, and a post-processing stage which produces a final pitch decision. The result of the first stage is a representation indicating at each instant time, the energy or probability of observing

⋆ This work has been supported by the Spanish MEC/FEDER project TEC2010-18009 and partially funded by the DIRHA European project FP7-ICT-2011-7-288121.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 197–206, 2012.

the different pitch values. We will call *tonegram* to this representation and different tools such as difference-function [2], comb-filter [4] or auto-correlogram [5] can be employed to obtain it. The post-processing stage tries to estimate the final pitch by employing this tonegram and rules which help to distinguish the target pitch from possible noise pitches. The continuity and smoothness of pitch lines is the most common rule for speech signals as it is shown by the Hidden Markov Models (HMMs) or mode filters which many of the pitch extractors have [5,8]. In addition, Computational Auditory Scene Analysis (CASA) rules, such as common limits (onset/offset) or even high level information [5], have been applied in order to group spectro-temporal pixels of the spectrogram and to obtain, as a result, a final pitch decision.

The goal of the paper is to show how the pitch lines can be extracted from a tonegram by means of a Dynamic Time Warping (DTW) approach, and how a final pitch decision can be obtained by means of a post-processing, inspired on CASA rules, of these lines. The advantage of working with pitch lines is that it let us associate to the lines different features (such as intensity, mean-pitch, space-localization, etc.) and later select the lines which fullfill the features of target speaker.

The structure of the paper is as follows. First, a block diagram gives an overview of the pitch extractor. Sec. 3 explains the proposed pitch extractor in greater detail. Sec. 4 presents the experimental framework and the Aurora-2 results by using a pitch-based Missing Data (MD) technique for ASR. The paper concludes with a summary and a discussion of future work.

2 System Overview

The pitch extractor (Fig. 1) has a noisy signal of an utterance (the sum of clean speech and noise, $y = x + n$) as input. This signal is segmented and the autocorrelation of each frame is obtained to produce a tonegram. High energy regions of the tonegram are identified and their maximum energetic trajectories, obtained by means of a DTW approach, result in many pitch lines. We select a set of Maximum Energy Lines (M.E.L.) and their octave factors regarding their fundamental lines are estimated by using CASA rules. We relocate these lines at its fundamental period position, and estimate the tonegram which should be observed if only M.E.L. were presented with the addition of the corresponding octaves. We estimate the mean pitch of the speaker by means of this tonegram estimate and the final pitch py is obtained by discarding and selecting those lines which must correspond to the target speaker.

3 Pitch Extractor

The most important blocks and functions of the proposed extractor are detailed below. Note that the parameters of the blocks were determined through preliminary experiments performed over a set of training sentences of Aurora-2 contaminated with noise.

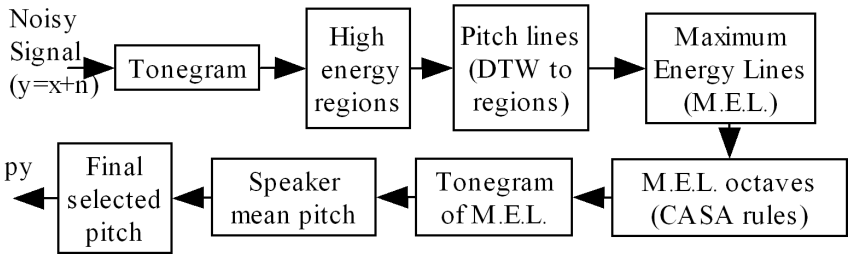


Fig. 1. Block diagram of the proposed pitch extractor

3.1 Tonegram

In order to estimate the tonegram, the unbiased autocorrelation is employed due to the following properties: fast computation (by means of Fast Fourier transform), concentration of noise at first coefficients (if it is not correlated [8]), capacity of representing the pitch energy, and capacity to define regions when a tone is presented. The power tonegram at pitch value p and frame time t is:

$$TG_{pow}(p,t) = \frac{1}{FL-p} \sum_{i=p}^{FL-1} y_t(i) y_t(i-p) \tag{1}$$

where y_t $(i = 0, .., FL-1)$ is the noisy signal in frame t (length $FL = 256$, sampling frequency $8kHz$). The frame shift is $FS = 80$ samples and $p \in [pl, ph]$, where $pl = 10$ and $ph = 160$ samples define the range of human pitch. The power tonegram is passed through a square root function and normalized to $[0, 1]$ in order to obtain the final tonegram ($TG(p,t)$), which is a more suitable representation of pitch magnitude energy. Fig. 2 shows a tonegram from an Aurora-2 utterance.

3.2 High Energy Regions

The mean and the standard deviation of each temporal frame of the tonegram ($\mathbf{TG}(t)$) increase when a tone is presented, so we can estimate the instantaneous energy of the tonegram as follows:

$$\mathbf{E}_{TG}(t) = \mu_{\mathbf{TG}(t)} + \sigma_{\mathbf{TG}(t)} \tag{2}$$

where $\mu_{\mathbf{TG}(t)}$ and $\sigma_{\mathbf{TG}(t)}$ denote the mean and the standard deviation of a tonegram vector at time t. The instantaneous background energy $\mathbf{Eb}_{TG}(t)$ is obtained by passing $\mathbf{E}_{TG}(t)$ through a smoothing mean filter of length $WL/5$ samples (diameter $2 * WL/5 + 1$) followed by a minimum filter of length $WL/2$ samples. WL is 30 frames and refers to the expected mean Word Length. A tonegram pixel is classified with a boolean high energy indicator if $TG(p,t) > \mathbf{Eb}_{TG}(t)$. The high energy regions consist of connected high energy pixels. Regions with an area lower than $2 * WL/5$ pixels are deleted. Fig. 2 shows the resulting high energy regions. In the following, the l^{th} region will be denoted as $TG^l(p,t)$.

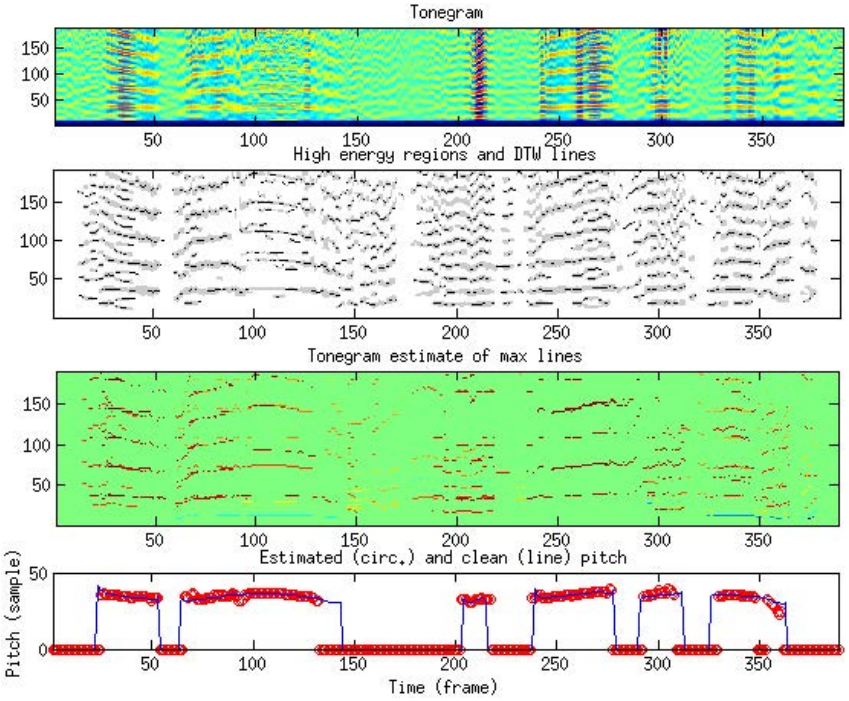


Fig. 2. Tonegram $TG(p,t)$ and its corresponding high energy regions with the DTW lines, tonegram estimate of Maximum Energy Lines (M.E.L) and pitch estimate for the FCJ_1396Z33A Aurora-2 utterance contaminated with babble noise at 0 dB

3.3 Estimation of Pitch Lines Based on DTW

Due to errors when estimating $\mathbf{Eb}_{TG}$, high energy regions can contain more than only one pitch line or even two or more crossed lines. An approach based on the maximum at each time in order to estimate the strongest energy line, can result in a discontinuous trajectory in these situations. Because of this, an approach based on searching for the path with maximum energy can be more suitable.

In other words, we can estimate a pitch trajectory through a region $TG^l(p,t)$ as the path that maximizes the global accumulated energy along axis t. For the sake of simplicity, in this section t will be a relative index ($t = 1,..,length^l$) where $length^l$ is the number of frames covered by the region. In order to find this path, we employ a method based on Dynamic Programming. The employed algorithm is quite similar to the well-known standard DTW technique [11], but introducing certain restrictions that we have found appropriate for pitch trajectory estimation.

Standard DTW is a pattern matching technique that has been used for decades in speech recognition, as well as in other areas, such as feature alignment in music [12]. Briefly, given a matrix $TG^l(p,t)$, DTW finds the warping path through the grid (p,t) that represents the "best" mapping between the two axes according to

$TG^l(p,t)$. This path is represented by a pair of warping vectors, $\mathbf{p}^l$ and $\mathbf{t}^l$, which give the coordinates of the path at every step i, that is, $\mathbf{p}^l = [p_1^l, p_2^l, \ldots, p_i^l, \ldots, p_I^l]$ and $\mathbf{t}^l = [t_1^l, t_2^l, \ldots, t_i^l, \ldots, t_I^l]$, where I is the number of steps in the path. In order to find the best path among all possible combinations, DTW minimizes the accumulated cost over the entire path. In our case, where $TG^l(p,t)$ represents energy (not cost), the optimal warping path can be defined as the one that maximizes the quantity $\sum_{i=1}^{I} TG^l(p_i^l, t_i^l)$, which measures the accumulated energy along the path.

In order to obtain a path that represents a meaningful pitch trajectory, some constraints must be imposed on the warping vectors. Firstly, the path must provide only a single pitch value for every frame, and secondly, the pitch trajectory must be smooth and continuous in frequency (and therefore, large hops in $\mathbf{p}^l$ should not be allowed). To satisfy both requirements, we impose the following local continuity constraints:

$$\mathbf{t}^l = [1, 2, \ldots, length^l] \tag{3}$$

$$\mathbf{p}^l = [p_1^l, \ldots, p_{length^l}^l] \quad \text{s.t.,} \quad |p_{i+1}^l - p_i^l| \leq h. \tag{4}$$

Clearly, the first constraint implies that each time frame will have only a single pitch, while the second one avoids pitch hops larger than h (in our experiments, we set $h = 3$ samples).

Taking into account these constraints, the DTW algorithm for finding the optimal trajectory through a region $TG^l(p,t)$ with size $P^l \times length^l$ can be summarized in two steps:

1. *Recursion*: For $1 \leq p \leq P^l$ and $2 \leq t \leq length^l$, compute

$$D(p,t) = \max_{p'}[D(p', t-1) + TG^l(p,t)], \tag{5}$$

whith initialization $D(p,1) = TG^l(p,1)$. Here, $D(p,t)$ can be interpreted as the maximum partial accumulated energy that can be obtained among all possible paths reaching the point (p,t). Observe that the maximization in (5) is performed only over the values p' from which (p,t) can be reached in a single step, in accordance with the constraint in (4). The best predecessor for each (p,t) is stored in ξ, i.e.,

$$\xi(p,t) = \arg\max_{p'}[D(p', t-1) + TG^l(p,t)]. \tag{6}$$

2. *Termination and Backtracking*: Finally, the optimal trajectory $\mathbf{p}^l$ is the path with higher global accumulated energy up to the end frame, yielding:

$$p_{length^l}^l = \arg\max_{p} D(p, length^l), \tag{7}$$

and the complete path is retrieved backwards as follows:

$$p_t^l = \xi(p_{t+1}^l, t+1), \quad \text{for } 1 \leq t \leq length^l - 1. \tag{8}$$

Fig. 2 shows the resulting DTW lines corresponding to each $TG^l(p,t)$ region.

3.4 Line Features

Once the pitch lines have been extracted we must store the following data vectors of length $length^l$ for every line associated to the region $TG^l(p,t)$: $\mathbf{t}^l$, $\mathbf{p}^l$, $\mathbf{E}^l$ vectors with the time, pitch positions and instantaneous energy. We will also note E^l_{mean} the mean line energy, and t^l_{max}, t^l_{min}, p^l_{max}, p^l_{min} the corresponding maxima and minima.

3.5 Selection of Maximum Energy Lines (M.E.L.)

A vector with the line labels, corresponding to maximum mean energy (E^l_{mean}) at each time, is obtained and passed through a mode filter of length $WL/10$. This filter avoids including lines which are maximum for a very short time and its length is related to the temporal masking effect. The different filtered labels indicate the M.E.L. set. In the case of an energy tie, the line with lower pitch is selected because we are looking for the lines corresponding to the fundamental period. This situation will be addressed in Sec. 3.8.

3.6 Octave Estimation

Any line corresponding to a fundamental pitch period should appear repeated at integer multiples, or horizontally in the tonegram. This can cause octave error when selecting M.E.L.s. The integer relation between the pitch of a maximum selected line lm and its fundamental line $lm0$ will be called the octave of lm ($o^{lm} = \mathbf{p}^{lm}/\mathbf{p}^{lm0}$) and is estimated by a grouping-line approach inspired on CASA [5] in these four steps:

1. Find horizontal lines close to lm: lines lh which fulfill this condition ($t^{lh}_{max} > t^{lm}_{min} \quad \& \quad t^{lh}_{min} < t^{lm}_{max}$) are selected.
2. Measure common movement, limit and intensity between lm and the horizontal lines lh as follows:

$$c^{lh}_{mov} = 1 - \frac{\sigma_{(\bar{\mathbf{p}}^{lm} - \bar{\mathbf{p}}^{lh}/f^{lh})}}{10} \tag{9}$$

$$c^{lh}_{lim} = 1 - \frac{|t^{lm}_{min} - t^{lh}_{min}| + |t^{lm}_{max} - t^{lh}_{max}|}{length^{lm}} \tag{10}$$

$$c^{lh}_{int} = 1 - \frac{|E^{lh} - E^{lm}|}{E^{lm}} \tag{11}$$

   where $\bar{\mathbf{p}}^{lm}$ and $\bar{\mathbf{p}}^{lh}$ indicate the common pitch part between lm and lh, and $f^{lh} = \mu_{(\bar{\mathbf{p}}^{lh}/\bar{\mathbf{p}}^{lm})}$ is the horizontal factor. Note that the maximum value for the common measures is always 1.
3. Select octave-related lines: the lines with common movement, limit and intensity bigger than $Th_o = (0.9, 0.9, 0.9)$ are the octave-related lines **lo** to lm. In case of not grouping lines, we try these other thresholds $Th_o = (0.7, 0.9, 0.9)$ and $Th_o = (0.9, 0.7, 0.9)$.

4. Estimate the octave of maximum line: If horizontal lines have not been selected, octave estimate is $\hat{o}^{lm} = 1$. If horizontal lines have been selected but not octave-related, $\hat{o}^{lm} = -1$. In other case, we estimate the octave considering that the f^{lh} of an octave-related line has to be an integer multiple of $1/o^{lm}$. For example, assuming $o^{lm} = 2$ the observed vector of octave lines should ideally be $\mathbf{f}^{lo} = 0.5, 1, 1.5, ...$ Taking this into account, the octave estimate is that which minimizes the distance between the observed and the ideal factor vector of an octave ($\hat{o}^{lm} = \arg\min_o(dist(\mathbf{f}^{lo}, \mathbf{f}_o^{ideal}))$). This distance is obtained by means of a clustering procedure and increases when the clustering error and the amount of not matched centroids (elements of $\mathbf{f}_o^{ideal}$) increases. The maximum possible tried octave is always $o_{max} = 6$.

3.7 Tonegram Estimation of M.E.L.

The tonegram of M.E.L. is estimated as follows: we fill an empty tonegram with the original M.E.L. of Sec. 3.5 but relocated to their correct new position using the octave estimate ($\mathbf{p}_{new}^{lm} = \mathbf{p}_{orig}^{lm}/\hat{o}^{lm}$) and with the same original instantaneous energy. Also, the corresponding octave lines are put at integer multiples of $\mathbf{p}_{new}^{lm}$ and with the same energy. The lines with $\hat{o}^{lm} = -1$ are not moved but some possible octave lines are put at integer multiples and divisions of p_{orig}^{lm} and with the corresponding energy of the original tonegram. We do so because the octave is unknown. The maximum integer number, for adding octaves, is always limited to o_{max} in order to avoid the inclusion of too many lines. The features of this new tonegram are extracted and loaded in a structure as in Sec. 3.4. Fig. 2 shows this tonegram estimate.

3.8 Mean Pitch Estimation of the Speaker

We select again the M.E.L. from the previous estimated tonegram in a similar way to Sec. 3.5 and a tonegram with these new M.E.L. is constructed. This tonegram will be denoted as TG_{perc} and can be considered as a representation of the perceived tones at each time if we are focusing our attention on maximum energy tones presented in the auditory scene. The total perceived energy of each tone ($E_{perc}(p)$) is obtained by summing neighboring channels separated one tone as follows:

$$\mathbf{E}_{perc}(p) = \sum_{t=1}^{nf} \sum_{\rho=[p*8/9]}^{[p*9/8]} TG_{perc}(\rho, t) \tag{12}$$

where nf is the number of frames and [] the round operator. Considering that, even at low SNRs, the majority of maximum tones correspond to the target speaker, we can say that the maximum of $\mathbf{E}_{perc}$ corresponds to the speaker mean pitch (p_{mean}).

3.9 Final Pitch Selection

If we suppose that the speaker pitch lines are concentrated around an interval of p_{mean} we can discard many lines from the M.E.L. tonegram of Sec. 3.7, so

the l lines which do not fulfill this condition ($p^l_{max} > (2/3)p_{mean}$ & $p^l_{min} < (3/2)p_{mean}$) are deleted. In a similar way to Sec. 3.5, we select the M.E.L. of this deleted-tonegram and the corresponding pitches at each time of the line with maximum total energy conform our previous pitch estimate.

The previous unvoiced frames are those where pitch has not been detected. In the case that unvoiced frames are not detected, we suppose unvoiced the first and last 10 frames. In a similar way to Sec. 3.2 we obtain $\mu_{\mathbf{E}^u_{TG}}$ and $\sigma_{\mathbf{E}^u_{TG}}$ (the mean and the standard deviation of the instantaneous energy $\mathbf{E}_{TG}$ of unvoiced frames) in order to obtain an unvoiced background threshold. The instantaneous energy of the voiced frames ($\mathbf{E}^v$) is smoothed with a mean filter of length $WL/10$ samples and the frames with $\mathbf{E}^v < \mu_{\mathbf{E}^u_{TG}} + 5 * \sigma_{\mathbf{E}^u_{TG}}$ are labeled as unvoiced. Finally, the value of the previous pitch is made null at unvoiced frames and this is our final pitch estimate py. Fig. 2 also compares this pitch extraction with the clean pitch (extracted from the corresponding clean utterance).

4 Experimental Framework and Results

4.1 Experimental Framework

The experiments reported here employ the Aurora-2 database which consists of digit utterances contaminated by different types of noises at different SNRs [9].

The evaluation of the pitch estimate will be done in a novel and useful way - by means of a pitch-based technique [6] for robust ASR. This technique has been presented in [7] and combines two complementary noises [a Voice Activity Detection noise (suitable for silence frames) and a tunnelling noise (suitable for voiced frames)] to estimate the noise spectrogram. This noise produces a soft Missing Data (MD) mask which is passed, together with the noisy spectrogram, to a marginalization MD recognizer. For the sake of simplicity, here, we will obtain a hard mask [3](instead of soft) which only requires the optimization of the threshold (and not also of the slope) to decide wether a feature is reliable or not. Clean train is always done and the HMM model features of the MD recognizer are the standards of Aurora-2 when the spectrogram is employed (9 Gauss/state, 23-LogMel-static+23-LogMel-delta feature vector, etc.. [7]).

4.2 Experimental Results

Tab. 1 shows the different word accuracies achieved by different systems tested over the whole (set A, B and C) Aurora-2 database.

FE+CMN is the ETSI Front End (FE) with Cepstral Mean Normalization and acts on a classical cepstral recognizer [9]. The rest of the systems act on the MD recognizer explained above with different pitch extractors. *PEFAC* employs the pitch extractor proposed in [4] but, in order to improve its results, we apply the following post-processing: frames with voiced probability lower than 0.8 are selected as unvoiced. This decision is later passed though a mode filter of length 1 frame. Finally, we make null the pitch at unvoiced frames. *Yin* uses the extractor

Table 1. Word accuracies obtained by different systems tested with Aurora-2 (set A, B and C) for different SNR values

Systems	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Mean (20-0 dB)
FE+CMN	99.12	97.17	92.53	76.15	44.16	23.02	13.00	66.61
PEFAC pitch	98.67	93.56	84.69	69.29	55.23	37.30	18.31	68.01
Yin pitch	98.89	94.93	89.32	80.07	66.47	39.56	14.36	74.07
DTW-lines pitch (proposal)	98.20	95.07	90.14	80.93	66.15	39.06	14.90	74.27

described in [2]. Frames with a normalized energy threshold lower than 0.8 and gross aperiodicity bigger than 0.95 are considered unvoiced. *DTW-Lines* employs the proposed pitch extractor. The optimum threshold of the masks was $-3dB$ in all cases, except for the *PEFAC* approach ($0dB$).

We can see that our pitch extractor outperforms all the extractors on average. In clean conditions, our pitch extractor does not obtain as good results as the others probably because the background energy thresholds of Sec. 3.2 and 3.9 avoid the detection of some weak regions and pitch values respectively.

5 Conclusions

This paper has proposed a pitch extractor for ASR based on the assumption that the most energetic pitch lines of the tonegram, around a speaker mean pitch estimate, correspond to the speaker pitch. The pitch lines have been extracted with a DTW approach and CASA rules have been employed to group and reject lines. The proposal has been evaluated on a robust ASR system showing high performance. Regarding future work, the results at clean and noisy conditions could be improved by means of a better estimation of the background energy threshold and a better application of CASA rules in the selection of the target speaker lines. Also we would like to test this scheme on another more robust tonegram (such as the difference function [2]), and on the two-talker recognition problem [5] by using the line features (such as the intensity, mean-pitch or even space-localization) together with high level information (provided by Speech Fragment Decoding [1,5]) in order to separate the pitch lines of the two speakers.

References

1. Barker, J., Cooke, M., Ellis, D.: Decoding speech in the presence of other sources. Speech Communication 45, 5–25 (2005)
2. De Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America 111(4), 1917–1930 (2002)
3. Cooke, M., Green, P., Josifovski, L., Vizinho, A.: Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34, 267–285 (2001)

4. Gonzalez, S., Brookes, M.: A pitch estimation filter robust to high levels of noise (pefac). In: EUSIPCO (2011)
5. Ma, N., Green, P., Barker, J., Coy, A.: Exploiting correlogram structure for robust speech recognition with multiple speech sources. Speech Communication 49, 874–891 (2007)
6. Morales-Cordovilla, J.A.: Pitch-based technique for robust speech recognition. PhD thesis, Dept. of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Spain (2011)
7. Morales-Cordovilla, J.A., Ma, N., Sánchez, V., Carmona, J.L., Peinado, A.M., Barker, J.: A pitch based noise estimation technique for robust speech recognition with missing data. In: ICASSP, May 22-27, pp. 4808–4811 (2011)
8. Morales-Cordovilla, J.A., Peinado, A.M., Sánchez, V., Gonzalez, J.A.: Feature extraction based on pitch-synchronous averaging for robust speech recognition. IEEE Trans. on Audio, Speech and Lang. Proc. 19(3), 640–651 (2011)
9. Pearce, D., Hirsch, H.G.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ICSLP, vol. 4, pp. 29–32 (2000)
10. Peinado, A.M., Segura, J.C.: Speech Recognition over Digital Channels. Wiley (2006)
11. Rabiner, L., Juang, B.-H.: Fundamentals of speech recognition. Prentice-Hall (1993)
12. Turetsky, R.J., Ellis, D.P.: Ground-truth transcriptions of real music from force-aligned midi syntheses. In: Int. Conf. Music Inf. Retrieval (ISMIR), pp. 135–141 (2003)

Speech Denoising Using Non-negative Matrix Factorization with Kullback-Leibler Divergence and Sparseness Constraints

Jimmy Ludeña-Choez and Ascensión Gallardo-Antolín

Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 - Leganés (Madrid), Spain
{jimmy,gallardo}@tsc.uc3m.es

Abstract. A speech denoising method based on Non-Negative Matrix Factorization (NMF) is presented in this paper. With respect to previous related works, this paper makes two contributions. First, our method does not assume a priori knowledge about the nature of the noise. Second, it combines the use of the Kullback-Leibler divergence with sparseness constraints on the activation matrix, improving the performance of similar techniques that minimize the Euclidean distance and/or do not consider any sparsification. We evaluate the proposed method for both, speech enhancement and automatic speech recognitions tasks, and compare it to conventional spectral subtraction, showing improvements in speech quality and recognition accuracy, respectively, for different noisy conditions.

Keywords: Non-Negative Matrix Factorization, Kullback-Leibler Divergence, Sparseness Constraints, Speech Denoising, Speech Enhancement, Automatic Speech Recognition.

1 Introduction

The quality of speech is degraded in the presence of noise. Noisy speech signals are a common problem in many applications, e.g. Automatic Speech Recognition (ASR), landline and mobile phone communications, etc. In ASR systems, the problem is harder because machine understanding is still far from humans and speech enhancement is sometimes performed as a preprocessing stage for those systems. In this paper, we have concentrated our efforts on enhancing speech for both, human consumption and ASR. Several methods for reducing the influence of noise have been proposed. Among them, it is worth mentioning the Wiener filtering technique [1] and Spectral Subtraction (SS) [2], which consists of subtracting an estimate of the noise spectrum from the noisy speech spectrum. Both of them produce a more intelligible signal but generate the so called musical noise as a side effect.

Recently, Non-Negative Matrix Factorization (NMF) has been successfully used in areas related to speech processing, including speech denoising [3], sound separation [4], speaker separation [5] and feature extraction [6]. NMF provides a

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 207–216, 2012.

way of decomposing a signal into a convex combination of nonnegative building blocks (also called basis vectors) by minimizing a cost function. Typical cost functions are the Euclidean distance and the Kullback-Leibler (KL) divergence. Therefore, NMF is capable of separating sound sources when their corresponding building blocks are sufficiently distinct, as is the case of speech and noise.

In this paper, we propose a NMF-based method for speech denoising which is very close to the one developed in [3] for speech enhancement tasks. The technique in [3] is based on a prior model of speech and noise, and therefore it assumes a priori knowledge of the type of noise which contaminates speech. In contrast, our method does not use this explicit information about noise, because it works with the only-noise segments of the current utterance to be denoised, after being detected with a Voice Activity Detector (VAD). Besides, we report results for both, speech enhancement and automatic speech recognition. On the other hand, several studies point out that it may be useful to explicit control the degree of sparsity in NMF decompositions for sound and speaker separation tasks. In this sense, the method for speaker separation proposed in [5] introduces a penalty term in the NMF with Euclidean distance that allows to control the sparsity of the solution. However, recent NMF-based techniques in speech processing report better results by using NMF with KL divergence [6], [4]. For this reason, in this paper, we propose a NMF-based method for speech denoising which combines the use of the KL divergence with sparseness constraints following the procedure described in [7].

This paper is organized as follows: Section 2 introduces the mathematical background of NMF; in Section 3 we present the speech denoising process using NMF. In Sections 4 and 5 we describe the application of the method to speech enhancement and automatic speech recognition, respectively, and end with some conclusions in Section 6.

2 Non-negative Matrix Factorization (NMF)

Given a matrix $V \in \mathbb{R}_+^{F \times T}$, where each column corresponds to a data vector, non-negative matrix factorization (NMF) approximates it as a product of two matrices of nonnegative low rank W and H, such that

$$V \approx WH \tag{1}$$

where $W \in \mathbb{R}_+^{F \times K}$ and $H \in \mathbb{R}_+^{K \times T}$ and normally $K \leq min\,(F, T)$. This way, each column of V can be written as a linear combination of the K basis vectors (columns of W), weighted with the coefficients of activation or gain located in the corresponding row of H. NMF can be seen as a dimensionality reduction of data vectors from an $F-$dimensional space to the $K-$dimensional space. This is possible if the columns of W uncover the latent structure in the data [8]. The factorization is achieved by an iterative minimization of a given cost function as, for example, the Euclidean distance or the generalized Kullbak Leibler (KL) divergence,

$$D_{\mathrm{KL}}\left(V\|WH\right)=\sum_{ij}\left(V_{ij}log\frac{V_{ij}}{\left(WH\right)_{ij}}-\left(V-WH\right)_{ij}\right) \tag{2}$$

In this work, we consider the KL divergence because it has been recently used with good results in speech processing tasks, such as sound source separation [4], speech enhancement [3] or feature extraction [6]. In order to find a local optimum value for the KL divergence between V and (WH), an iterative scheme with multiplicative update rules can be used as proposed in [8] and stated in (3)

$$W \leftarrow W \otimes \frac{\frac{V}{WH}H^T}{1H^T} \qquad H \leftarrow H \otimes \frac{W^T\frac{V}{WH}}{W^T1} \tag{3}$$

where 1 is a matrix of size V, whose elements are all ones and the multiplications $\otimes$ and divisions are component wise operations.

The NMF algorithm does not assume any sparsity or mutual statistical independence between columns of W. However, NMF usually provides sparse decomposition [8]. There are several ways to achieve some control of the sparsity. In this paper, we follow the approach proposed in [7] and [9] for KL cost functions, in which the NMF is regularized using non-linear projections based on (3). Applying this procedure, the regularized learning rules are the following,

$$W \leftarrow \left[W \otimes \frac{\left[\frac{V}{WH}H^T\right]^{\omega}}{1H^T}\right]^{(1+\alpha_w)} \qquad H \leftarrow \left[H \otimes \frac{\left[W^T\frac{V}{WH}\right]^{\omega}}{W^T1}\right]^{(1+\alpha_h)} \tag{4}$$

where α_w and α_h are the regularization parameters or sparse factors and ω is a relaxation parameter which also controls the sparsity and, in addition, speeds up the algorithm convergence. Note that with the sparse factors, the exponent of the learning rules are greater than one, which implies that the small values in the non-negative matrix tend to zero as the number of iterations increase [9]. In this paper, we only consider sparsification on the matrix H.

3 Speech Denoising Using NMF

NMF-based methods perform speech denoising under the hypothesis that noisy speech signals are the additive mixture of two sufficiently different sources: speech and noise. NMF is applied to magnitude spectra as it is assumed that the short-time magnitude spectra of a noisy signal, $|V_{\mathrm{mix}}|$ can be expressed as a linear combination of several distinct components, those representing only-speech spectra (W_{speech}) and those representing only-noise spectra (W_{noise}). These components are called Spectral Basis Vectors (SBV). The NMF representation of a noisy signal is shown in Fig. 1, wherein the speech SBVs (W_{speech}) and their corresponding speech activation coefficients (H_{speech}) can be used to reconstruct the clean speech signal ($|V_{\mathrm{speech}}| \approx W_{\mathrm{speech}}H_{\mathrm{speech}}$), while the noise SBVs (W_{noise}) and their corresponding noise activation coefficients (H_{noise}) can also be used to reconstruct the noise signal ($|V_{\mathrm{noise}}| \approx W_{\mathrm{noise}}H_{\mathrm{noise}}$) if required.

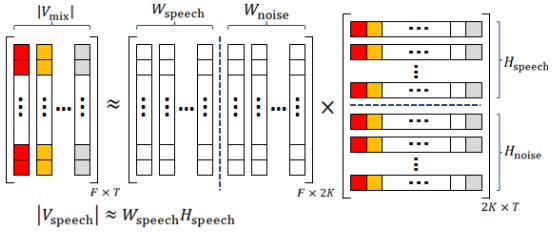


Fig. 1. NMF representation of noisy speech signals

The speech enhancement process consists of two different stages, training and denoising itself, as detailed below.

Training Stage. In the training stage, the SBVs representing speech and noise signals are determined. This is done by separately performing NMF on clean speech and noise data. First, the spectrum magnitude of both, clean speech ($|V_{\text{speech}}|$) and noise ($|V_{\text{noise}}|$) is computed. Afterwards, the KL divergence between the magnitude spectra and their corresponding factored matrices (($W_{\text{speech}}H_{\text{speech}}$) and ($W_{\text{noise}}H_{\text{noise}}$)) is minimized using the learning rules in (3). Since it is an iterative algorithm, it is important to perform a proper initialization of the matrices. Note that the spectral basis vectors contained in W_{speech} and W_{noise} are used in the next stage as speech and noise models.

For building the speech model, it is assumed that enough clean speech data is available. For the noise model, we have explored two different alternatives:

- *Offline Noise Data (OND).* In this approach, a priori knowledge about the type of the noise is assumed as in [3]. Therefore, a separate noise model for each of the noise types considered is trained using some offline available noise data. This approach will provide an upper limit of the proposed NMF-based denoising method performance.
- *Voice Activity Detector Noise Data (VADND).* In this approach, a VAD is used in order to explicit detect the only-noise segments of the utterance to be denoised. Afterwards, the noise model is built using these noise frames. Therefore, one noise model is trained for each utterance to be enhanced. This approach is more computational costly, but it avoids the need of the a priori knowledge about the type of noise, which it is not always possible.

Denoising Stage. As W_{speech} and W_{noise} are assumed to be good spectral basis functions to describe speech and noise, in the denoising stage they are kept fixed and are concatenated to form a single set of SBVs called W_{all}. Given the magnitude spectrum of the noisy speech signal ($|V_{\text{mix}}|$), we compute its factorization $|V_{\text{mix}}| \approx W_{\text{all}}H_{\text{all}}$ by minimizing the KL divergence between $|V_{\text{mix}}|$ and ($W_{\text{all}}H_{\text{all}}$), updating only the activation matrix H_{all} with the learning rules in (4). In order to control the sparseness of H_{all}, appropriate values for the regularization parameters (ω and α_h) need to be chosen (see subsection 4.2).

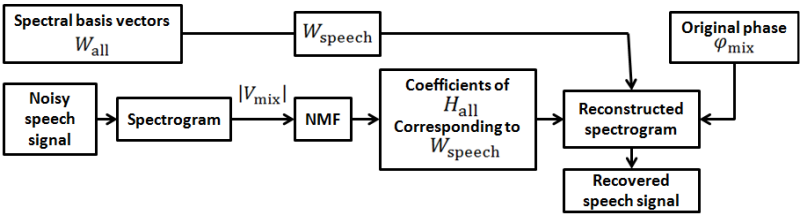


Fig. 2. Block diagram of the speech denoising process using NMF

The magnitude spectrum of the denoised speech is estimated as $|V_{\text{speech}}| \approx W_{\text{speech}}H_{\text{speech}}$, being H_{speech} the rows of H_{all} corresponding to the activation coefficients of W_{speech}. Finally, the spectrogram is recovered using the phase spectrum of the original noisy signal and the denoised speech signal is transformed into the time domain by means of a conventional overlap-add method. The whole process of speech denoising is shown in the block diagram of Fig. 2.

4 Application to Speech Enhancement

In this section, the evaluation of the proposed methods (OND and VADND) on a speech enhancement task is presented.

4.1 Database and Experimental Setup

The evaluation of speech enhancement was conducted on the AURORA-2 database [10], which is based on the TIDIgits database and it contains the recordings of 52 male and 52 females US-American adults pronouncing sequences of digits. Originally the database was recorded in clean conditions and subsequently contaminated with several types of noises at different SNRs. The sampling frequency is 8KHz. The database was end-pointed using the G.729 VAD.

For training the speech SBVs we used around 420 clean files belonging to the clean training set of the AURORA-2 database. In the OND method, the specific noise models were trained using the corresponding noise signals included in the database. In the VADND approach, the noise model for each utterance was trained using the initial only-noise frames detected by the VAD. In order to perform the study in subsection 4.2 we used 1,200 files from the test set A, which correspond to different noisy versions of 200 arbitrarily selected files with car noise added at SNRs from −5dB to 20dB with 5dB step. Finally, experiments in subsection 4.3 were conducted over 4,800 files from the test set A containing speech contaminated with subway, babble, car and exhibition hall noises at the SNRs previously mentioned. These files are noisy versions of 200 arbitrarily selected speech signals different from the ones used in subsection 4.2.

To evaluate the performance of the proposed methods, we use the so-called *Perceptual Evaluation of Speech Quality (PESQ)*, which is a measure recommended by the ITU-T for end-to-end speech quality assessment. The PESQ

score is able to predict subjective quality with good correlation in a very wide range of conditions (noise, filtering, coding distortions, etc.) [11] and uses a 5-point scale with 1 the worst and 5 the best values. PESQ values were computed using the code available in [12] and considering the clean speech signal as the reference. Results are presented using the following relative measure,

$$Ef_{\text{rel}} = \frac{PESQ_{\text{denoised}} - PESQ_{\text{noisy}}}{PESQ_{\text{noisy}}} \times 100\% \tag{5}$$

where $PESQ_{\text{noisy}}$ and $PESQ_{\text{denoised}}$ are the PESQ scores before and after applying the speech enhancement process, respectively. Increments imply a quality improvement and decrements a degradation with respect to the noisy signal.

4.2 Study on the Influence of the NMF Parameters

This set of experiments was performed in order to study the impact of several NMF parameters on the quality of the enhanced speech. The considered parameters were the analysis window length and the frame shift used for spectrograms computation, the number of spectral basis vectors and the values of the regularization factors, ω and α_h. In all cases, NMF was initialized by running 10 times the Alternating Least Squares NMF (ALS NMF) algorithm [9], in such a way that the factorization with the smallest euclidean distance between V and (WH) was chosen for initialization. Then, these initial matrices were refined by minimizing the KL divergence with sparseness constraints as indicated en Section 2. Preliminary experiments considering the Euclidean distance as cost function instead of the KL divergence produced worse results in terms of PESQ. The main experiments and results are summarized in next paragraphs:

- The window length was varied from 10ms to 45ms with 5ms step. From this set of experiments, it was observed that PESQ scores decreased with the window length, obtaining the best results in the range from 25ms to 45ms.
- The frameshift was studied in the range from 1ms to 10ms. In this case, the speech quality improved as the frameshift became smaller. Best PESQ scores were found in the range from 1ms to 5ms.
- The number of SBVs was varied from 10 to 80 with 10 step. Results showed that the quality of the denoised speech degraded when using a small number of SBVs (below 30), whereas best PESQ scores were obtained in the range from 40 to 80 SBVs. This result indicates that for an adequate representation of speech signals in NMF, it seems necessary to consider more than 30 SBVs.
- With respect to the regularization parameters, several experiments were performed varying α_h from 0 to 1.2 (fixing $\omega = 1$) and varying ω from 1 to 2.5 (fixing $\alpha_h = 0$). Results for the OND approach are shown in Fig. 3a and Fig. 3b, respectively. Similar trends were observed for the VADND method. As it can be observed, PESQ scores degrade when no regularization is used (this case corresponds to $\alpha_h = 0$ in Fig. 3a and $\omega = 1$ in Fig. 3b). However, when the values of the regularization parameters increase, the speech quality improves, being the best performance found for the combination of

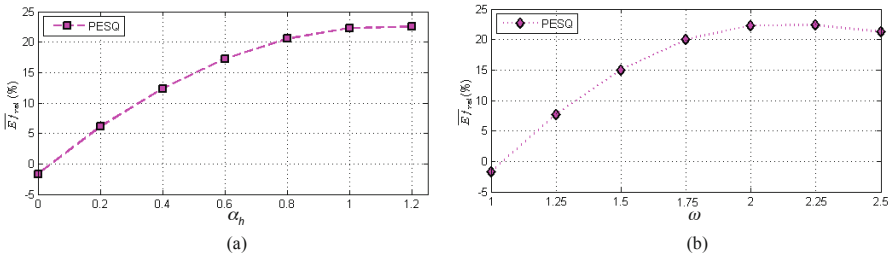


Fig. 3. Relative PESQ measure for the OND approach and a) $\omega = 1$ with different values of α_h and b) $\alpha = 0$ with different values of ω

α_h around 1 and $\omega = 1$ or the combination of ω around 2 and $\alpha_h = 0$). Other combinations of these parameters were tried, not obtaining significant improvements with respect to these PESQ values.

4.3 Experimental Results

In this subsection, we compare the performance of the two NMF-based denoising approaches (OND and VADND) with the conventional Spectral Subtraction (SS) in terms of the relative PESQ measure. According to the results achieved in the previous study, for the NMF-based methods, we used a window length of 40ms, a frameshift of 2.5ms, 50 SBVs, $\omega = 1$ and $\alpha_h = 1$. For a fair comparison, in SS we considered the same values for the window length and the frameshift.

Fig. 4 shows the relative PESQ measure with respect to the noisy signal for the four types of noise considered at several SNRs. For subway, babble and exhibition hall noises, the NMF-based methods clearly outperform SS at low and medium SNRs (from -5 dB to 10 dB). For SNR = 15 dB, results obtained with OND, VADND and SS are rather similar. However, at higher SNR (20 dB), SS produces better results than the NMF-based techniques. For the car noise, OND is better than SS at low and medium SNRs (-5 dB, 0 dB and 5 dB). Nevertheless, SS outperforms OND for higher SNRs. VADND produces worse results than SS at all SNRs, being more noticeable the differences for SNRs over 15 dB. In general, results show that OND and VADND are more suitable than SS for low and medium ranges of SNR.

With respect to the comparison between OND and VADND, it can be observed that the quality of the enhanced signal is better with OND in all cases. This result is expectable because OND uses more information than VADND in the denoising process. In fact, it needs to know the type of noise (not the SNR) presented in the noisy utterances. Nevertheless, VADND is capable of effectively denoise the speech signal using only the information contained in the only-noise segments of each utterance.

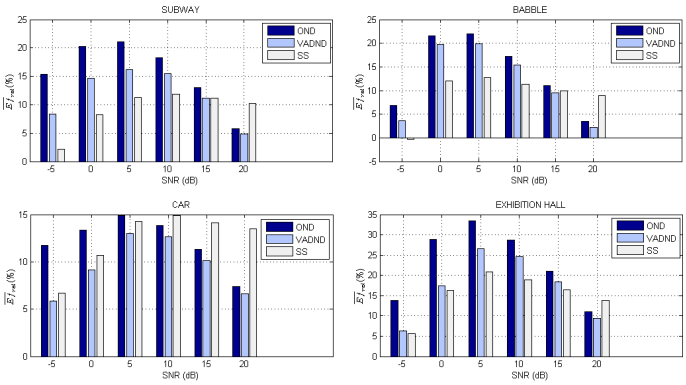


Fig. 4. Relative PESQ measure for SS, OND and VADND techniques

5 Application to Automatic Speech Recognition (ASR)

In this section, we present the evaluation of the proposed techniques on an ASR task. In this case, firstly noisy signals are denoised using the NMF-based techniques (OND or VADND) and then, these enhanced signals are fed into the ASR system.

5.1 Database and Experimental Setup

The experiments were conducted over the AURORA-2 database [10] as for the speech enhancement task. The recognizer was based on HTK (Hidden Markov Model Toolkit) software package with the configuration included in the standard experimental protocol of the database. Acoustic models were obtained from the clean training set of the database, whereas test files correspond to the complete test set A. Results are shown in terms of the recognition accuracy.

Acoustic features consist of the conventional Mel-Frequency Cepstrum Coefficients (MFCC). In particular, twelve MFCCs were computed every 10 ms using a Hamming analysis window of 25 ms long and 23 mel-spaced spectral bands. The log-energy of each frame and the corresponding delta and acceleration coefficients were also computed and added, yielding feature vectors of 39 components. Finally, cepstral mean normalization (CMN) was applied.

5.2 Experimental Results

Fig. 5 shows the recognition results achieved by the different NMF-based denoising techniques as well as for Spectral Subtraction (SS) and the baseline system (without denoising). For SS, OND and VADND, the same configuration parameters as in the case of speech enhancement were used, except for the regularization parameters, that were set to $\omega = 1.25$ and $\alpha_h = 0.2$, after a preliminary experimentation.

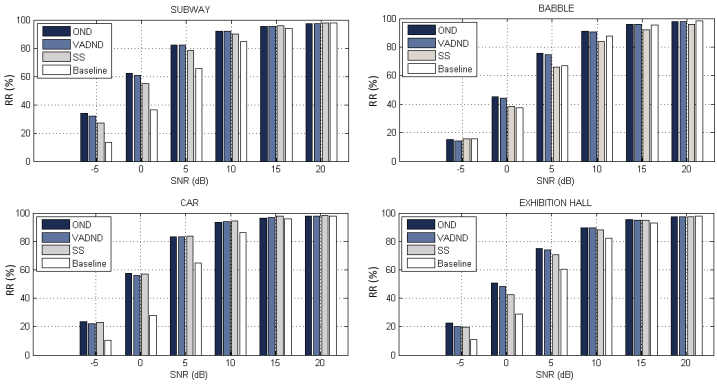


Fig. 5. Recognition Rates (%) for the baseline, SS, OND and VADND techniques

As it can be observed, for subway, babble and exhibition hall noises, both NMF-based techniques achieve better results than SS and the baseline for low and medium SNRs (from -5 dB to 10 dB). For higher SNRs, all the algorithms present a similar behaviour except for SS in the babble noise. In this case, the recognition accuracy obtained with SS is lower than the other techniques (including the baseline), probably due to the distortions introduced by SS in the denoising process. For the car noise, similar results are achieved with all techniques. On the other hand, comparing the two NMF-based methods for all noises, OND outperforms slightly VADND in most cases, being these performance differences less noticeably than in the speech enhancement task.

Table 1. Average Recognition Rates (%) for the four types of noise

Noise	OND	VADND	SS	Baseline
Subway	77.12	76.62	73.95	65.34
Babble	70.19	69.66	65.35	66.83
Car	75.29	74.94	75.72	63.86
Exhibition Hall	71.81	70.66	68.83	62.23

Table 1 shows the recognition rates averaged over all SNRs for each type of noise. It can be observed that OND and VADND outperforms SS for all noises, except for the car noise in which the results are very similar.

6 Conclusions and Future Work

In this paper we have presented a NMF-based method for speech denoising which combines the use of the Kullback-Leibler divergence with sparseness constraints on the activation matrix and it does not assume a priori knowledge about the

nature of the noise. In addition, an exhaustive study on the influence of different NMF parameters (window length, frameshift, number of spectral basis vectors and regularization parameters) on the quality of the enhanced speech has been carried out. We have compared the proposed method to conventional spectral subtraction for both, speech enhancement and automatic speech recognitions tasks, under different noisy conditions, obtaining significant improvements especially at low and medium SNRs.

For future work, we plan to experiment on real noisy signals instead of the artificially distorted ones used in this paper. Other future lines include the unsupervised learning of auditory filter banks by means of NMF and the use of the activation coefficients as acoustic parameters in ASR tasks.

Acknowledgments. This work has been partially supported by the Spanish Government grants TSI-020110-2009-103 and TEC2011-26807. Financial support from the Fundación Carolina (Jimmy Ludeña-Choez) is thankfully acknowledged.

References

1. Scalart, P., Filho, J.: Speech enhancement based on a priori signal to noise estimation. In: ICASSP 1996, pp. 629–632 (1996)
2. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: ICASSP 1979, pp. 208–211 (1979)
3. Wilson, K., Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. In: ICASSP 2008, pp. 4029–4032 (2008)
4. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. on Audio, Speech and Language Processing 15(3), 1066–1074 (2007)
5. Schmidt, M., Olsson, R.: Single-channel speech separation using sparse nonnegative matrix factorization. In: INTERSPEECH 2006 (2006)
6. Schuller, B., Weninger, F., Wollmer, M., Sun, Y., Rigoll, G.: Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In: ICASSP 2010, pp. 4562–4565 (2010)
7. Cichocki, A., Zdunek, R., Amari, S.: New algorithms for non-negative matrix factorization in applications to blind source separation. In: ICASSP 2006, pp. 621–625 (2006)
8. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
9. Cichocki, A., Zdunek, R., Phan, A., Amari, S.: Nonnegative matrix and tensor factorizations. John Wiley and Sons, United Kingdom (2009)
10. Pearce, D., Hans, G.: The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In: ICSLP 2000 (2000)
11. Beerends, J., Hekstra, A., Rix, A., Hollier, M.: Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II. Psychoacoustic model. Journal of the Audio Engineering Society 50(10), 765–778 (2002)
12. Hu, Y., Loizou, P.: Matlab software (2011), http://www.utdallas.edu/~loizou/speech/software.htm

MMSE Feature Reconstruction Based on an Occlusion Model for Robust ASR

José A. González*, Antonio M. Peinado, and Ángel M. Gómez

Dpt. Teoría de la Señal, Telemática y Comunicaciones,
Centro de Investigación en Tecnologías de la Información y de las Comunicaciones,
18071-Granada, Spain
{joseangl,amp,amgg}@ugr.es http://tstc.ugr.es,http://citic.ugr.es

Abstract. This paper proposes a novel compensation technique developed in the log-spectral domain. Our proposal consists in a minimum mean square error (MMSE) estimator derived from an occlusion model [1]. According to this model, the effect of noise over speech is simplified to a binary masking, so that the noise is completely masked by the speech when the speech power dominates and the other way round when the noise is dominant. As for many MMSE-based techniques, a statistical model of clean speech is required. A Gaussian mixture model is employed here. The resulting technique has clear similarities with missing-data imputation techniques although, unlike these ones, an explicit model of noise is employed by our proposal. The experimental results show the superiority of our MMSE estimator with respect to missing-data imputation with both binary and soft masks.

Keywords: robust ASR, feature reconstruction, MMSE estimation, occlusion model.

1 Introduction

Automatic speech recognition (ASR) is currently moving toward new ubiquitous and pervasive applications where it allows an efficient and natural way for human-machine interaction. However, these scenarios may reduce the performance of ASR systems due to several reasons. Undoubtedly, an adverse acoustic environment and, in particular, environmental noise, is the main of these reasons. Thus, the robustness of ASR systems against noise is a desirable feature that must be addressed.

In order to reduce the effect of the acoustic noise over speech recognizers there exist multiple approaches, but two of them stand out from others [2]: feature compensation and model adaptation. While the first one tries to *denoise* the speech features employed for recognition, the second one modifies the acoustic model parameters to reduce the mismatch with the noisy input features. The advantages of feature compensation is that it can be developed independently

* This work has been supported by the FPU fellowship program from the Spanish Ministry of Education and project MICINN TEC2010-18009.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 217–226, 2012.

from the recognition engine and, also, that it can be implemented more efficiently than adaptation.

This paper proposes a novel compensation technique developed in the log-spectral domain. Our proposal consists in a minimum mean square error (MMSE) estimator derived from an occlusion model [1]. According to this model, the effect of noise over speech is simplified to a binary masking, so that the noise is completely masked by the speech when the speech power dominates and the other way round when the noise is dominant. In order to model the clean speech log-spectra, we follow the classical approach based on a Gaussian mixture model (GMM). Section 2 is devoted to present and derive the proposed estimator. We will see that the application of MMSE along with the hard-decision occlusion model will yield a graceful soft-decision estimate which is a linear combination of the observed (noisy) feature vector and an estimate of the clean feature vector for the case of speech totally occluded by noise. The resulting estimator will resemble other techniques derived from a missing-data (MD) framework. The similarities and differences with these techniques are discussed is section 3. Section 4 is devoted to the experimental results. A summary of this work can be found in section 5.

2 MMSE Estimation from an Occlusion Model

2.1 Occlusion Model

We will note as $\boldsymbol{y}$ the feature vectors corresponding to the observed (noisy) log-Mel filterbank energies. Also, $\boldsymbol{x}$ and $\boldsymbol{n}$ will represent the same type of spectral features for the clean speech and the noise, respectively. The relationship between these variables is accurately represented by the following model [3],

$$\boldsymbol{y} = \boldsymbol{x} + \log\left(\mathbf{1} + e^{\boldsymbol{n}-\boldsymbol{x}}\right) + \boldsymbol{r} \tag{1}$$

where $\boldsymbol{r}$ is a residual vector that depends on the phase relationship between clean speech and noise.

Although accurate, the above distortion model does not allow an easy derivation of the MMSE estimator that we want to obtain, so some approximations must be introduced. In the case of the occlusion model (OM), the residual $\boldsymbol{r}$ is neglected and, also, the *log-max* approximation (that is, $\log(e^a + e^b) \approx \max(a, b)$) is applied to (1) (see [4] for a detailed derivation of this model). The resulting model can be finally expressed as,

$$\boldsymbol{y} \approx \max(\boldsymbol{x}, \boldsymbol{n}) \tag{2}$$

where operator **max** is applied feature by feature.

This model was first proposed in [1] and involves that some parts of the clean speech spectrogram are completely masked by noise, while others are almost unaffected (noise is masked by speech). Our proposal uses this assumption and the spectral correlations represented by the GMM to provide clean speech feature estimates.

2.2 MMSE Estimation Based on the Occlusion Model

MMSE estimation is a Bayesian tool frequently employed in feature compensation techniques. The MMSE estimate of the clean feature vector given the observed (noisy) one can be expressed as,

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\boldsymbol{y}] = \int_{-\infty}^{\infty} \boldsymbol{x} p\left(\boldsymbol{x}|\boldsymbol{y}\right) d\boldsymbol{x} \tag{3}$$

In first place, the MMSE estimator requires a clean feature model λ_X which allows the computation of the posterior needed in (3). This is usually carried out through a mixture of pdf's defined by,

$$p(\boldsymbol{x}|\lambda_X) = \sum_{k=1}^{M} P(k|\lambda_X) p(\boldsymbol{x}|k, \lambda_X) \tag{4}$$

The typical choice is a GMM where the pdf's $p(\boldsymbol{x}|k, \lambda_X) = p_X(\boldsymbol{x}|k)$ are Gaussians $\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ with means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ $(k = 1, \ldots, M)$.

The proposed MMSE estimator will also require an statistical model λ_N of noise. We will do the common assumption that the noise statistics are available at every instant. These statistics must be obtained from a previous estimation applied to the observed utterance. We will consider a single Gaussian model (for every time instant),

$$p\left(\boldsymbol{n}|\lambda_N\right) = p_N(\boldsymbol{n}) = \mathcal{N}\left(\boldsymbol{n}; \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right) \tag{5}$$

The posterior $p(\boldsymbol{x}|\boldsymbol{y}) \equiv p(\boldsymbol{x}|\boldsymbol{y}, \lambda_X, \lambda_N)$ required in equation (3) can be derived from (4) and (5),

$$p\left(\boldsymbol{x}|\boldsymbol{y}, \lambda_X, \lambda_N\right) = \sum_{k=1}^{M} p\left(\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N\right) P\left(k|\boldsymbol{y}, \lambda_X, \lambda_N\right) \tag{6}$$

so the MMSE estimate can be finally expressed as,

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{M} P\left(k|\boldsymbol{y}, \lambda_X, \lambda_N\right) \underbrace{\int \boldsymbol{x} p\left(\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N\right) d\boldsymbol{x}}_{E[\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N]} \tag{7}$$

As usual in MMSE feature compensation, the above estimate requires the computation of the posterior $P(k|\boldsymbol{y}_u, \lambda_X, \lambda_N)$ and the partial estimate $E[\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N]$ for every Gaussian component k. In both cases, we have to solve multivariate integrals. We will see next how the OM model can help us to do this. As previously mentioned, this model keeps the maximum between $\boldsymbol{x}$ and $\boldsymbol{n}$ feature by feature. Thus, in order to ease its application to our estimation problem, we will assume statistical independence among features. That is, all Gaussians in λ_X and λ_N will be diagonal and the required integrals can be correspondingly factorized. Statistical independence between speech and noise is also assumed.

For the sake of simplicity, models λ_X and λ_N will be removed from the notation. In the case that not both but only one model applies, this will be indicated with the corresponding subscript ($p_X(\cdot)$ or $p_N(\cdot)$).

Posterior Computation. In order to obtain the required posterior, we first apply the Bayes' rule,

$$P\left(k|\boldsymbol{y}\right) = \frac{p\left(\boldsymbol{y}|k\right) P_X(k)}{\sum_{k'=1}^{M} p\left(\boldsymbol{y}|k'\right) P_X(k')} \tag{8}$$

Therefore, our problem becomes that of computing $p(\boldsymbol{y}|k)$. If we now apply the statistical independence assumptions, this pdf can be factorized into the product of probabilities $p(y|k)$, where y represents a given observed feature.

Thus, we focus now on the computation of,

$$p\left(y|k\right) = \iint p\left(x, n, y|k\right) dxdn \tag{9}$$

$$= \iint p(y|x, n, k) p_X(x|k) p_N(n) dxdn \tag{10}$$

where x and n denote the corresponding clean speech and noise feature, respectively. In this equation, densities $p_X(x|k)$ and $p_N(n)$ are known, but $p(y|x, n, k)$ must be determined. Since the occlusion model forces y to be the maximum of x and n, it can be expressed as,

$$p(y|x, n, k) = p(y|x, n) = \delta\left(y - \max(x, n)\right) \tag{11}$$

where $\delta(\cdot)$ is the Dirac delta function.

On the other hand, the joint speech-noise space $\{(x, n)\}$ can be split into two subsets,

$$\begin{aligned} \mathcal{X} &= \{(x, n) | x \geq n\} \\ \mathcal{N} &= \{(x, n) | n > x\}, \end{aligned} \tag{12}$$

which yields the following expression for (10),

$$p\left(y|k\right) = \iint_{\mathcal{X}} \delta(y - x) p_X(x|k) p_N(n) dxdn + \iint_{\mathcal{N}} \delta(y - n) p_X(x|k) p_N(n) dxdn$$

Now, the integrations over variables x and n can be separated,

$$\begin{aligned} p\left(y|k\right) = &\int_{-\infty}^{\infty} p_X(x|k) \delta(y - x) dx \int_{-\infty}^{x} p_N(n) dn \\ &+ \int_{-\infty}^{\infty} p_N(n) \delta(y - n) dn \int_{-\infty}^{n} p_X(x|k) dx \end{aligned}$$

and it is finally obtained that

$$p\left(y|k\right) = p_X(y|k) C_N(y) + p_N(y) C_X(y|k) \tag{13}$$

where $C_X(y|k)$ and $C_N(y)$ are the cumulative density functions (cdf) corresponding to $p_X(y|k)$ and $p_N(y)$, respectively. Since, in our case, $p_X(y|k)$ and $p_N(y)$ are

Gaussians, these cdfs can be easily computed through the corresponding error functions as,

$$C_X(y|k) = \Phi\left(\frac{y-\mu_k}{\sigma_k}\right), \qquad C_N(y) = \Phi\left(\frac{y-\mu_N}{\sigma_N}\right). \tag{14}$$

It must be pointed out that the resulting posterior of eqn. (13) is the same as that proposed by Varga and Moore in [1] to perform speech recognition in noise. However, while Varga and Moore propose a 3-dimensional Viterbi algorithm to decode speech employing separate hidden Markov models (HMMs) for speech and noise, our proposal is oriented to feature compensation.

Partial Estimate Computation. Now, we must obtain the partial MMSE estimate $E[\boldsymbol{x}|\boldsymbol{y},k]$ (defined in (7)) applying the OM model. Considering the statistical independence assumptions, we must solve the following expectation,

$$E\left[x|y,k\right] = \int_{-\infty}^{\infty} xp\left(x|y,k\right)dx = \iint xp\left(x,n|y,k\right)dxdn \tag{15}$$

In this case, the pdf required for the integration is $p\left(x,n|y,k\right)$, which can be suitably developed by applying the Bayes' rule as,

$$p\left(x,n|y,k\right) = \frac{p\left(y|x,n\right)p_X\left(x|k\right)p_N\left(n\right)}{p\left(y|k\right)} \tag{16}$$

The integration can be now carried out in a similar way as performed for posterior $P(k|y)$, that is, considering $p(y|x,n) = \delta\left(y - \max(x,n)\right)$ and splitting the speech-noise space (and, therefore, the integral) into the same subsets $\mathcal{X}$ and $\mathcal{N}$ as those defined in (12). Thus, it is finally obtained that,

$$E\left[x|y,k\right] = w_k y + (1-w_k)\tilde{\mu}_k(y) \tag{17}$$

where

$$w_k = \frac{p_X(y|k)C_N(y)}{p(y|k)} \tag{18}$$

$$1 - w_k = \frac{p_N(y)C_X(y|k)}{p(y|k)} \tag{19}$$

$$\tilde{\mu}_k(y) = \int_{-\infty}^{y} x\frac{p_X(x|k)}{C_X(y|k)}dx = \mu_k - \sigma_k\frac{p_X(y|k)}{C_X(y|k)} \tag{20}$$

Discussion. The partial estimate of eqn. (17) is a linear combination of two feature estimates. The first one is the observed feature y, which can be interpreted as an estimate of the clean feature for high SNR values. The second one $\tilde{\mu}_k(y)$ can be interpreted as an estimate of the clean speech when it is completely masked by noise. In this case, we only know that the clean feature is somewhere

between $-\infty$ and y, so $\tilde{\mu}_k(y)$ is the mean value of Gaussian $p_X(x|k)$ truncated at y. Probability w_k acts as a weight which indicates how much y is affected by noise.

The final feature estimate can be expressed as,

$$\hat{x} = \left(\sum_{k=1}^{M} P(k|\boldsymbol{y}) w_k \right) y + \sum_{k=1}^{M} P(k|\boldsymbol{y})(1 - w_k)\tilde{\mu}_k(y) \tag{21}$$

which reflects again a linear combination of the observed feature and an estimate for the case of speech completely masked by noise. This former estimate is obtained as linear combination of the truncated Gaussian means.

3 Comparison with Related MD Techniques

The OM model has already been employed for feature compensation in previous works. This section is devoted to the comparison between these previous techniques and our proposal. In particular, we will focus on missing-data (MD) imputation techniques where the OM model is employed for spectral reconstruction [5,6,7].

The starting point of the MD techniques is a binary mask representing the reliability of the observed features. This mask has the same size as the input utterance spectrogram and each pixel m in it indicates whether the corresponding feature y is reliable ($m = 1$) or not ($m = 0$). Considering the OM model, $m = 1$ means $x \geq n$ while $m = 0$ means that the corresponding feature is completely occluded by noise, that is, $n > x$. Then, the conditional probability of eqn. (11) can be compactly written now as,

$$p\,(y|x, n) = m\delta(y - x) + (1 - m)\delta(y - n) \tag{22}$$

so that $p(y|k)$, which is required to compute $P(k|\boldsymbol{y})$ in (8), can be obtained as,

$$p\,(y|k) = m p_X(y|k) C_N(y) + (1 - m) p_N(y) C_X(y|k) \tag{23}$$

The expected value of eqn. (7) can be derived in a similar way,

$$E\,[x|y, k] = my + (1 - m)\tilde{\mu}_k(y) \tag{24}$$

The feature estimate can be finally obtained as,

$$\hat{x} = my + (1 - m) \sum_{k=1}^{M} P(k|\boldsymbol{y})\tilde{\mu}_k(y) \tag{25}$$

Let us compare now the MMSE estimator defined by eqns. (23)-(25) with that obtained in the previous section (eqns. (13), (17) and (21)). It is clear that the MD approach introduces a hard decision (reliable / not reliable) which is a consequence of the binary mask. On the other hand, our proposal avoids this and introduces a soft decision by considering the probability of feature occlusion. As

a result, it can be expected that the proposed technique will be more robust to errors in noise estimation since, in the case of MD, this may lead to erroneous masks and, therefore, to incorrect feature reliability classification.

The perjudicial effect that erroneous masks have over recognition performance is usually mitigated through soft masks [8,9] where $m \in [0, 1]$. This involves that $m \in [0, 1]$, that is, a continuous degree of reliability from 0 (fully unreliable) until 1 (fully reliable). Formulae (22)-(25) can be kept for this case and can be also found in [8].

We can observe that when soft masks are applied in the MD approach, the resulting estimator has clear similarities with the proposed one, although it must be noticed that our OM-based technique does not require any *a priori* knowledge about the feature reliability (that is, the mask). In particular, comparing the MD and OM estimators of equations (21) and (25), we could consider that our proposal provides a method to estimate the soft mask values as,

$$m = \sum_{k=1}^{M} P\left(k|\boldsymbol{y}\right) w_k \tag{26}$$

This last equation allows a direct comparison of (21) and (25). In both equations we have an estimate $\boldsymbol{y}$ (with weight m) for the case of speech-masking-noise which is linearly combined with an estimate for the case of noise-masking-speech. Although the terms on $\boldsymbol{y}$ are equivalent in both estimators, the comparison also reveals that the noise-masking-speech terms are clearly different.

Fig. 1 shows examples of log-Mel spectrograms for clean and noisy versions of the utterance *eight six zero one one six two* extracted from the Aurora-2 database [10]. Subway noise at 0dB was artificially added to the clean utterance in order to obtain the noisy one. Noise was estimated through linear interpolation of initial noise estimates obtained from the first and last frames of the utterance. It can be seen that the proposed technique effectively compensate for the noise degradation (enhanced speech plot) and also it is able to estimate feature reliability (estimated mask plot).

4 Experimental Results

In order to test our proposal and other reference techniques, we have employed the Aurora-2 [10] (connected digits) and Aurora-4 [11] (sentences from WSJ) databases and experimental frameworks. Aurora-2 has 3 test sets: A, B and C. Sets A and B consist of speech artifically contaminated by 4 different types of additive noise in each case (set A: the same noises as in training; set B: different from the training ones), and at 7 different SNRs (-5 to 20 dB, plus clean condition). Set C uses only two types of additive noise and also introduces channel distortion. Aurora-4 is a large vocabulary database with 14 test sets. The first seven sets (T-01 to T-07) artificially add six different noise types (T-01 is the clean condition) with SNR values between 5 dB and 15 dB. The last seven sets are obtained in the same way, but the utterances have been recorded with

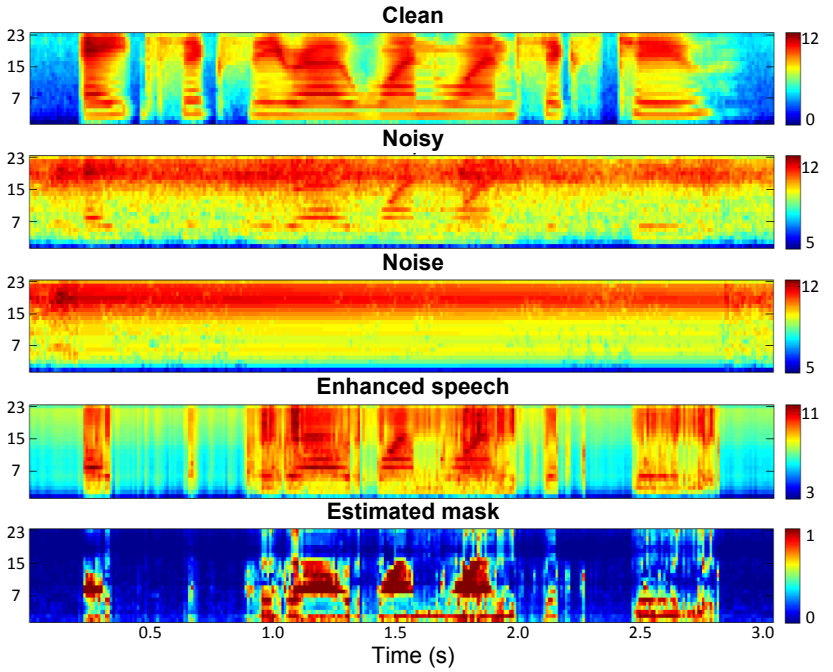


Fig. 1. Example of speech reconstruction and mask estimation (eqn. (26)) from the proposed OM- MMSE-based estimator

microphones different than those of training. For both databases, the acoustic models are trained with the usual scripts provided with the databases and using only clean speech.

The final feature vector employed for recognition consist of 13 Mel-frequency cepstral coefficients (MFCCs) ($C0$ is included instead of log-Energy) enlarged with Δ and $\Delta\Delta$ coefficients. Feature compensation is carried out over the 23 log-outputs of the Mel filterbank, which are DCT-transformed to obtain MFCCs. Also, cepstral mean normalization (CMN) is applied in order to mitigate channel distortions.

The clean spectral features were modeled with a 256-component GMM with diagonal covariance matrices, which has been trained through the expectation-maximization algorithm over the corresponding training set. The required noise estimates are obtained as follows: the first and last T frames ($T = 20$ for Aurora-2, $T = 35$ for Aurora-4) of every utterance have been averaged and the estimates for the intermediate frames are obtained through linear interpolation between the former ones. The noise model at every frame is completed with a covariance matrix fixed for all frames and computed from the first and last frames.

In order to assess our proposal in comparison with other techniques, the MD estimators described in the previous section has been also evaluated. Both, binary and soft masks have been considered. The binary masks are simply obtained

Table 1. Word accuracy results (%) for Aurora-2 at different SNRs

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Baseline	99.11	97.29	92.55	75.56	42.82	22.69	12.92	66.18
Oracle	99.11	99.01	98.74	97.84	95.72	89.64	73.79	96.19
BMD	98.88	97.45	95.32	90.01	78.47	54.99	25.55	83.25
SMD1	98.90	98.04	96.51	92.15	80.62	56.70	26.89	84.80
SMD2	98.91	97.91	96.32	91.74	79.77	55.30	26.20	84.21
SRO	98.91	98.08	96.69	92.77	82.18	58.76	27.21	85.70

by SNR thresholding at 0 dB. Soft masks are obtained by two different methods: a relaxation of the binary mask through a sigmoid function (its center and slope parameters has been optimized for a validation subset) and the mask defined by equation (26) and based on the proposed models (OM as well as clean speech and noise models). This last mask allows a direct comparison between MD imputation and MMSE estimation (both based on the occlusion model).

The word accuracy results for Aurora-2 are shown in table 1. The baseline corresponds to MFCCs with CMN. Three MD imputation techniques with three different types of masks are considered: masks obtained from the actual noise (Oracle), binary masks (BMD) and soft-masks (SMD1 and SMD2 for sigmoid-based and model-based masks, respectively). The oracle results can be considered an upper bound of the MD techniques (since the feature reliability is perfectly known). Our spectral reconstruction based on the OM model will be denoted as SRO. The results correspond to the average score over sets A, B and C for every SNR. Also, the average (Avg.) for SNRs from 0 to 20 dB is shown.

As it could be expected, the Oracle experiment achieves the best performance, but SRO provides the best results with estimated noise. Therefore, since the different techniques can be all considered variants of MD imputation which mainly differ in the way the mask values are computed (as explained in the previous section), we can say that SRO is more robust against noise estimation errors and, therefore, to mask errors. In this regard, the worst behavior corresponds to BMD. In this case, when a mask error ocurrs, an unreliable feature can be classified as reliable and the other way round. In the first case, the observed unreliable feature is kept. The second case is even worse, since it involves that reliable feature are treated as unreliable, being degraded by the estimation processing. In the case of MD with soft masks, the use of a mask looks *redundant* with the estimation based on a noise model as it is evident in equations (22) and (23), since the estimator can obtain its own mask (eqn. (26)). SMD1 yields results slightly better than SMD2 since SMD1 involves an optimization of the sigmoid parameters while the masks in SMD2 are completely extracted from statistical and OM models.

The results for Aurora-4 can be found in table 2. The proposed SRO technique outperforms again the MD imputation techniques with binary or soft masks, with relative improvements of 10.90 % and 1.77 %, respectively.

Table 2. Word accuracy results (%) for the different test sets of Aurora-4

	T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Avg.
Baseline	87.69	75.30	53.24	53.15	46.80	56.36	45.38	77.04	64.24	45.30	42.07	36.15	47.43	36.67	54.77
Oracle	87.69	86.74	84.46	84.44	83.19	85.90	82.38	79.13	77.86	74.03	73.45	70.48	75.04	71.77	79.75
BMD	86.96	80.78	58.47	52.74	59.63	56.14	61.42	79.39	74.13	54.83	46.76	50.55	51.26	56.17	62.09
SMD2	87.52	83.65	66.62	63.78	63.48	69.19	65.31	81.00	75.64	60.98	55.02	54.89	62.39	57.74	67.66
SRO	87.54	83.28	69.23	64.49	64.88	70.63	66.93	80.52	76.48	63.53	55.67	56.62	63.87	60.38	68.86

5 Conclusions

In this work we have proposed a technique for the MMSE estimation of log-spectral features corrupted by additive noise. The starting point is a simplification of a general noise distortion model through the *log-max* approximation, which yields the so-called occlusion model. This modeling involves that either the speech feature dominates the noise or, on the contrary, the speech is completely masked by noise. The resulting estimator has clear similarities with some MD imputation techniques. Indeed, it can be considered an MD technique or equivalently, a way of computing soft masks for MD imputation. Our experimental results have shown the superiority of our proposal over the reference MD techniques.

References

1. Varga, A.P., Moore, R.K.: Hidden Markov model decomposition of speech and noise. In: Proc. ICASSP, pp. 845–848 (April 1990)
2. Huang, X., Acero, A., Hon, H.: Spoken language processing: A guide to theory, algorithm, and system development. Prentice Hall (2001)
3. Deng, L., Droppo, J., Acero, A.: Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features. IEEE Trans. Speech Audio Process. 12(3), 218–233 (2004)
4. Reddy, A.M., Raj, B.: Soft Mask Methods for Single-Channel Speaker Separation. IEEE Trans. Audio Speech and Language Process. 15(6), 1766–1776 (2007)
5. Cooke, M., Green, P., Josifovski, L., Vizinho, A.: Robust automatic speech recognition with missing and unreliable data. Speech Comm. 34(3), 267–285 (2001)
6. Raj, B., Seltzer, M.L., Stern, R.M.: Reconstruction of missing features for robust speech recognition. Speech Comm. 48(4), 275–296 (2004)
7. González, J.A., Peinado, A.M., Gómez, A.M., Ma, N., Barker, J.: Combining missing-data reconstruction and uncertainty decoding for robust speech recognition. In: Proc. ICASSP, pp. 4693–4696 (March 2012)
8. Raj, B., Singh, R.: Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition. In: Proc. ASRU, pp. 275–296, 65–70 (2005)
9. Faubel, F., Raja, H., McDonough, J., Klakow, D.: Particle filter based soft-mask estimation for missing-feature reconstruction. In: Proc. IWAENC (2008)
10. Hirsch, H.G., Pearce, D.: The Aurora experimental framework for the performance evaluations of the speech recognition systems under noisy conditions. In: ISCA ITRW ASR 2000, Paris, France (2000)
11. Hirsch, H.G.: Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task. Tech. Rep., STQ AURORA DSR Working Group (2002)

Automatic Speech Recognition Based on Ultrasonic Doppler Sensing for European Portuguese

João Freitas[1], António Teixeira[2], Francisco Vaz[2], and Miguel Sales Dias[1,3]

[1] Microsoft Language Development Center, Lisboa, Portugal
{i-joaof,miguel.dias}@microsoft.com
[2] Dep. Electronics Telecommunications & Informatics/IEETA, University of Aveiro, Portugal
{ajst,fvaz}@ua.pt
[3] ISCTE-Lisbon University Institute/ADETTI-IUL, Lisboa, Portugal

Abstract. Conventional Automatic Speech Recognition systems solely rely on acoustic information, making them susceptible to problems like environmental noise, privacy, information disclosure and also excluding users with speech impairments. An Ultrasonic Doppler Sensing (UDS) based interface may be used to tackle these issues since it does not rely on audio signal information. This paper describes the first speech recognition experiments based on UDS for European Portuguese (EP). The work here presented analyzes the UDS signal and explores the recognition of EP digits and minimal pairs of words that only differ on nasality of one of the phones. The results of our experiments show a best word error rate of 27.8% using data collected with the device at different distances from the speaker in an isolated word recognition problem.

Keywords: Ultrasonic Doppler Sensing, Silent Speech, European Portuguese, Nasality.

1 Introduction

Ultrasonic Doppler Sensing (UDS) of speech is one of the approaches reported in literature that is suitable for implementing a Silent Speech Interface (SSI) [1]. A SSI performs ASR in the absence of an intelligible acoustic signal and can be used to tackle problems such as environmental noise, privacy, information disclosure and in aiding users with speech impairments. This technique is based on the emission of a pure tone in the ultrasound range towards the speaker's face that is received by an ultrasound sensor tuned to the transmitted frequency. The reflected signal will contain Doppler frequency shifts proportional to the movements of the speaker's face. Based on the analysis of the Doppler signal, patterns of movements of the facial muscles, lips, tongue, jaw, etc., can be extracted [2]. The main advantages of this approach are the following: Non-invasive nature, since the device is completely non-obtrusive and it has been proven to work without requiring any attachments; the signal is not affected by environment noise in the audible frequency range; no acoustic audio signal is required, since this technique is based on the signal that contains Doppler frequency shifts caused by facial movements; the hardware used on this approach is commercially available and is very inexpensive.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 227–236, 2012.

The work here presented is a follow up of a preliminary analysis presented in [3] and focus on the development of an SSI on ultrasonic sensors for European Portuguese. We start by describing the used hardware and analyzing the obtained signal. Afterwards, we conduct several recognition experiments where we analyze the use of the sensor in two distinct positions and also the detection of nasality. Nasality is a relevant characteristic of EP and, as previous studies with different approaches have shown [4], it can be a relevant source of error.

The remainder of this document is structured as follows: Section 2 presents a description of previous work in ASR based on ultrasonic sensors. Section 3 describes the methodology used in this experiment. Section 4 describes an exploratory analysis of the signal and multiple recognition experiments. Finally, in Section 5 the conclusions of this paper are presented.

2 Background

Doppler Effect is the modification of the frequency of a wave when the observer and the wave source are in relative motion. If v_s and v_o are the speed of the source and the observer measured on the direction and sense observer-source, c is the propagation velocity of the wave on the medium and f0 the source frequency, the observed frequency will be:

$$f = \frac{c + v_o}{c + v_s} f_0 \tag{1}$$

Considering a standstill observer $v_o = 0$ and $v_s \ll c$ the following approximation is valid:

$$f = \left(1 - \frac{v_s}{c}\right) f_0 \text{ or } \Delta f = -\frac{v_s}{c} f_0 \tag{2}$$

We are interested in echo ultrasound to characterize the moving articulators of a Human speaker. In this case a moving body with a speed v (positive when the object is moving towards the emitter/receiver) reflects an ultrasound wave which frequency is measured by a receiver placed closely to the emitter. The observed Doppler shift will then be the double:

$$\Delta f = \frac{2v}{c} f_0 \tag{3}$$

Considering $c = 340 m/s$ as the sound air speed, a maximum articulator speed of 1m/s and 40 kHz ultrasound primary wave, the maximum frequency shift will be 235Hz.

2.1 State-of-the-Art

Ultrasonic sensors are used in a variety of applications that range from industrial automation to medical ultrasonography. In the area of speech, UDS have been previously used in voice activity detection [5], speaker identification [6], and synthesis [2]. Regarding speech recognition, ultrasonic devices were first applied to ASR in 1995 using an ultrasonic lip motion detector by Jennings and Ruck [7]. In this work, an experiment where the "Ultrasonic Mike", as they call it, is used as an input to an automatic lip reader with the aim of improving ASR in noisy environments by combining it with a conventional ASR system. The used hardware is constituted by an emitter and a receiver based on piezoelectric material and a 40 KHz oscillator to create a continuous wave ultrasonic signal. In the feature extraction phase, 10 LPC cepstral coefficients are extracted from the acoustic signal. The classification is based on Dynamic Time Warping (DTW) distances between the test utterances and the ones selected as ground truth. Best results for this work include an accuracy of 89% for the ultrasonic input alone using 4 template utterances, in a speaker dependent isolated digit recognition task, considering 5 test sessions and each session containing 100 utterances. For the cross-session scenario no higher than a 12.6% accuracy was achieved.

It was only a few years later, in 2007, that UDS was again applied to speech recognition by Zhu [8]. In their work an ASR experiment is conducted based on a statistical approach and a continuous speech recognition task are considered. In terms of hardware, Zhu, used an ultrasonic transmitter and receiver tuned to a resonant frequency of 40 kHz. The received signal is then multiplied by a 35.6 kHz sinusoid causing it to be centered at 4.4 kHz. This study collected 50 sequences of ten random digits of twenty speakers at a 15.2 cm distance relative to the sensors. For what feature extraction is concerned, the authors split the signal in frequency and magnitude sub bands and then features based on energy-band frequency centroids and frequency sub-band energy averages are extracted for each frame. The features were later projected to a lower dimensional space using Principal Component Analysis. The experiments were conducted using a landmark-based speech recognizer. The accuracy results for the ultrasonic approach were very similar across multiple noise levels, with a best result of 70.5% Word-Error Rate (WER).

In 2010, Srinivasan et al. [1], was able to improve previous results and achieved an overall accuracy of 33% also on a continuous digit recognition task. In this work Srinivasan et al. use similar hardware to the one previously described, however it added the synchronization of the two-channel (audio and ultrasound) output, the carrier was located at 8kHz and the sensor was positioned at 40.46cm. In terms of features, the authors have applied a FFT over the pre-processed signal and applied a Discrete Cosine Transform to the bins corresponding to the frequencies between 7 kHz and 9.5 kHz. For classification, HMM models with 16 states and one Gaussian per state were used. Best results for fast speech presented an accuracy of 37.75% and 18.17% for slow speech.

In the literature several other types of sensors that enable ASR in noisy environments or Voice Activity Detection (VAD) can be found such as, Bone conduction microphones, Physiological microphones (P-mics), throat microphones and the non-acoustic glottal electromagnetic sensors (GEMS). However, these devices have the drawback of needing to be mounted on the jaw bone, speaker's face or throat, restricting their appli-

cability or leaving the user uncomfortable. Hence, when compared with other secondary sensors the Ultrasonic Doppler sensors have the advantage of not needing to be mounted on the speaker and although their measurements is not as detailed as in P-mics or GEMS [9], the results for mutual information between UDS and acoustic speech signals is very similar to the ones reported for the other secondary devices [9]. When compared with vision devices such as, cameras, these sensors present a much lower cost, since an ultrasonic sensing setup can be arranged for less than $10.

The results for ultrasound-only approaches are still far from audio-only performance. Nonetheless, latest studies reveal viability and margin for improvement of this approach.

3 Methodology

3.1 Corpora

The European Portuguese UDS data collected in this study can be split into 2 corpora which we named: PT-DIGIT-UDS and PT-NW-UDS. The first corpus is similar to what was used in previous studies [1, 8] that address ASR based on UDS and consists in ten digits – um [ũ], dois [dojʃ], três [tɾeʃ], quatro [kwatɾu], cinco [sĩnku], seis [sɐjʃ], sete [sɛti], oito [ojtu], nove [nɔvi], dez [dɛʃ] - with the difference that we are using EP digits instead of English and that only isolated digits are considered. The second corpus consists in 4 pairs of EP common words that only differ on nasality of one of the phones (minimal pairs, e.g. Cato/Canto [katu]/[kɐ̃tu] or Peta/Penta [petɐ]/[pẽtɐ] – see [4] for more details). For the first corpus we have recorded 6 speakers – 4 male and 2 female - and each speaker recorded an average of 6 utterances for each prompt. For the second corpus we have recorded the same 6 speakers and each speaker recorded 4 observations for each prompt, giving a total of 552 utterances for both corpora. Most of the recordings occurred at a distance of approximately 40cm from the speaker to the sensor with exception for an extra session of 40 utterances using the PT-DIGIT-UDS prompts which was recorded at a distance of 12cm for comparison and analysis.

3.2 Simultaneous Acquisition of Speech and Doppler

To study the Doppler effect of a speaker a dedicated circuit board was developed based on the work of Zhu [8]. It includes 1) the ultrasound transducers (400ST and 400SR working at 40kHz) and a microphone to receive the speech signal; 2) a crystal oscillator at 7.2MkHz and frequency dividers to obtain 40 and 36 kHz; 3) all amplifiers and linear filters needed to process the echo signal and the speech. The board is placed in front of the speaker and the echo signal is the sum of the contributions of all the articulators. If the ultrasound generated is a sine wave $sin2\pi f_0\, t$, an articulator with a velocity v_i will produce an echo wave that can be characterized by:

$$x_i = a_i \sin 2\pi f_0 \left(t + \frac{2}{c} \int_0^t v_i \, d\tau + \varphi_i \right) \tag{4}$$

a_i, φ_i are parameters defining the reflection and are function of the distance. Although they are also function of time they are slow varying and are going to be considered constants. The total signals will be the sum for all articulators and the moving parts of the face of the speaker

$$x = \sum_i a_i \sin 2\pi f_0 \left(t + \frac{2}{c} \int_0^t v_i \, d\tau + \varphi_i \right) \tag{5}$$

The signal is a sum of frequency modulated signals. It was decided to make a frequency translation: multiplying the echo signal by a sine wave of a frequency $f_a = 36kHz$ and low passing the result it is obtained a similar frequency modulated signal centered at $f_1 = f_0 - f_a$,i.e., $f_1 = 4kHz$

$$d = \sum_i a_i \sin 2\pi f_1 \left(t + \frac{2}{c} \int_0^t v_i \, d\tau + \varphi_i \right) \tag{6}$$

This operation is made on the board and it was used an analog multiplier AD633. The Doppler echo signal and speech are now digitized at 44.1 kHz and the following process is digital and implemented in Matlab.

3.3 Signal Pre-processing

After some exploratory analysis and based on previous work [1] the acquired signal is first zero-meaned, then the signal is passed through a 3 sample moving average filter to suppress the 4 kHz carrier and later a difference operator is applied. Fig. 1 shows the resulting spectrogram after this preprocessing.

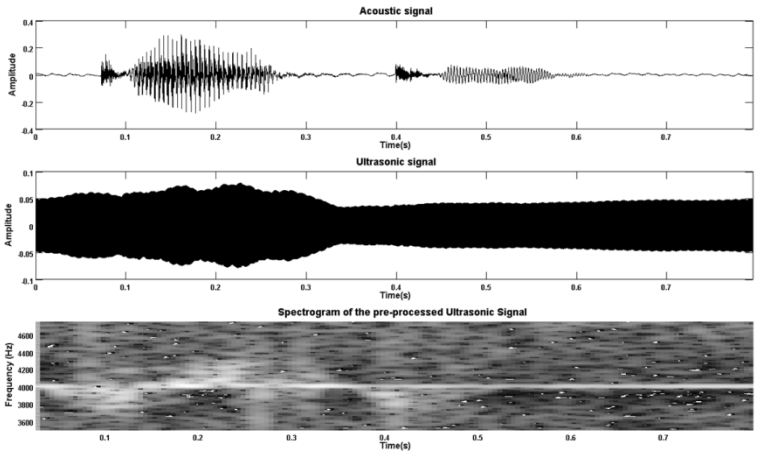


Fig. 1. Signals for the EP word *Cato*, from top to bottom: Acoustic signal, raw ultrasonic signal and spectrogram of the pre-processed ultrasonic signal

3.4 Feature Extraction and Classification

Before conducting feature extraction, we preprocess the signal as described earlier. After the pre-processing stage, we split the signal into 50ms frames and extract a set of features based on the work of Srinivasan et al. [1]. We start by calculating a Discrete Fourier transform (DFT) with second-order Goertzel algorithm over the preprocessed signal for the interval of 3500 Hz to 4750 Hz. Finally, a DCT is applied to the DFT results to de-correlate the signal and extract the first 38 coefficients, which contain most of the signal energy.

In our recognition pipeline, we start by pre-processing the acquired UDS signal and extracting the features described in the previous section. After the feature extraction phase we need to classify to which class they belong. Based on the number of available observations and considering the limited vocabulary, we have chosen to use DTW, a technique which was also employed by 7 to classify this type of signals. The classification algorithm uses a 10-fold cross-validation for partitioning the data and for each observation from the test group, compares the representative example and selects the word that provides the minimum distance in the feature vector domain.

4 Experimental Results

The following section describes the first recognition experiments based on UDS which are not applied to English. These experiments analyze the recognition of EP digits, the minimal pairs described in section 3.1, and a combination of both based on a fixed group of features. Results regarding the effect of the sensor distance to the speaker are also reported.

4.1 Exploratory Analysis

After the pre-processing stage, a first exploratory analysis of the signal shows a clear difference between EP digits. If a discriminative analysis of the signal depicted in Fig. 2 is performed, it is noticeable that the digits that require more movement from visible articulators present a more distinguishable signal. For example, if we compare an observation of *um* (one) with an observation of *quatro* (four) a clear magnitude difference is visible across time.

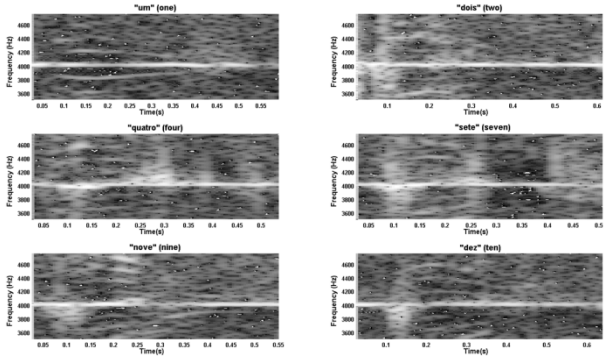


Fig. 2. Spectrogram of the pre-processed signal for 6 EP digits and frequencies between 3500 and 4750 for a single speaker

Fig. 3 shows a similar analysis performed for the words "*cato*" and "*canto*", a minimal pair where the only difference is the presence of nasality. In this case, dissimilarities are more subtle, but nonetheless they seem to be present between the two words. Regarding the cross speaker signals, the signals seem to have relevant differences between them.

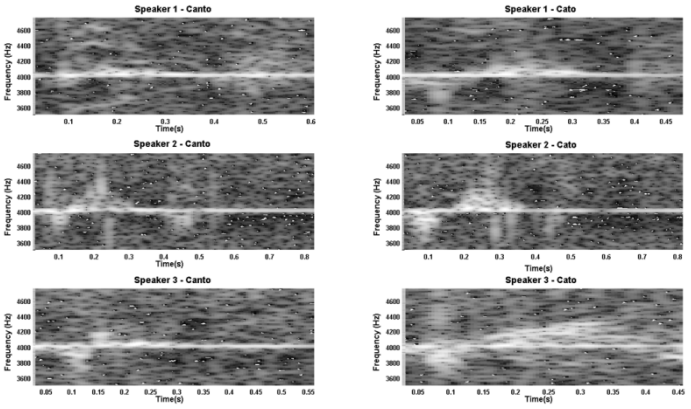


Fig. 3. Spectrogram of the words "Cato" and "Canto" for 3 speakers

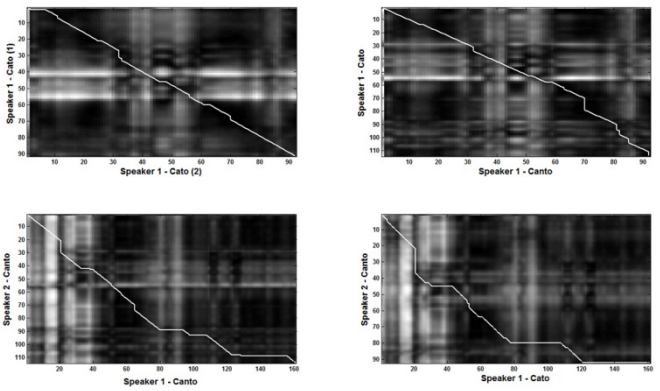


Fig. 4. Distance comparison between the word *canto* and *cato* for different speakers using the Matlab algorithm from [10]. The white line represents the lower DTW distances found across time.

If we analyze the signal using Dynamic Time Warping (DTW), which provides temporal alignment to time varying signals that have different durations, differences between minimal pairs can be noticed. In Fig. 4, the DTW is applied to several pairs of words observations and we depict the DTW distance results, by means of gray scale coding of such results. The similarity between two words is given by the

smallest DTW distance between them across time, thus when two words are the same, the lowest distance will lay in the image's diagonal. A highest discrepancy is found when we compare different speakers, again showing that the signal is highly dependent of the speaker. We have also noticed that these differences are more accentuated when we compare different genders.

4.2 Recognition Results

Table 1 presents the results of three test conditions differing on the used corpus.

Table 1. Classification results for the following sets of data: PT-DIGITS-UDS, PT-NW-UDS and a combination of both

	PT-DIGITS-UDS	**PT-NW-UDS**	**Both**
Word Error Rate	**36.1%**	42.7%	45.3%

Results show that we are able to achieve the best word error rate of 36.1% for a vocabulary of 10 digits (PT-DIGITS-UDS) on an isolated word recognition problem. It is also noticeable that if we consider a smaller vocabulary with only 8 words based on minimal pairs (PT-NW-UDS) the error rate increases to 42.7%. If we analyze the confusion matrix for this run depicted in Table 2, a large error incidence can be found in the minimal pairs. For instance, in the case of the word *mato*, 87.5% of the incorrect observations were classified as *manto*. The reverse case (*manto* being classified as *mato*) is also noticeable since 75% of the incorrect observations were classified as *mato*. This problem is also evident for the case of *cato* being confused with *canto* and for the pair *peta/penta*. Nonetheless, for the minimal pair *tito/tinto* this is not verified.

Additionally, we have also run an experiment using a joint vocabulary of 18 words, based on the previous two vocabularies, obtaining a slight worse error rate of 45.3%. A considerable part of the error stills occurs in the *mato/manto* and *peta/penta* minimal pairs (for the case of *mato* 42.9% of the incorrect observations were classified as *manto* and for the case of *penta* 55.6% of the incorrect observations were classified as *peta*).

Table 2. Confusion matrix for the recognition experiment with the PT-NW-UDS corpus

		Output							
	Words	**Cato**	**Canto**	**Mato**	**Manto**	**Peta**	**Penta**	**Tito**	**Tinto**
Input	**Cato**	**21**	**6**	1	1	2	0	2	3
	Canto	0	12	2	2	2	2	1	6
	Mato	1	0	**15**	**7**	0	0	0	0
	Manto	0	0	**6**	**13**	1	0	1	0
	Peta	0	4	0	1	**9**	**5**	4	0
	Penta	0	0	0	0	**6**	**15**	2	2
	Tito	1	0	0	0	4	1	13	1
	Tinto	1	2	0	0	0	1	1	12

4.3 Distance Effect

As mentioned before, we have also recorded an extra session considering a closer distance with the device positioned 12cm from the speaker. The goal is to investigate the mismatch effect caused by the distance change in the traininig patterns and the test conditions, further analysing distance limitations. Thus, we have ran the previous experiment using the following data distributions: 1) Use only the PT-DIGIT-UDS data recorded at 12cm; 2) Use the PT-DIGIT-UDS data recorded at 12cm as a test group, creating a mismatch between train and test; 3) Use the PT-DIGIT-UDS data from the previous experiment plus the PT-DIGIT-UDS data recorded at 12cm for train and test. The obtained results are presented in Table 3.

Table 3. Classification results for three data sets: 1) Only PT-DIGIT-UDS data recorded at 12cm. 2) Use only PT-DIGIT-UDS data recorded at 12m in the test group and data recorded at 40cm in the train group. 3) Use all PT-DIGIT-UDS data for classification.

	12cm data only	**12cm data as test**	**All data**
Word Error Rate	35.0%	35.0%	**27.8%**

Results for this experiment include 35% error rate for the first two data distributions and 27.8% for the last distribution using all acquired data.

4.4 Discussion

The exploratory analysis of the signal has shown differences between the selected words, especially in those where the articulators' movement is more intense. It is also visible a difference across speaker, which corroborates the results achieved by Jennings and Ruck [7] where the performance of the system has a drastic performance reduction when cross-speaker recognition is considered.

Previous recognition results have achieved a WER of 67% in a continuous speech recognition task of 10 English digits. Although in our case we are considering an isolated digit recognition task on the same vocabulary size and the tests conditions are not the same, if a direct comparison was made with the best result of 27.8% WER, we find a relative improvement of 58.6%. Additionally, the error analysis seems to indicate that minimal pairs such as *mato/manto* and *peta/penta* may cause recognition problems for an interface based on this approach. The results show viability for a vocabulary increase beyond the 10 words previously presented. The distance comparison experiment seems to indicate that data collected at 12cm and 40cm present similar features since the error rate was close to the recognition experiment that only used 40cm data when using the 12cm data for testing and the 40cm data for training. It is also worth noting the relative improvement in the error rate of 23.1% by joining both types of data.

5 Conclusion and Future Work

In this work, a first experiment of UDS-based speech recognition for EP is presented. It describes the device used in the acquisition of this type of data and an analysis to the signal that shows viability for using this type of approach in speech recognition experiments for EP. The results can be arguably placed at the level of the state of the art, with a best word error rate of 27.8% in an isolated word recognition problem across several speakers. This result was achieved using data acquired at different distances, also demonstrating that UDS data collected with the device at closer distances might be beneficial for the recognition performance.

Acknowledgements. This work was partially funded by Marie Curie Golem (ref.251415, FP7-PEOPLE-2009-IAPP), by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology in the context of the project FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011). The authors would also like to thank all the speakers involved in the experiment.

References

1. Srinivasan, S., Raj, B., Ezzat, T.: Ultrasonic sensing for robust speech recognition. In: Internat. Conf. on Acoustics, Speech, and Signal Processing (2010)
2. Toth, A.R., Kalgaonkar, K., Raj, B., Ezzat, T.: Synthesizing speech from Doppler signals. In: IEEE International Conference on Acoustics Speech and Signal Processing, pp. 4638–4641 (2010)
3. Freitas, J., Teixeira, A., Dias, M.S., Bastos, C.: Towards a Multimodal Silent Speech Interface for European Portuguese. In: Ipsic, I. (ed.) Speech Technologies. InTech (2011) ISBN: 978-953-307-996-7
4. Freitas, J., Teixeira, A., Dias, M.S.: Towards a Silent Speech Interface for Portuguese: Surface Electromyography and the nasality challenge. In: Int. Conf. on Bio-inspired Systems and Signal Processing, Vilamoura, Algarve, Portugal (2012)
5. Kalgaonkar, K., Raj, B., Hu, R.: Ultrasonic doppler for voice activity detection. IEEE Signal Processing Letters 14(10), 754–757 (2007)
6. Kalgaonkar, K., Raj, B.: Ultrasonic doppler sensor for speaker recognition. In: Internat. Conf. on Acoustics, Speech, and Signal Processing (2008)
7. Jennings, D.L., Ruck, D.W.: Enhancing automatic speech recognition with an ultrasonic lipmotion detector. In: Internat. Conf. on Acoustics, Speech, and Signal Processing, Detroit (1995)
8. Zhu, B.: Multimodal speech recognition with ultrasonic sensors. Master's thesis. Massachusetts Institute of Technology, Cambridge, Massachusetts (2008)
9. Hu, R., Raj, B.: A Robust Voice Activity Detector Using an Acoustic Doppler Radar. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 171–176 (2005)
10. Ellis, D.: Dynamic Time Warp (DTW) in Matlab. Web resource (2003), `http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/`

Integrating a State-of-the-Art ASR System into the Opencast Matterhorn Platform

Juan Daniel Valor Miró, Alejandro Pérez González de Martos, Jorge Civera, and Alfons Juan

Universitat Politècnica de València
Camino de Vera s/n, 46022 Valencia, Spain
{juavami,alpegon2}@upv.es, {jcivera,ajuan}@dsic.upv.es
http://prhlt.iti.upv.es/

Abstract. In this paper we present the integration of a state-of-the-art ASR system into the Opencast Matterhorn platform, a free, open-source platform to support the management of educational audio and video content. The ASR system was trained on a novel large speech corpus, known as poliMedia, that was manually transcribed for the European project transLectures. This novel corpus contains more than 115 hours of transcribed speech that will be available for the research community. Initial results on the poliMedia corpus are also reported to compare the performance of different ASR systems based on the linear interpolation of language models. To this purpose, the in-domain poliMedia corpus was linearly interpolated with an external large-vocabulary dataset, the well-known Google N-Gram corpus. WER figures reported denote the notable improvement over the baseline performance as a result of incorporating the vast amount of data represented by the Google N-Gram corpus.

Keywords: Speech Recognition, Linear Combination, Language Modeling, Google N-Gram, Opencast Matterhorn.

1 Introduction

Online educational repositories of video lectures are rapidly growing on the basis of increasingly available and standardized infrastructure. Transcription and translation of video lectures is needed to make them accessible to speakers of different languages and to people with disabilities. Automatic transcription in these domains is however a challenging task due to many factors such as unfavourable recording quality, high rate out-of-vocabulary words or multiplicity of speakers and accents. Therefore, human intervention is needed to achieve accurate transcriptions. Recently, approaches to hybrid transcription systems have been proposed based on fully manual correction of automatic transcriptions, which are not practical nor comfortable to the users who perform this time-consuming task. In this paper we present an intelligent user interactive semi-automatic speech recognition system to provide cost-efficient solutions to produce accurate transcriptions. This speech recognition system is being developed within the

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 237–246, 2012.

framework of the European *transLectures* project [1], along the lines of other systems, such as JANUS-II [2], UPC RAMSES [3] or SPHINX-II [4]. Resulting transcriptions may be translated into other languages, as it is the case of the transLectures project, or other related project, such as SUMAT [5].

Initial results are reported on the recently created poliMedia corpus using a linear combination of language models [6,7,8,9]. This linear combination aims at alleviating the problem of out-of-vocabulary words in large-scale vocabulary tasks with a great variety of topics. The baseline automatic speech recognition (ASR) system is based on the RWTH ASR system [10,11] and the SRILM toolkit [12], both state-of-the-art software in speech and language modeling, respectively. In this work, we present significant improvements in terms of WER over the baseline when interpolating the baseline language model with a language model trained on the well-known *Google n-gram* dataset [13]. Furthermore, details about the integration of this speech recognition system into the open-source videolecture platform Matterhorn are also provided. The integration into Matterhorn enables user-assisted corrections and therefore, it guarantees high quality transcriptions.

The rest of this paper is organised as follows. First, the novel freely available poliMedia corpus is presented in Section 2. Secondly, the Opencast Matterhorn platform is introduced in Section 3. In Section 4, the backend RWTH ASR system is described, and initial results are reported in Section 5. Finally, conclusions are drawn and future lines of research are depicted in Section 6.

2 The poliMedia Corpus

poliMedia [14] is a recent, innovative service for creation and distribution of multimedia educational content at the *Universitat Politècnica de València* (UPV). It is mainly designed for UPV professors to record courses on video lectures lasting 10 minutes at most. Video lectures are accompanied with time-aligned slides and recorded at specialised studios under controlled conditions to ensure maximum recording quality and homogeneity. As of today, poliMedia catalogue includes almost 8000 videos accounting for more than 1000 hours. Authors retain all intellectual property rights and thus not all videos are accessible from outside the UPV. More precisely, about 2000 videos are openly accessible.

poliMedia is one the two videolectures repositories along with Videolectures.NET[1] that are planned to be fully transcribed in the framework of the European project transLectures[2]. To this purpose, 704 videolectures in Spanish corresponding to 115 hours were manually transcribed using the tool Transcriber [15], so as to provide in-domain dataset for training, adaptation and internal evaluations in the transLectures project (see Table 1). These transcribed videolectures were selected so that authors had granted open access to their content. This fact guarantees that the poliMedia corpus can be used by the research community beyond the scope of the transLectures project.

[1] http://videolectures.net
[2] http://translectures.eu

Most of the videos in poliMedia were annotated with topic and keywords. More precisely, 94% of the videos were assigned a topic and 83% were described with keywords. However, these topics and keywords were not derived from a thesaurus, such as EuroVoc. Speakers were also identified for each transcription.

Table 1. Basic statistics on the poliMedia corpus

Videos	704
Speakers	111
Hours	115
Sentences	40K
Running words	1.1M
Vocabulary (words)	31K
Singletons (words)	13K

3 The Opencast Matterhorn Platform

Matterhorn[3] is a free, open-source platform to support the management of educational audio and video content. Institutions will use Matterhorn to produce lecture recordings, manage existing video, serve designated distribution channels, and provide user interfaces to engage students with educational videos.

Matterhorn is an open source; this means that the product is fully based on open source products. The members of the Opencast Community have selected Java as programming language to create the necessary applications and a Service-Oriented Architecture (SOA) infrastructure. The overall application design is highly modularised and relies on the OSGi (dynamic module system for Java) technology. The OSGi service platform provides a standardised, component-oriented computing environment for cooperating network services.

Matterhorn is as flexible and open as possible and further extensions should not increase the overall complexity of building, maintaining and deploying the final product. To minimise the coupling of the components and third party products in the Matterhorn system, the OSGi technology provides a service-oriented architecture that enables the system to dynamically discover services for collaboration. Matterhorn uses the Apache Felix [16] implementation of the OSGI R4 Service Platform [17] to create the modular and extensible application.

The main goal in transLectures is to develop tools and models for the Matterhorn platform that can obtain accurate transcriptions by intelligent interaction with users. For that purpose, an HTML5 media player prototype has been built in order to provide a user interface to enable interactive edition and display of video transcriptions (see Figure 1). This prototype offers a main page where available poliMedia videolectures are listed according to some criteria. Automatic video transcriptions are obtained from the ASR system when playing a particular video.

[3] http://opencast.org/matterhorn

Since automatic transcriptions are not error free, an interactive transcription editor allows intelligent user interaction to improve transcription quality. However, as users may have different preferences while watching a video, the player offers two interaction models depending on the user role: simple user and collaborative user (prosumers).

Fig. 1. HTML5 player and interactive transcription editor for collaborative users

Simple users are allowed to interact in a very simplistic manner, just showing their liking about the transcriptions. However, collaborative users may provide richer feedback to correct transcriptions. As shown in Figure 1, collaborative users have an *edit transcription* button available on the player control bar that enables the transcription editor panel. The editor panel is situated next to the video. It basically contains the transcription text, which is shown synchronously with the video playback. Clicking on a transcription word or sentence enables the interactive content modification. User corrections are sent to the speech recognition module through a web service, so corrections are processed and new transcription hypothesis are offered back to the user. Some other user-friendly features such as keyboard shortcuts and useful editing buttons are also available. Simple users have no edit transcription button available as they are not expected

to be working on transcription editing. Instead, a *low quality transcription* button appears so they can report that the transcription quality is not good enough.

The current HTML5 prototype is a proof-of-concept version that works with pre-loaded transcriptions, however the version currently being developed communicates with the ASR system through a web service implemented for that purpose. Figure 2 illustrates the system architecture and the communication process.

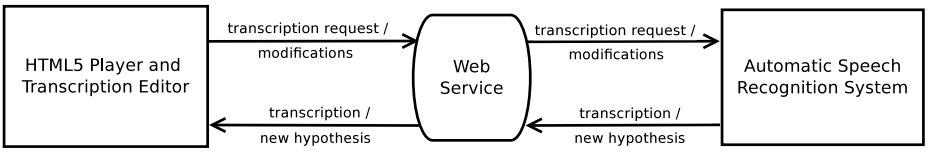


Fig. 2. HTML5 player and ASR system communication

The next step is to integrate the developed interactive ASR system into the Matterhorn infrastructure. There are many different approaches to perform this integration. Our proposal lets an external system manage all the transcriptions, so there will not be necessary to add nor store them in any way into the current Matterhorn system[4]. In addition, two primary tasks are involved in the integration process into Matterhorn. Both of them require an interface to enable communication between Matterhorn and the ASR system. For that purpose, a RESTful Web Service has been implemented to allow media uploading, retrieve the processing status of a particular recording, request a video transcription, send transcription modifications and other functionalities.

The first task would be to define a new Matterhorn workflow operation to transfer the audio data of the new media to the ASR system through the REST service mentioned before, so as to obtain automatic transcriptions for every recording uploaded to the Matterhorn platform. This task will involve the implementation of a new Matterhorn service.

The second part is to replace or adapt the Matterhorn Engage Player to enable transcription edition, along the lines of the HTML5 player prototype indicated previously. The player must obtain and transmit every transcription-related information through the REST Web Service in a similar way as the HTML5 prototype did (see Figure 2). Here the main problem is the addition of new features to the Flash-based Matterhorn player, since it is not straightforward to implement the transcription functionalities provided by the HTML5-based player. Our solution is to use an alternative open-source Matterhorn engage player based on HTML5 called Paella Engage Player[5].

[4] http://opencast.jira.com/wiki/display/MH/MediaPackage+Overview

[5] http://unconference.opencast.org/sessions/paella-html5-matterhorn-engage-player

4 The RWTH ASR System

Our baseline ASR system is the RWTH ASR system [10,11] along with the SRILM toolkit [12]. The RWTH ASR system includes state-of-the-art speech recognition technology for acoustic model training and decoding. It also includes speaker adaptation, speaker adaptive training, unsupervised training, a finite state automata library, and an efficient tree search decoder. SRILM toolkit is a widespread language modeling toolkit which have been applied to many different natural language processing applications.

In our case, audio data is extracted from videos and preprocessed to extract the normalized acoustic features obtaining the Mel-frequency cepstral coefficients (MFCCs) [18]. Then, triphoneme acoustic models based on a prebuilt cart tree are trained adjusting parameters such as number of states, gaussian components, etcetera on the development set. The lexicon model is obtained in the usual manner by applying a phonetic transliteration to the training vocabulary. Finally, n-gram language models are trained on the transcribed text after filtering out unwanted symbols such as punctuation marks, silence annotations and so on.

In this work, we propose to improve our baseline system by incorporating external resources to enrich the baseline language model. To this purpose, we consider the linear combination of an in-domain language model, such as that trained on the poliMedia corpus, with an external large out-domain language model computed on the Google N-Gram corpus [13]. A single parameter λ governs the linear combination between the poliMedia language model and the Google N-Gram model, being optimised in terms of perplexity on a development set.

5 Experimental Results

In order to study how the linear combination of language models affects the performance, in terms of WER, of an ASR system in the poliMedia corpus, a speaker-independent partition in training, development and test sets was defined. The statistics of this partition can be found in Table 2. Topics included in the development and test sets range from technical studies such as architecture, computer science or botany, to art studies such as law or marketing.

The baseline system, including acoustic, lexicon and language models, was trained only on the poliMedia corpus. System parameters were optimised in terms of WER on the development set. A significant improvement of more than 5 points of WER was observed when moving from monophoneme to triphoneme acoustic models. Triphoneme models were inferred using the conventional CART model using 800 leaves. In addition, the rest of parameters to train the acoustic model were 2^9 components per Gaussian mixture, 4 iterations per mixture and 5 states per phoneme without repetitions. The language model was an interpolated trigram model with Kneser-Ney discount. Higher order n-gram models were also assessed, but no better performance was observed.

Provided the baseline system, a set of improvements based on the language model were proposed and evaluated. The baseline language model solely trained

Table 2. Basic statistics on the poliMedia partition

	Training	Development	Test
Videos	559	26	23
Speakers	71	5	5
Hours	99	3.8	3.4
Sentences	37K	1.3K	1.1K
Vocabulary	28K	4.7K	4.3K
Running words	931K	35K	31K
OOV (words)	-	4.6%	5.6%
Perplexity	-	222	235

on poliMedia corpus was interpolated with the Google N-Gram corpus [13]. To this purpose, we unify all Google N-Gram datasets, which are initially splitted by years, in a single, large file. Then, we train a trigram language model using Google N-Gram that was interpolated with the poliMedia language model. These two language models were interpolated to minimise perplexity on the development set. This interpolation was performed using a particular vocabulary in the case of Google N-Gram, ranging from that vocabulary matching that of poliMedia (poliMedia vocab), over the 20.000 most frequent words in the Google N-Gram corpus (20K vocab), to the 50.000 most frequent words (50K vocab). In this latter experiment, approximate values of interpolation weights are 0.65 for the poliMedia language model and 0.35 for the Google N-Gram language model.

The idea behind these experimental setups was to evaluate the effects, in terms of WER, of an increasing vocabulary coverage using external resources in the presence of a comparatively small in-domain corpus such as poliMedia. Experimental results are shown in Table 3.

Table 3. Evolution of WER above the baseline for the RWTH ASR system, as a result of interpolating the poliMedia language model with an increasingly larger vocabulary language model trained on the Google N-Gram corpus

System	WER	OOV
baseline	39.4	5.6%
poliMedia vocab	34.6	5.6%
20K vocab	33.9	4.4%
50K vocab	33.7	3.5%

As reported in Table 3, there is a significant improvement of 5.7 points of WER over the baseline when considering a language model trained with the 50K most frequent words in the Google N-Gram corpus. As expected, the decrease in WER is directly correlated with the number of Out-Of-Vocabulary words (OOVs) in the test set, since the Google N-Gram corpus provides a better vocabulary coverage.

A similar trend is observed when comparing perplexity figures between the baseline and poliMedia vocab systems. Perplexity significantly drops from 235 to 176 just by interpolating our baseline poliMedia language model with the Google N-Gram language model that only considers the poliMedia vocabulary. Perplexity figures with 20K and 50K vocab are not comparable to the previous ones, since the size of the vocabulary is not the same. Note that by adding more vocabulary from the Google N-Gram dataset, the number of OOVs is reduced, but also more useless words are added to the final language model. This causes that the improvement in terms of WER is not so significant when going from 20K to 50K vocabulary. Further experiments with 2-gram and 4-gram language model were carried out. WER figures with 2-gram were two points below on average, while 4-gram results were similar to those obtained with 3-grams.

6 Conclusions and Future Work

In this paper we have presented the integration of a state-of-the-art ASR system into the Opencast Matterhorn platform. This system was trained on a novel large speech corpus, known as poliMedia, that was manually transcribed for the European project transLectures. This novel corpus contains more than 115 hours of transcribed speech that will be available for the research community.

Initial results on the poliMedia corpus are also provided to compare the performance of different systems based on the linear interpolation of language models. To this purpose, the in-domain poliMedia corpus was linearly interpolated with an external large-vocabulary dataset, the well-known Google N-Gram corpus. WER figures reported denote the notable improvement over the baseline performance as a result of incorporating the vast amount of data contained in the Google N-Gram corpus.

Regarding the backend ASR system, various aspects need to be considered for future research. A simple manner to improve our initial results is to perform an intelligent data selection from the Google N-Gram corpus based on a chronological criteria such as the year of publication, or inspired on a simple, yet effective, method such that presented in [19]. In this latter case, only infrequent n-grams in poliMedia will be enriched with counts computed in large external resources such as the Google N-Gram corpus. Obviously, the extension of the vocabulary size to 100K words or greater may provide little reductions in WER values, but not significants compared to the computational cost required to run such an experiment.

In any case, ASR accuracy is still far from producing fully automatic high-quality transcriptions, and human intervention is still needed in order to improve transcriptions quality. However, user feedback can be exploited to minimise user effort in future interactions with the system [20]. New features need to be developed and integrated into the Matterhorn platform to achieve an effective user interaction. The resulting prototype will not only be evaluated under controlled laboratory conditions, but also in real-life conditions in the framework of the transLectures project.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 287755. Also supported by the Spanish Government (MIPRCV "Consolider Ingenio 2010" and iTrans2 TIN2009-14511) and the Generalitat Valenciana (Prometeo/2009/014).

References

1. UPVLC, XEROX, JSI-K4A, RWTH, EML, DDS: transLectures: Transcription and Translation of Video Lectures. In: Proc. of EAMT, p. 204 (2012)
2. Zhan, P., Ries, K., Gavalda, M., Gates, D., Lavie, A., Waibel, A.: JANUS-II: towards spontaneous Spanish speech recognition 4, 2285–2288 (1996)
3. Nogueiras, A., Fonollosa, J.A.R., Bonafonte, A., Mariño, J.B.: RAMSES: El sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC. In: VIII Jornadas de I+D en Telecomunicaciones, pp. 399–408 (1998)
4. Huang, X., Alleva, F., Hon, H.W., Hwang, M.Y., Rosenfeld, R.: The SPHINX-II Speech Recognition System: An Overview. Computer, Speech and Language 7, 137–148 (1992)
5. Speech and Language Technology Group. Sumat: An online service for subtitling by machine translation (May 2012), http://www.sumat-project.eu
6. Broman, S., Kurimo, M.: Methods for combining language models in speech recognition. In: Proc. of Interspeech, pp. 1317–1320 (2005)
7. Liu, X., Gales, M., Hieronymous, J., Woodland, P.: Use of contexts in language model interpolation and adaptation. In: Proc. of Interspeech (2009)
8. Liu, X., Gales, M., Hieronymous, J., Woodland, P.: Language model combination and adaptation using weighted finite state transducers (2010)
9. Goodman, J.T.: Putting it all together: Language model combination. In: Proc. of ICASSP, pp. 1647–1650 (2000)
10. Lööf, J., Gollan, C., Hahn, S., Heigold, G., Hoffmeister, B., Plahl, C., Rybach, D., Schlüter, R., Ney, H.: The rwth 2007 tc-star evaluation system for european english and spanish. In: Proc. of Interspeech, pp. 2145–2148 (2007)
11. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., Ney, H.: The rwth aachen university open source speech recognition system. In: Proc. of Interspeech, pp. 2111–2114 (2009)
12. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proc. of ICSLP (2002)
13. Michel, J.B., et al.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176–182
14. Turro, C., Cañero, A., Busquets, J.: Video learning objects creation with polimedia. In: 2010 IEEE International Symposium on Multimedia (ISM), December 13-15, pp. 371–376 (2010)
15. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication Special Issue on Speech Annotation and Corpus Tools 33(1-2) (2000)
16. Apache. Apache felix (May 2012), http://felix.apache.org/site/index.html
17. Osgi alliance. osgi r4 service platform (May 2012), http://www.osgi.org/Main/HomePage
18. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition 54(4), 543–565 (2012)

19. Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., Casacuberta, F.: Does more data always yield better translations? In: Proc. of EACL, pp. 152–161 (2012)
20. Sánchez-Cortina, I., Serrano, N., Sanchis, A., Juan, A.: A prototype for interactive speech transcription balancing error and supervision effort. In: Proc. of IUI, pp. 325–326 (2012)

Speech Reconstruction by Sparse Linear Prediction*

Ján Koloda, Antonio M. Peinado,
and Victoria Sánchez

Dpt. Teoría de la Señal, Telemática y Comunicaciones,
Centro de Investigación en Tecnologías de la Información y de las Comunicaciones,
18071-Granada, Spain
{janko,amp,victoria}@ugr.es
http://tstc.ugr.es, http://citic.ugr.es

Abstract. This paper proposes a new variant of the least square autoregressive (LSAR) method for speech reconstruction, which can estimate via least squares a segment of missing samples by applying the linear prediction (LP) model of speech. First, we show that the use of a single high-order linear predictor can provide better results than the classic LSAR techniques based on short- and long-term predictors without the need of a pitch detector. However, this high-order predictor may reduce the reconstruction performance due to estimation errors, especially in the case of short pitch periods, and non-stationarity. In order to overcome these problems, we propose the use of a sparse linear predictor which resembles the classical speech model, based on short- and long-term correlations, where many LP coefficients are zero. The experimental results show the superiority of the proposed approach in both signal to noise ratio and perceptual performance.

Keywords: Speech reconstruction, error concealment, sparse linear prediction, least squares, autoregressive model.

1 Introduction

Speech Reconstruction is a subject that has been widely treated in the speech community and which has a number of applications. Thus, we can mention audio restoration, where short signal segments completely degraded must be recovered from adjacent segments as it frequently occurs in old recordings. Also, in Voice-over-IP (VoIP) systems based on intraframe codecs, the real time constraints imposed by the transmission protocols may cause a packet loss problem which finally results in the loss of speech segments.

In order to perform the reconstruction of a lost signal segment, some sort of sample interpolation or extrapolation using adjacent and correctly received samples must be applied [1]. This can be a difficult task. Fortunately, in the case of speech there exists a well-known signal production model based on linear prediction (LP)

* This work has been supported by the Spanish MEC/FEDER project TEC 2010-18009.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 247–256, 2012.

which is employed by many reconstruction methods. Thus, we have the least square autoregressive (LSAR) method [2], which carries out an iterative interpolation of the lost samples from the adjacent ones applying the LP model and a least squares (LS) estimation. Other methods also based on LP focus on the estimation of the LP excitation (LP residual) [3,4]. Also, the LP spectrum has been combined with sinusoidal models of the excitation for signal extrapolation [5].

In this paper we will focus on the class of LSAR signal interpolators, where the missing samples are LS-estimated according to a previous estimation of the LP model. Although the basic LSAR [2] just uses a short-term predictor, better results can be obtained when long-term prediction is also considered as it is common practice in speech coding [6]. A first drawback of this approach is that it requires the use of a pitch detector which may be affected by detection errors. This can be avoided using a single high-order predictor which accounts for both short- and long-term correlations. The prediction order must be large enough as to cover the longest possible correlations (due to the longest possible pitch). Although this approach increases the computational cost, we will show that it results in a better reconstruction performance.

The use of a single high-order predictor for LSAR is a simple and compact solution. However, it does not follow the classical speech model based on short- and long-term predictors. This involves that many LP coefficients that are forced to be zero by this speech model can have now non-zero values, which can be interpreted as a sort of estimation noise. Also, it must be considered that a high-order predictor may be more affected by non-stationarity. For example, and as it is shown later, this effect can degrade the performance for the case of relatively small pitch values since the LP order is likely much larger than necessary. This problem has been recently addressed by the application of sparse linear prediction (SLP) [7,8]. The SLP idea consists in the optimization of a single high-order linear predictor which maintains as much as possible the high sparsity level involved by the classical speech model. The underlying philosophy of SLP is that of predicting the missing samples by employing as few adjacent samples as possible. This idea has already been successfully applied by the authors to video packet loss concealment [9] and will be adapted here to speech reconstruction by LSAR methods.

The paper is organized as follows. Section 2 is devoted to the review and analysis of LSAR techniques. Then, the proposed SLP method is developed in Section 3 and the simulation results are shown and commented in Section 4. Finally, the main conclusions are summarized.

2 Least Square Autoregressive (LSAR) Interpolation

Let us review now the basic LSAR interpolation algorithm of reference [2]. According to the linear prediction model of speech signals, a sample $x(m)$ is modeled as,

$$x(m) = \sum_{k=1}^{P} a_k x(m-k) + e(m) \tag{1}$$

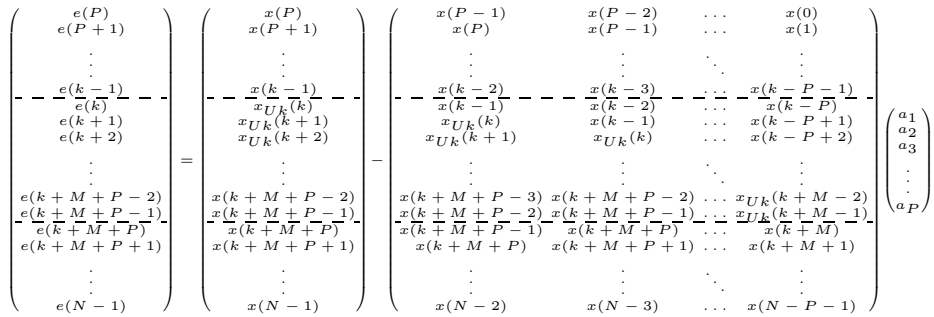


Fig. 1. Matrix form of the residual for the LSAR algorithm

where a_k are the model coefficients and $e(m)$ is a zero mean excitation signal.

Let us assume that a received signal segment $\boldsymbol{x} = (x(0), x(1), ..., x(N-1))^T$ contains a series of lost (unknown) samples $\boldsymbol{x}_{Uk} = (x(k), ..., x(k+M-1))$. The objective is to reconstruct the missing samples $\boldsymbol{x}_{Uk}$ using the remaining known samples and the LP model of the signal (1). Rearranging the LP model and expanding it to a matrix notation we obtain the formulation displayed in Fig. 1, which can be rewritten in a compact notation as,

$$\boldsymbol{e}(\boldsymbol{x}_{Uk}, \boldsymbol{a}) = \boldsymbol{x} - \boldsymbol{X}\boldsymbol{a} \tag{2}$$

The missing samples $\boldsymbol{x}_{Uk}$ are then reconstructed by minimizing the squared error expressed as

$$\varepsilon = \|\boldsymbol{e}\|_2^2 = \boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{x}^T\boldsymbol{x} + \boldsymbol{a}^T\boldsymbol{R}_x\boldsymbol{a} - 2\boldsymbol{a}^T\boldsymbol{r}_x \tag{3}$$

where $\boldsymbol{R}_x = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{r}_x = \boldsymbol{X}^T\boldsymbol{x}$. Note that ε is a function of two unknown variables, the predictor coefficients $\boldsymbol{a}$ and the unknown segment $\boldsymbol{x}_{Uk}$, whose reconstruction is the objective of this problem. Since (3) involves unknown terms of fourth and cubic order, solving the problem by differentiating ε with respect to the unknown vectors $\boldsymbol{x}_{Uk}$ and $\boldsymbol{a}$ would be mathematically impractical. An estimation-maximization (EM) procedure is used instead. First, Eq. (2) is linearized by setting the unknown samples to zero (estimation). This makes the squared error $\boldsymbol{e}$ to be a function of the LP-coefficients $\boldsymbol{a}$ only. The coefficients are then computed by minimizing ε, that is, by solving the usual set of normal equations, which yields,

$$\hat{\boldsymbol{a}} = \boldsymbol{R}_x^{-1}\boldsymbol{r}_x. \tag{4}$$

Finally, the unknown samples are reconstructed using the estimated LP coefficients. This approach can be iterated several times, although in most cases very few iterations are needed.

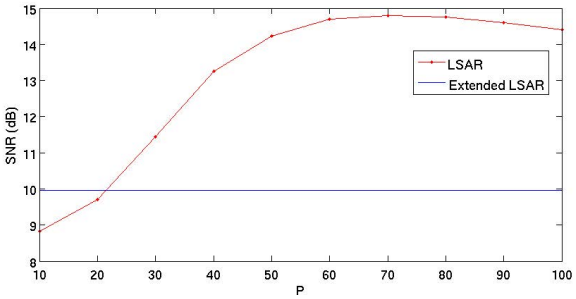


Fig. 2. SNR performance of LSAR (red) and Extended LSAR (blue). The Extended LSAR is applied with P = 10 and Q = 1.

Given that voiced speech signals are quasi-periodic, a speech sample is highly correlated with the neighboring ones as well as with the samples shifted by one (or several) pitch period. In order to exploit these longer correlations, a modification of the basic LSAR (Extended LSAR) which introduces a long-term predictor was proposed in [6]. The speech model involved by the Extended LSAR is,

$$x(m) = \sum_{k=1}^{P} a_k x(m-k) + \sum_{k=-Q}^{Q} p_k x(m-T-k) + e(m), \tag{5}$$

where Q is the order of the long term LP and T is the pitch period. This is the underlying speech model employed somehow by many speech codecs and can be solved again through the corresponding set of normal equations. An interesting feature of this model is that we can consider that equation (5) contains a single predictor with a high level of sparsity. This feature will be exploited in our proposal.

As mentioned in the introduction section, the long-term correlations can be also exploited by the basic LSAR if a prediction order P, large enough to cover the longest possible correlations, is used. The main advantage of this solution is that no pitch estimation is needed.

In order to asses both the basic LSAR and the Extended LSAR, Fig. 2 shows the average SNR values obtained by both techniques for gaps of 6 ms separated 30 ms. The corresponding experimental setup will be described in Section 4. The basic LSAR performance is plotted versus the LP order P, while the extended one is only shown for typical values ($P = 10$, $Q = 1$, 13 coefficients). A first comparison can be made for this typical LP orders. In this case, the Extended LSAR not only outperforms LSAR for $P = 10$, but also for 20 coefficients. This makes clear the need of including long-term correlations. However, it is also observed that the performance can be meaningfully increased with the basic LSAR by simply increasing the LP order. The order increase does not make sense for the Extended LSAR since this would simply imply that the short-term predictor would *absorb* the long-term one.

The main conclusion that can be extracted from the above discussion is that the basic LSAR must be employed if there are no strong computational constraints. However, it still has two problems:

1. Many LP coefficients, which are forced to be zero when the classical speech model is applied, can have now non-zero values. In principle, this may especially affect the inter-pitch and post-pitch coefficients and could be interpreted as a sort of estimation noise.
2. When a large order P is applied, the estimated coefficients can be more affected by the non-stationarity of the speech signal since more autocorrelation coefficients are used in (4).

The effect of these problems over the SNR plot of Fig. 2 is a SNR decay for the higher LP orders. In average, this decay starts after the average pitch value of the speech corpus (57.80 samples).

In this paper, we propose a modification of the LSAR algorithm oriented to mitigate the above problems by applying sparse linear prediction (SLP) for LP predictor estimation. We can consider that this proposal combines the best features of the basic LSAR with large P and Extended LSAR since it uses a single compact predictor which does not require pitch estimation and tries to keep the sparsity of Extended LSAR. SLP and the proposed modification to LSAR are presented in the next section.

3 LSAR by Sparse Linear Prediction

As discussed in the previous section, our goal is the development of a new variant of LSAR with a single large-order predictor which is, at the same time, highly sparse. Thus, we have to minimize the squared error in (3), with respect to $\boldsymbol{a}$, with a sparsity constraint, that is,

$$\begin{aligned} &\text{minimize } \epsilon(\boldsymbol{a}) = \left\|\boldsymbol{a}^T \boldsymbol{R}_x \boldsymbol{a} - 2\boldsymbol{a}^T \boldsymbol{r}_x\right\|_2^2 \\ &\text{subject to } \|\boldsymbol{a}\|_0 \leq \delta_0. \end{aligned} \tag{6}$$

where the term $\boldsymbol{x}^T\boldsymbol{x}$ is not included in the optimization procedure since it comprises the DC component of the squared error. The main problem that arises when solving (6) is that the ℓ_0-norm is non convex so that the global minimum is usually found by exhaustive search and is therefore computationally prohibitive. This problem has been thoroughly studied in compressive sensing theory and can be efficiently solved by applying convex relaxation [10], i.e.

$$\begin{aligned} &\text{minimize } \epsilon(\boldsymbol{a}) = \left\|\boldsymbol{a}^T \boldsymbol{R}_x \boldsymbol{a} - 2\boldsymbol{a}^T \boldsymbol{r}_x\right\|_2^2 \\ &\text{subject to } \|\boldsymbol{a}\|_1 \leq \delta_1. \end{aligned} \tag{7}$$

The objective function, as well as the constraints, are both convex and the optimization problem can be efficiently solved by a convex optimization algorithm. In our simulations, we apply the primal-dual interior point (IP) method [11].

The LP-coefficients obtained in the previous step are then used to re-estimate the unknown samples $\boldsymbol{x}_{Uk}$. This is carried out by inserting the obtained coefficients $\boldsymbol{a}$ into Eq.(2) and minimizing the squared error ε with respect to $\boldsymbol{x}_{Uk}$, which is the only unknown variable the squared error now depends on. Note that only the equations within the dashed lines in Fig. 1 are involved in the minimization since the remaining ones are constant with respect to $\boldsymbol{x}_{Uk}$. These equations can be rearranged so that the excitation signal is a combination of known and unknown samples:

$$\boldsymbol{e} = \boldsymbol{A}_1\boldsymbol{x}_{Uk} + \boldsymbol{A}_2\boldsymbol{x}_{Kn} \tag{8}$$

where the matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are both constructed using the LP-coefficients $\boldsymbol{a}$ and $\boldsymbol{x}_{Kn} = (x(k-P), \ldots, x(k+M+P-1))^T$ (see ref. [2] for more details). The total squared error is then given by,

$$\|\boldsymbol{e}\|_2^2 = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{A}_1\boldsymbol{x}_{Uk} + \boldsymbol{A}_2\boldsymbol{x}_{Kn})^T(\boldsymbol{A}_1\boldsymbol{x}_{Uk} + \boldsymbol{A}_2\boldsymbol{x}_{Kn}) \tag{9}$$

The unknown samples $\boldsymbol{x}_{Uk}$ that minimize the squared error are obtained by setting the derivative of the squared error function with respect to $\boldsymbol{x}_{Uk}$ to zero

$$\frac{\partial \boldsymbol{e}^T\boldsymbol{e}}{\partial \boldsymbol{x}_{Uk}} = 2\boldsymbol{A}_1^T\boldsymbol{A}_1\boldsymbol{x}_{Kn} + 2\boldsymbol{A}_1^T\boldsymbol{A}_2\boldsymbol{x}_{Kn} \tag{10}$$

Finally, from Eq.(10) we have

$$\hat{\boldsymbol{x}}_{Uk} = -(\boldsymbol{A}_1^T\boldsymbol{A}_1)^{-1}(\boldsymbol{A}_1^T\boldsymbol{A}_2)\boldsymbol{x}_{Kn} \tag{11}$$

The sparsity restriction does not make sense in this case since $\boldsymbol{x}_{Uk}$ is not sparse in general, although it can be solved via convex optimization with no restrictions.

In order to better illustrate the differences between the sparse approach and the classic LSAR, let us analyze two particular cases of missing segment reconstruction. The first case involves a voiced segment with pitch period equal to 32 samples. The pitch period is calculated over the clean (original) signal using the Yin pitch detector [12]. In the second case, an unvoiced segment is reconstructed. For both cases, we perform a reconstruction with 100 LP-coefficients using LSAR and the proposed technique. The results are shown in Fig. 3. Figure 3(a) shows the obtained coefficients for the voiced segment. As expected, the SLP-coefficients are much sparser than the coefficients obtained by LSAR while providing a reconstruction with lower squared error. Moreover, the significant elements of the LP-vector are concentrated around the position of 32, 64 and a small contribution around 96. Note that the pitch period of the original signal has been determined to be 32. Thus, the proposed SLP predictor adaptively encounters the pitch value. The case of the unvoiced segment reconstruction is shown in Fig. 3(b). Again, the LSAR coefficients vector is much less sparse while generating a reconstruction with larger squared error. In this case, SLP automatically concentrates the weights in the close proximity of the lost segment which is coherent with the assumption that in unvoiced segments the most correlated samples are the closest ones.

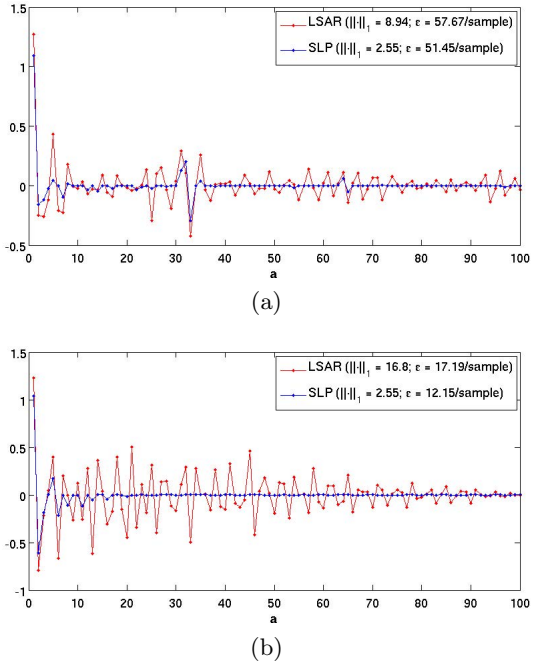
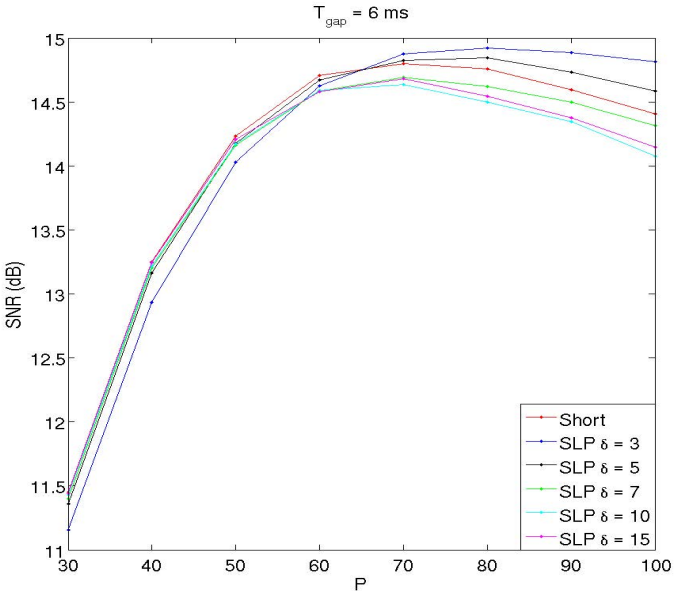


Fig. 3. Example of distribution of the LP-coefficients obtained by LSAR (red) and SLP (blue). (a) Voiced segment. (b) Unvoiced segment.

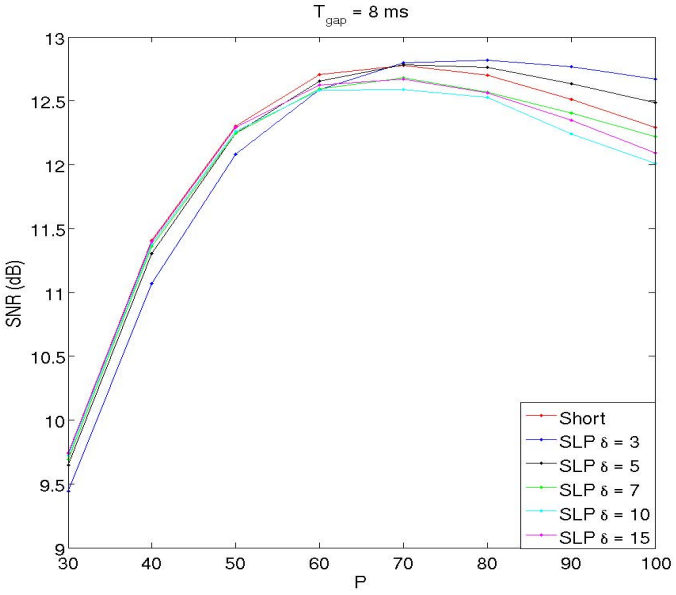
4 Simulation Results

The speech utterances used for testing comprise a subset of 400 sequences extracted from the geographic corpus of the *Albayzin* database [13]. All the speakers, used for recording the database, are also present in our tested subset. Figure 4 shows a comparison of LSAR and the proposed SLP-based algorithm in terms of SNR. The average SNR value over the 400 utterances is shown for different LP orders. The proposed SLP technique is also tested for different levels of sparsity (controlled with parameter δ_1). Moreover, the comparison is carried out for missing segment lengths (T_{gap}) of 6 ms and 8 ms. The losses are produced every 30 ms and a 32 ms window, centered over the missing segment, is used for estimating the predictor. The window is eventually extended (up to the required length) in cases where the sum of LP order and the duration of the gap is larger than the 32 ms window. Finally, two iterations are employed for both cases.

The simulations reveal that the larger the LP order, the sparser it should be in order to obtain better quality reconstructions. The average pitch period of the tested subset is 57,80 so, for shorter LP orders, there is no need to impose sparsity over the LP estimator. Note that weak sparsity restrictions approximate the LSAR behavior for low order LP estimators. For LP orders above the average pitch period, the performance of the LSAR technique starts to decay while



(a)



(b)

Fig. 4. Performance comparison, in terms of SNR, of LSAR and SLP with different values of δ. Both algorithms are tested for different values of LP order and two iterations are applied. (a) $T_{gap} = 6$ ms. (b) $T_{gap} = 8$ ms.

our proposal eventually rises and then practically maintains the reconstruction quality (with a very slight decay). This results confirm that the basic LSAR with large prediction order may be affected by noise estimation and non-stationarity and that the sparsity constraint helps to palliate these problems. Also, we can conclude that the sparsity parameter δ_1 could be set according to the LP order, although in this paper we focused on obtaining a fixed estimator suitable for the majority of pitch periods. Thus, a high order and highly sparse (small δ_1) LP estimator is preferred.

Table 1 shows the average values of SNR and PESQ (Perceptual Evaluation of Speech Quality) obtained for different lengths of lost segments. The LP order is set to 100 in order to include all possible pitch values in the database and 5 iterations are used. The proposed method outperforms the basic LSAR in all cases and the difference in perceptual quality has an increasing trend with the gap length.

Table 1. Average SNR and PESQ values for SLP and LSAR for different lost segment lengths. The simulation is carried out for P = 100 with 5 iterations.

SNR	T_{gap} = 4ms	T_{gap} = 6ms	T_{gap} = 8ms	T_{gap} = 10ms
SLP	18.10	15.13	13.03	11.38
LSAR	17.47	14.41	12.34	10.67
PESQ				
SLP	4.00	3.81	3.63	3.47
LSAR	3.91	3.72	3.51	3.34

5 Conclusions

We have proposed a modification of the LSAR speech reconstruction algorithm which uses sparse linear prediction. The proposed approach has several advantages as avoiding the use of pitch detectors, a better approximation to the sparse classical model employed in speech coding, a better behavior for large pitch values (reducing the estimation noise) and less sensitivity to non-stationarities. Applying convex relaxation allows to solve the minimization problem with sparsity constraint in a relatively efficient way. The proposed technique outperforms the classic LSAR both at objective and perceptual level. Future work includes the dynamic adaptation of the sparsity parameter δ_1 to the instantaneous pitch values and the LP order.

References

1. Vaseghi, S.: Multimedia signal processing. John Wiley (2007)
2. Janssen, A., Veldhuis, R., Vries, L.: Adaptive interpolation of discrete-time signals that can be modeled as AR processes. IEEE Transactions on Acoustics, Speech and Signal Processing, 317–330 (1986)

3. Jauppinen, I., Roth, K.: Audio signal restoration - theory and applications. In: Proceedings of the 5th Int. Conf. on Digital Audio Effects (2002)
4. Esquef, P., Biscainho, L.: An efficient model-based multirate method for reconstruction of audio signals along long gaps. IEEE Transactions on Speech and Audio Processing 14, 1391–1400 (2006)
5. Lindblom, J., Hedelin, P.: Packet loss concealment based on sinusoidal modelling. In: Proceedings of ICASSP 2002 (2002)
6. Vaseghi, S., Rayner, P.: Detection and suppression of impulsive noise in speech communication systems. IEE Proceedings 1, 38–46 (1990)
7. Giacobello, D., Christensen, M., Dahl, J., Jensen, S., Moonen, M.: Sparse linear prediction of speech. In: Proceedings of Interspeech 2008 (2008)
8. Giacobello, D., Christensen, M., Murthi, M., Jensen, S., Moonen, M.: Speech coding based on sparse linear prediction. In: Proceedings of Eusipco 2009 (2009)
9. Koloda, J., Østergaard, J., Jensen, S., Peinado, A., Sanchez, V.: Sequential error concealment of video/images via weighted template matching. In: Proceedings of DCC 2012 (2012)
10. Romberg, J.: Imaging via compressive sensing. IEEE Signal Processing Magazine 25 (March 2008)
11. Vandenberghe, L., Boyd, S.: Semidefinite programming. Society for Industrial and Applied Mathematics (1996)
12. Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America 111(4), 1917–1930 (2002)
13. Díaz-Verdejo, J., Peinado, A., Rubio, A., Segarra, E., Prieto, N., Casacuberta, F.: ALBAYZIN: a task-oriented spanish speech corpus. In: First International Conference on Language Resources and Evaluation, vol. 1, pp. 487–502 (May 1998)

Steganographic Pulse-Based Recovery for Robust ACELP Transmission over Erasure Channels

Domingo López-Oller[1], Angel M. Gomez[1], José Luis Pérez Córdoba[1], Bernd Geiser[2], and Peter Vary[2]

[1] Departamento de Teoría de la Señal, Telemática y Comunicaciones
University of Granada, Spain
{domingolopez,amgg,jlpc}@ugr.es
[2] Institute of Communication Systems and Data Processing
RWTH Aachen University, Germany
{geiser,vary}@ind.rwth-aachen.de

Abstract. This paper presents an ACELP-based speech transmission scheme that is robust to frame erasures. The scheme is based on the steganographic transmission of media-specific FEC codes. These FEC codes are intended to prevent the adaptive codebook desynchronization frequently found in the decoder after a frame erasure. They are based on a multipulse representation of the previous frame excitation. By means of steganographic methods, the FEC codes are embedded into the codec bitstream, thus causing no bit rate increase. In particular, an ACELP-specfic steganography approach exploits the inefficiencies in the ACELP codebook search and imposes certain algebraic restrictions which allow the hiding of data in the ACELP codewords. Effectively, side information can be transmitted without compromising the codec speech quality. The performance of our proposal is evaluated with the well-known AMR ACELP codec, both in terms of speech quality and intelligibility. To this end, objective measures, i.e. PESQ and STOI, are applied. The proposed coding scheme achieves a noticeable improvement over the legacy codec under adverse channel conditions without consuming any additional bit rate.

Keywords: ACELP, speech coding, data hiding, steganography, multipulse, frame erasure.

1 Introduction

Modern speech codecs are based on the CELP [1] paradigm that provides a high-quality synthesis at a remarkably low bit-rate. Nevertheless, due to the extensive use of predictive filtering, CELP codecs are relatively vulnerable to the frame erasures which frequently appear in packet-based transmissions. One of these filters, the long-term prediction (LTP) filter, which is used to build up the adaptive codebook (ACB), can exceed frame boundaries and it is primarily responsible for undesired error propagation effects, cf. [2–4]. Error propagation occurs when

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 257–266, 2012.

the excitation samples within a lost speech frame are synthetically replaced with the help of a concealment algorithm. Such erroneous excitation patterns cause a desynchronization of the ACB at the decoder. Until synchronicity is regained, degradations in the synthesized speech can propagate over several frames [3, 4]. Indeed, error propagation is currently being considered as a notable source of degradation and several authors have proposed a number of techniques to minimize or even avoid it in packet-switched telephony, e.g. [5–8]

In [3], we have proposed a novel FEC technique to alleviate such propagation error problems. This technique was derived from a multipulse encoding approach and consists of the computation and transmission of a pulse-based representation of the previous frame excitation. Hence, an alternative representation of ACB is available after a frame loss. Nevertheless, those pulses (position-amplitude pairs) are transmitted in the form of media-specific FEC bits, causing a small but unavoidable increase of the transmitted bit rate. The resulting bitstream format is therefore incompatible with the original codec standard.

In order to avoid this overhead and eventual incompatibilities, we propose to use data hiding or steganography to embed the FEC information into the bitstream, allowing a similar protection to that achieved with FEC codecs but without any bit rate increase. Bitstream data hiding has typically been performed for various media coding schemes such as JPEG [9], H.264 [10] or MPEG2 [11] with different purposes. Regarding speech transmission, the principles of CELP-oriented data hiding were proposed by Z.M. Lu et al. [12] where a rather low steganographic capacity of 37 bit/s was achieved. This capacity was further extended by the steganographic ACELP codec proposed in [13] allowing bit rates of several 100 bit/s without compromising the quality of the coded speech signal. Here, we will exploit a recently proposed data hiding technique for ACELP codecs [14, 15], which allows to embed various bit rates from 200 bit/s up to 2 kbit/s.

The remainder of this paper is organized as follows. Sections 2 and 3 explain the fundamentals of the FEC codes based on multipulse resynchronization and the steganographic method used to embed them into the ACELP bitstream, respectively. In Section 4 we present our experimental framework and results. Finally, in Section 5 conclusions are drawn.

2 Basic Multipulse Approach for Propagation Error Recovery

Modern speech codecs are based on the linear prediction model. Under this model, speech is obtained by filtering an excitation signal, $e(n)$, by means of the inverse short-term linear prediction (LP) filter, i.e. with the system function $1/A(z)$. In CELP codecs, the excitation $e(n)$ is the result of a long-term prediction (LTP) filter which applied over a residual signal, also known as the code vector, $e_c(n)$. Formally, the excitation signal, $e(n)$, is obtained as:

$$e(n) = g_a \sum_{k=-(q-1)/2}^{(q+1)/2} p_k e(n-(T+k)) + g_c e_c(n) = g_a e_a(n) + g_c e_c(n), \qquad (1)$$

where T, p_k, g_a and g_c are the parameters of the LTP filter (q is the prediction order), namely lag-delay, long-term coefficients, and adaptive and code gains, respectively. Those parameters are chosen in order to minimize the (spectrally weighted) error between the synthesized speech signal and the original one. Alternatively, the excitation signal can be seen as a weighted sum of two different components, $e_c(n)$, chose from a fixed codebook, that represents the residual signal remaining after removing the long-term redundancy, and $e_a(n)$ an adaptive codebook (ACB), that models the long-terms correlations in excitation.

When a frame erasure happens, a concealment algorithm tries to minimize the degradation on the perceptual quality by extrapolating and gradually muting (in case of consecutive lost frames) the speech signal. The excitation corresponding to this concealment is used by the LTP filter to compute the excitation in the next frame. Since the concealed signal is not identical to the transmitted one, the decoder gets desynchronized from the encoder and a distortion appears that can propagate over several subsequent correctly received frames.

In order to alleviate this problem we proposed a media-specific FEC coding scheme to combat the aforementioned error propagation effect. This FEC code represents the previous excitation as in multipulse coding [16], that is, by a few pulses with different amplitudes and positions. Thus, after a frame loss, this alternative excitation is used instead of that provided by the previous packet loss concealment algorithm in order to prevent the LTP desynchronization.

The proposed FEC is computed by considering the previous frame samples as a memory where some pulses can be set. By means of a procedure derived from multipulse coding, the position and amplitude of the pulses are optimized in order to minimize the least square error (LSE) between the synthesized signal and the original one [4]:

$$\varepsilon = \sum_{n=0}^{N-1} \left(s(n) - \hat{s}(n)\right)^2 = \sum_{n=0}^{N-1} \left(s(n) - h(n) * \hat{e}(n)\right)^2 \tag{2}$$

$$\text{with} \qquad \hat{e}(n) = f(T, g_a, g_c, e_c(n), b_0, p_0, b_1, p_1, ..., b_{L-1}, p_{L-1}, n), \tag{3}$$

where $h(n)$ is the impulse response of the LP filter, $s(n)$ is the target signal and N the number of samples per frame. It can be shown [3, 4] that contributions from perceptual filtering (commonly used in CELP coding) and code vector (i.e. g_c and $e_c(n)$) can be removed from the target vector $s(n)$, easing the computation of the pulses. Here, only the perceptual filter contribution is considered while the code vector one will be neglected. This is because the algebraic code remains unknown until steganography has been applied, see Section 3.

2.1 Pulse Computation and Coding

In many CELP-based codecs, frames are split into subframes applying different filters in each one so the resulting coefficients are variable in time [17, 18]. Due to this, for a complete frame we deal with a linear but time variant system. Given a unit pulse at position p producing an output $h_{LT}(n, p)$, the same pulse

with amplitude b will provide the same output but scaled by b. Therefore, we can relate the excitation with the sum for every L pulse contribution as:

$$\hat{e}(n) = \sum_{l=0}^{L-1} b_l h_{LT}(n, p_l) \tag{4}$$

where we will refer to $h_{LT}(n, p_l)$ as shape signals that are obtained by placing a unit pulse at position p_l in the previous frame (filter memory) and LTP filtering the current frame with a zero excitation. Here we are assuming that $e(n)$ only depends on LTP filter, i.e. the adaptive gain and pulses, as the perceptual filter is removed from the target signal and the code vector contribution is neglected as previously mentioned. Then, it can be proved [3] that the synthesized signal is given by:

$$\hat{s}(n) = \sum_{l=0}^{L-1} b_l \cdot g_{p_l}(n) \tag{5}$$

$$\text{with} \qquad g_p(n) = h(n) * h_{LT}(n, p)$$

where $g_p(n)$ is the LP response to the shape signal at position p (in the LTP filter memory) and can also be computed on a subframe basis so we can cope with the existence of different LP sets for each subframe.

It can be proved [4] that, provided a set of pulse positions p_k, optimal amplitudes b_k are obtained by means of the following set of equations:

$$\sum_{k=0}^{L-1} b_k^* \phi_{p_k p_j} = c_{pj} \tag{6}$$

where

$$\phi_{p_k p_j} = \mathbf{\Phi}[k, j] = \sum_{n=0}^{N-1} g_{p_k(n) g_{p_j}(n)}$$

$$c_{p_j} = c[j] = \sum_{n=0}^{N-1} s(n) g_{p_j}(n) \tag{7}$$

Since the LSE can be also obtained from $s(n)$ and $g_p(n)$, pulse positions could be found by choosing that combination which provides the lowest error (with optimal amplitudes). However, this solution is impracticable due to the high number of possible combinations of L pulses in the LTP filter memory. In order to reduce complexity, we can consider only a single pulse, as in [4], so the set of equations (6) reduces to:

$$b^* \cdot \phi_{pp} = c_p \tag{8}$$

In addition, the LSE between the original signal and the synthesized one, obtained from a single pulse at position p with optimum amplitude b^*, is given as [4]:

$$\varepsilon^* = \sum_{n=0}^{N-1} s^2(n) + b^{*2}\phi_{pp} - 2b^* c_p = \sum_{n=0}^{N-1} s^2(n) - c_p^2/\phi_{pp} \tag{9}$$

As we can see, we only require the diagonal elements of Φ ($\phi_{ii} = \Phi[i,i]$), and the position which provides the optimal pulse-position i^* and its corresponding optimal amplitude b_i^* are finally obtained as:

$$i^* = \underset{i}{\operatorname{argmax}}\,(c_i^2/\phi_{ii}), \qquad b_i^* = c_{i^*}/\phi_{i^*i^*} \tag{10}$$

After obtaining the position and amplitude, we can jointly quantize them in order to be transmitted as a FEC. This pulse is used in the decoder to regenerate the excitation when the previous frame is lost.

3 Steganography for ACELP Codecs

In CELP coders, steganography is generally realized by embedding the steganographic bits in the less important parts of the encoded bitstream, i.e., for the AMR codec, in the fixed codebook (FCB) contribution. For the application of data hiding to CELP coders, it turns out to be advantageous to integrate the watermarking procedure into the analysis-by-synthesis loop for the fixed codebook (FCB) search. The embedding of K steganographic bits per frame is achieved by partitioning the fixed excitation code book $\mathcal{C}$ into $M = 2^K$ disjoint subcodebooks $\mathcal{C}_m$. If we consider that m is the message to be embedded, $c \in \mathcal{C}_m$ are the examined candidate codevectors, and $\mathcal{X}(c)$ is the CELP cost function, the FCB search with information-embedding can be formulated as [14]:

$$\mathcal{X}(c) = ||v||^2 - \frac{(v^T Hc)^2}{||Hc||^2}$$

$$\hat{c} = \underset{c \in \mathcal{C}_m}{\operatorname{argmin}}\,\mathcal{X}(c) \tag{11}$$

with the target vector v (pitch removed prediction residual) and the perceptually weighted filter matrix H. The hidden message is decoded by identifying the subcodebook that contains the received vector $\hat{c}$:

$$m = m' \;:\; \hat{c} \cap \mathcal{C}_{m'} = \hat{c} \tag{12}$$

Considering the described embedding scheme, the number of examined FCB entries is decreased by a factor of M for each frame, and the inevitable consequence would be a decreased quality of the coded speech. So in order to maintain the speech quality, a joint implementation of the speech encoding and data hiding operations must be used [19].

The key to this ACELP steganographic technique is a modified search strategy for the ACELP codebook since typically only a small heuristically selected subset

$\mathcal{C}' \subset \mathcal{C}$ is examined during FCB search. If we introduce additional FCB entries in every sub-codebook $\mathcal{C}_m$, i.e. $|\mathcal{C}_m| \geq |\mathcal{C}'|$, M "equally good" sub-codebooks can be established and the data hiding procedure will not degrade the resulting speech quality. In this paper we will use the implementation in [14], which modifies the FCB search procedure for the 12.2 kbit/s mode of the AMR codec. The alternative steganographic search strategy allows different hidden data rate from 200 bit/s to 2 kbit/s. For our purposes, 4 steganographic bits are transmitted per 5-ms subframe, i.e. the obtained steganographic bit rate is 4 bits/5 ms=800 bit/s. Therefore, a concrete message m is defined as a 4 bit binary sequence whose individual bits are denoted by, $(m)_k$ with $k = 0, \ldots, 3$. To enable the transmission of $K = 4$ steganographic bits per subframe, the FCB is partitioned into $M = 2^K$ sub-codebooks that uniquely identify the selected message m. Based on the standard ACELP search method from [17], the proposed steganographic algorithm has been derived in two steps:

1. **Codebook Partitioning.** First, M disjoint sub-codebooks are established by appropriately restricting the set of admissible codevectors. In particular, a specific parity condition is imposed on certain parts of the AMR bitstream:

$$(m)_k = \left[\mathcal{G}\left(\lfloor \frac{i_k}{5} \rfloor \right) \oplus \mathcal{G}(\lfloor \frac{i_{k+5}}{5} \rfloor) \right] \mod 2 \tag{13}$$

   for the ACELP pulse positions i_k with $k \in \{0, ..., 3\}$, where $X \oplus Y$ is the bitwise exclusive disjunction (XOR) of two binary strings and $\mathcal{G}$ represents the standardized Gray encoding of the ACELP pulse position codewords. At the decoder, the hidden information can be retrieved directly from the AMR bitstream using (13). Solving the above bitstream parity condition for the position i_{k+5} of the *second* pulse in ACELP track k, the admissible indices (and thus the possible positions) for this pulse can be computed:

$$\lfloor \frac{i_{k+5}}{5} \rfloor = \mathcal{G}^{-1}(\mathcal{G}(\lfloor \frac{i_k}{5} \rfloor) \oplus (m)_k + 2 \cdot j) \tag{14}$$

   with $j \in \{0, \ldots, 3\}$. Hence, the first four (out of five) pulse tracks are restricted to have four (out of eight) admissible pulse positions. The fifth pulse track is not restricted here.
2. **Search Space Expansion**. Based on the chosen codebook partitioning, an FCB search strategy is devised that provides a good trade-off between speech quality and computational complexity. Thereby, the admissible values for the pulse positions i_{k+5} can be computed using (14). More details on this steganographic FCB search can be found in [14].

4 Experimental Framework and Results

In order to evaluate the performance of our proposal we use the PESQ (Perceptual Evaluation of Speech Quality) algorithm [20] and the STOI (Short-Term Objetive Intelligibility) [21] test with the AMR 12.2 kbit/s standard speech codec. The

speech corpus is a subset of TIMIT database [22] downsampled at 8 kHz, composed by a total of 1328 sentences uttered by a balanced number of male and female speakers. The scores obtained for every test sentence are weighted by their relative length in the overall score. Finally, frame erasures are simulated by a random packet loss model with a frame per packet, that provides 9 channel conditions with packet loss rates from 0% to 23%. Although this approach is known to be insufficient in modelling realistic packet loss scenarios in packet-switched networks, it has been used in this work as in many other papers [2–5, 7], since we are interested in the error propagation after a frame erasure, and not in the loss itself.

In this paper we have tested three different schemes (see Fig. 1), the AMR codec at 12.2 kbit/s mode, as baseline (AMR), the same AMR codec but with additional FEC information comprising a recovery pulse as described in Section 2, and the steganographic codec embedding the same FEC information into the bitstream. We encode the recovery pulse with 10 bits, which results in a bit rate increase of 500 bits/s. The steganographic transmission with 800 bits/s therefore leaves some headroom, e.g. for additional error protection.

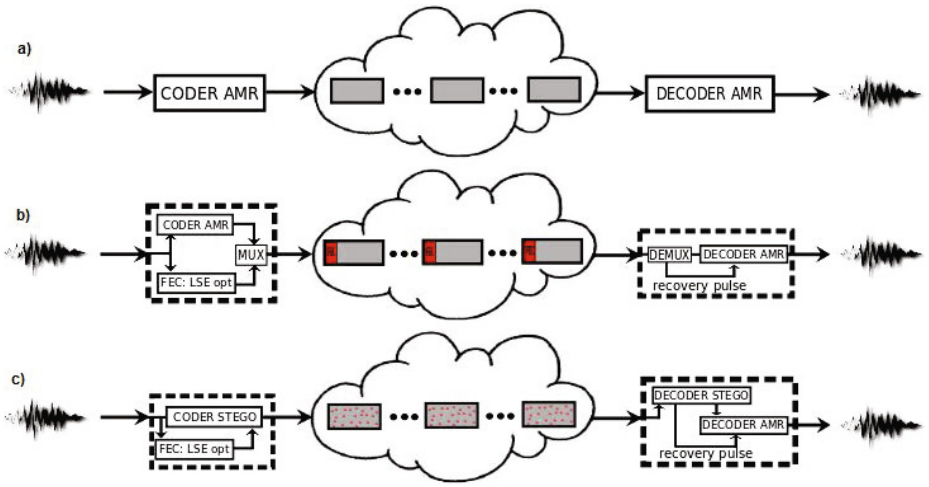


Fig. 1. Schemes of transmission for the implemented experiments: (a) AMR legacy codec (AMR), (b) AMR codec plus a FEC code (AMR+FEC) and (c) AMR codec using data hiding for the FEC code (STEGO)

4.1 Speech Quality

In this paper, an objective quality measurement tool has been selected in order to exhaustively evaluate the proposed schemes. The ITU PESQ algorithm [20] is one of the most popular algorithms for speech quality assessment. It is applied over each utterance in the testing database, providing a score about the speech signal quality within a range from 0.5 (bad) to 4.5 (excellent). Table 1 shows the results obtained by the PESQ test for the considered channel conditions.

Table 1. Average PESQ scores obtained for AMR 12.2 kbit/s (AMR) codec and the multipulse (AMR+FEC) and steganographic (STEGO) approaches under different channel conditions

	bit rate	Frame erasure ratio								
		0%	4%	7%	10%	13%	16%	18%	21%	23%
AMR	12.2	4.003	3.212	2.936	2.690	2.531	2.321	2.200	2.108	2.024
AMR+FEC	12.75	4.003	3.417	3.193	2.997	2.878	2.720	2.626	2.556	2.497
STEGO	12.2	3.991	3.398	3.170	2.974	2.852	2.693	2.597	2.527	2.466

We can observe that, in case of frame losses, the STEGO technique achieves a better performance than the BASE AMR. Therefore, the use of steganography in case of packet loss, using the multipulse technique as FEC information allows an improved performance compared with the results of BASE AMR, with the advantage in the STEGO technique of not increasing the bit rate and without incurring in a significant performance loss in comparison with the AMR+FEC technique (average difference 0.026).

4.2 Speech Intelligibility

Improving quality might not necessarily lead to improvements in terms of intelligibility. In fact, it is well known that some speech processing algorithms which achieve a significant improvement in quality might be accompanied by a decrease in intelligibility [23]. Due to this, it is noteworthy to also perform an intelligibility test to our techniques. To this end, we will use an objective metric, the Short-Time Objective Intelligibility metric (STOI) [24]. This technique has been recently proposed for speech intelligibility assessment and it is based on the linear correlation between a time-frequency representation of clean and damaged speech over time frames. STOI provides a very high correlation (> 0.9) with intelligibility scores provided by human listeners [24, 25], allowing us to make reasonable conclusions.

We have applied this algorithm over each utterance in the testing database, providing a score about the speech signal quality within a range from 0 (unintelligible) to 1 (fully intelligible). Table 2 shows the results obtained by STOI tests for the considered analyzed channel conditions.

Table 2. STOI scores obtained for base AMR 12.2 kbit/s (AMR) codec and the multipulse (AMR+FEC) and steganographic (STEGO) under different channel conditions

	bit rate	Frame erasure ratio								
		0	4%	7%	10%	13%	16%	18%	21%	23%
AMR	12.2	0.908	0.862	0.829	0.796	0.769	0.731	0.707	0.687	0.667
AMR+FEC	12.75	0.913	0.891	0.874	0.857	0.843	0.824	0.810	0.800	0.789
STEGO	12.2	0.914	0.888	0.871	0.855	0.841	0.822	0.808	0.798	0.788

We can observe that STEGO technique also achieves a better performance in terms of intelligibility than BASE AMR experiment. From the obtained results we can draw similar conclusions as from the previous PESQ test.

5 Conclusions

In this paper we have presented a robust speech transmission scheme that combines an steganographic technique for ACELP codecs with media-specific FEC codes. The aim is to reduce detrimental error propagation effects in CELP-based speech codecs. FEC codes are based on a multipulse representation of the previous excitation frame, allowing to retrieve the ACB codebook synchronicity after a frame erasure. On the other hand, steganography enables this data overhead to be hidden within the algebraic code, causing no bit rate increase and maintaining full bitstream compatibility with the codec standard.

Two objective metrics have been used to test the performance of the proposed scheme, one referred to speech quality, the PESQ algorithm, and the other referred to the speech intelligibility, the STOI measure. Both metrics confirm a significant increase in terms of robustness in adverse channel conditions with frame losses. Similar scores are obtained by the non-steganographic FEC code transmission and by the steganographic transmission, despite the fact htat the latter does not take the code vector contribution into account, as the algebraic code is modified by the steganographic procedure itself (cf. Section 2). In addition, the proposed scheme achieves almost identical performance to that offered by the legacy codec in clean channel conditions.

References

1. Shroeder, M., Atal, B.: Code-excited linear prediction (celp): high-quality speech at very low bit rates. In: IEEE International Conference on Acoustics, Speech and Signal Proccessing, vol. 10 (1985)
2. Carmona, J., Gomez, A., Peinado, A., Perez-Cordoba, J., Gonzalez, J.: A multipulse fec scheme based on amplitude estimation for celp codecs over erasure channels. In: INTERSPEECH (2010)
3. Gomez, A., Carmona, J., Peinado, A., Sanchez, V.: A multipulse-based forward error correction technique for robust celp-coded speech transmission over erasure channels. IEEE Trans. Audio Speech Lang. Process. (2010)
4. Gomez, A., Carmona, J., Gonzalez, J., Sanchez, V.: One-pulse fec coding for robust celp-coded speech transmission over erasure channels. IEEE Trans. Multimedia (2011)
5. Serizawa, M., Ito, H.: A packet loss recovery method using packet arrived behind the playout time for celp decoding. In: IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 1, pp. 169–172 (2002)
6. Carmona, J.L., Perez-Cordoba, J.L., Peinado, A.M., Gomez, A.M., Gonzalez, J.A.: A scalable coding scheme based on interframe dependency limitation. In: IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 4, pp. 4805–4808 (April 2008)

7. Vaillancourt, T., Jelinek, M., Salami, R., Lefebvre, R.: Efficient frame erasure concealment in predictive speech codecs using glottal pulse resynchronisation. In: IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 4, pp. 1113–1116 (April 2007)
8. Lamel, L.F., Kassel, R.H., Seneff, S.: Speech database development: Design and analysis of the acoustic-phonetic corpus. In: Speech Recognition Workshop (DARPA), pp. 100–110 (February 1986)
9. Wong, P., Au, O.: A blind watermarking technique in jpeg compressed domain. In: ICIP (September 2002)
10. Noorkami, M., Mersereau, R.: Compressed-domain video watermarking for h.264. In: ICIP (September 2005)
11. Siebenhaar, F., Neubauer, C., Herre, J.: Combined compression/watermarking for audio signals. In: 110th Conv. of the AES (May 2001)
12. Lu, Z., Yan, B., Sun, S.: Watermarking combined with celp speech coding for authentication. IEICET Trans. on Inf. and Systems E88-D(2), 330–344 (2005)
13. Geiser, B., Vary, P.: Backwards compatible wideband telephony in mobile networks: Celp watermarking and bandwidth extension. In: ICASSP (April 2007)
14. Geiser, B., Vary, P.: High rate data hiding in acelp speech codecs. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2008)
15. Vary, P., Geiser, B.: Steganographic wideband telephony using narrowband speech codecs. In: Conference Record of Asilomar Conference on Signals, Systems, and Computers (November 2007)
16. Atal, B., Remde, J.: A new model of lpc excitation for producing natural-sounding speech at low bit rates. In: IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 7, pp. 614–617 (May 1982)
17. 3GPP TS 26.090, Mandatory speech codec speech processing functions; adaptive multi-rate (amr) speech codec (1999)
18. ITU-T G.729 Recommendation, Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (cs-acelp) (1993)
19. Geiser, B., Mertz, F., Vary, P.: Steganographic packet loss concealment for wireless voip. In: Conference on Voice Communication (SprachKommunikation), pp. 1–4 (October 2008)
20. ITU-T Recomendation P.862, Perceptual evaluation of speech quality (pesq) (2001)
21. Tang, Y., Cooke, M.: Subjetctive and objetive evaluation of speech intelligibility enhancement under constant energy and duration constraints. In: INTERSPEECH (2011)
22. Garofolo, J.S.: The structure and format of the darpa timit cd-rom prototype
23. Goldsworthy, R., Greenberg, J.: Analysis of speech-based speech transmission index methods with implications in nonlinear operations. J. Acoust. Soc. Am. (116), 3679–3689 (2004)
24. Taal, C., Hendriks, R., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. on Audio, Speech, and Language Processing 19, 2125–2136 (2011)
25. Gomez, A., Schwerin, B., Paliwal, K.: Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio. Speech Communication (54), 503–515 (2012)

A Proposal for a Visual Speech Animation System for European Portuguese

José Serra[1,2], Manuel Ribeiro[3,4], João Freitas[3,4], Verónica Orvalho[1,2], and Miguel Sales Dias[3,4]

[1] Instituto de Telecomunicações, Porto, Portugal
[2] Department of Computer Science, Faculty of Science, University of Porto, Porto, Portugal
{jserra,veronica.orvalho}@dcc.fc.up.pt
[3] Microsoft Language Development Center, Lisbon, Portugal
{t-manrib,i-joaof,miguel.dias}@microsoft.com
[4] ISCTE-Lisbon University Institute/ADETTI-IUL, Lisboa, Portugal

Abstract. Visual speech animation, or lip synchronization, is the process of matching speech with the lip movements of a virtual character. It is a challenging task because all articulatory movements must be controlled and synchronized with the audio signal. Existing language-independent systems usually require fine tuning by an artist to avoid artefacts appearing in the animation. In this paper, we present a modular visual speech animation framework aimed at speeding up and easing the visual speech animation process as compared with traditional techniques. We demonstrate the potential of the framework by developing the first automatic visual speech automation system for European Portuguese based on the concatenation of visemes. We also present the results of a preliminary evaluation that was carried out to assess the quality of two different phoneme-to-viseme mappings devised for the language.

Keywords: visual speech animation, phoneme-to-viseme mapping, European Portuguese, virtual characters.

1 Introduction

Speech is the most natural way of conveying the ideas and thoughts of the personality of a virtual character. However, speech communication is not only composed of sounds but also of the corresponding articulatory movements and facial expressions. These poses and expressions have an important impact on the naturalness and believability of virtual characters. If speech animation is not done well, i.e. if the facial movements of the virtual character are not human-like or if the synchronization of lip movements with the audio is poor, viewers will find the animation awkward, even if they are not able to pinpoint the source of the problem.

Speech is commonly represented as a sequence of discrete sounds, or phones ('beads-on-a-string') [1]. Each phone and its abstract definition (phoneme) can be

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 267–276, 2012.

associated with a viseme, i.e. the position and orientation of the visible part of the vocal tract articulators comprising the lips, teeth, jaw, tongue and cheeks. All articulators can influence the production of a given phone but not all changes are visible; therefore different phones may be associated with the same viseme. In computer animation, when manually animating speech events, digital artists have to create each viseme by hand. Later, they can concatenate the visemes according to the utterances they want to animate, using an interpolation scheme. Thus, manual speech animation is time-consuming and tedious. As a result, several automatic approaches have been proposed for synchronizing the audio with the visemes and for modeling co-articulation [2].

Visual speech animation can be divided into two main areas according to the way the speech input and the articulatory movements are mapped to each other: (i) phoneme-to-viseme mapping and (ii) sub-phonetic mapping. In the first case, the phonemes are obtained using text or audio analysis, mapped to visemes and organized in a timeline. The actual mapping between phonemes and visemes is important for the end result; if it is not good, the animation can appear exaggerated or have unexpected visual effects. However, a good mapping is not sufficient for high-quality speech animation, and techniques relying on diphones and triphones [3, 4] have been proposed for solving the co-articulation problem – at the cost of larger visual speech databases. Another common technique to tackle this problem involves creating a model for simulating the co-articulation effect [2, 5]. Sub-phonetic approaches, on the other hand, try to simulate continuous co-articulated speech by automatically mapping speech (represented, for instance, as feature vectors) to articulatory movements [6, 7, 8]. Using such automatic approaches makes visual speech automation faster because the individual phonemes in speech do not need to be identified as the approaches rely on a regular discretization of the continuous signal. The main problem with sub-phonetic approaches is their high sensitivity to noise. Some work in the area of visual speech animation has been done for Brazilian Portuguese [9]. However, to the best of our knowledge, research in the area has not yet been published for European Portuguese (hereafter EP).

Current challenges in the field of visual speech animation include the selection of visemes, their synchronization with audio and the modeling of co-articulation. To try to tackle any of these issues, a researcher typically has to implement a visual speech animation system from scratch, which is laborious and time-consuming process [10]. Sutton et al. [11] and Berger et al. [10] introduce the first steps towards creating modular visual speech animation frameworks. Our contribution in this area involves introducing a new concept in visual speech animation, by dividing the process into several modules. We also present the first steps towards the definition of a visual speech animation system for EP. The remainder of this paper is organized as follows. In section 2, we present two different schemes of phoneme-to-viseme mappings for EP. In section 3, we introduce our proposal for a modular visual speech animation

framework. In section 4, we present results of a preliminary evaluation study that analyzed user preferences of the two mappings proposed in section 2. Finally, in section 5, we draw our conclusions and discuss our lines for further research.

2 Phoneme-to-Viseme Mapping

Phonemes are the smallest units of speech that can form contrasts between utterances. For instance, in the English minimal pair "pie" and "bye" (pronounced /paɪ/ and /baɪ/, respectively), the first consonantal sounds cause the two words to have different meanings. Therefore, we can assume that they are two distinct phonemes. The same concept can be applied in the visual domain. The visual counterpart of a phoneme is the viseme, which describes the facial and oral postures during the production of a phone. Visemes are related to the production of specific phones and are influenced by their features. Some of those features are distinctive during the production of a phone, but irrelevant in the visual domain. Nasality and voicing are examples of such features [12]. Thus, phonemes usually have a "many-to-one" relationship with visemes.

In this section, we describe our approach of mapping a 35-symbol phoneme set for EP to several classes of consonantal and vocalic visemes. Following different strategies, we created two mappings with different numbers of visemes.

The first mapping (Table 1) grouped consonants into nine different viseme classes, distinguishable primarily by the place and manner or articulation. In an attempt to reduce the number of visemes, we also created a second mapping (Table 2), which was mainly based on the place, rather than the manner, of articulation. The guttural phonemes (Table 2, Class E), whose place of articulation is near the back of the mouth, were all mapped to the same viseme, since their place and manner of articulation do not produce any relevant changes in the visual domain. The first mapping attempted to group the following types of vowels together: back vowels (Table 1, Class N), close front vowels (Table 1, Class J), close central vowels (Table 1, Class K), close-mid front and open/open-mid central vowels (Table 1, Class L), and open and open-mid front vowels (Table 1, Class M). In the second mapping, we grouped close and close-mid vowels together (Table 2, Class F), while maintaining the distinction between open-mid (Table 2, Class I) and open (Table 2, Class H) vowels. For the back vowels, we grouped close and close-mid vowels together (Table 2, Class G), and kept the open vowel separate (Table 2, Class J). This resulted in a slightly different classification of vowels, although the number of vocalic viseme classes remained unchanged. In both mappings, glides were grouped with their vocalic counterparts. So, we grouped the glide /j/ with /i/ and the glide /w/ with /u/. A final viseme, appearing as a class of its own, was considered to represent a neutral stance, or silence.

The visemes themselves were created by an experienced digital artist based on the articulatory movements made when uttering a given phoneme both on its own and in the context of other phonemes.

Table 1. First phoneme-to-viseme mapping

Viseme Class	Phonemes
A	/m/, /b/, /p/
B	/f/, /v/
C	/d/, /n/, /t/
D	/s/, /z/
E	/ʃ/, /ʒ/
F	/ɾ/
G	/l/, /ʎ/, /ɲ/
H	/g/, /k/
I	/ʀ/
J	/ĩ/, /j/, /i/
K	/ɨ/
L	/œ/, /œ̃/, /e/, /ẽ/
M	/a/, /ɛ/
N	/ɔ/, /o/, /õ/, /u/, /ũ/, /w/
S	silence/neutral

Table 2. Second phoneme-to-viseme mapping

Viseme Class	Phonemes
A	/m/, /b/, /p/
B	/f/, /v/
C	/d/, /n/, /t/, /l/, /ɾ/, /s/, /z/
D	/ʃ/, /ʒ/
E	/g/, /k/, /ʎ/, /ɲ/, /ʀ/
F	/ɨ/, /e/, /ẽ/, /i/, /ĩ/, /j/
G	/o/, /õ/, /u/, /ũ/, /w/
H	/œ/, /œ̃/, /a/
I	/ɛ/
J	/ɔ/
S	silence/neutral

3 Framework Description

Manually animating a talking 3D character that synchronizes facial movements with an audio signal quickly becomes impractical when the length of the utterances to be animated increases. We have created a system that can automatically handle utterances of different lengths based on a new framework for visual speech animation (see Figure 1 for the data pipeline).

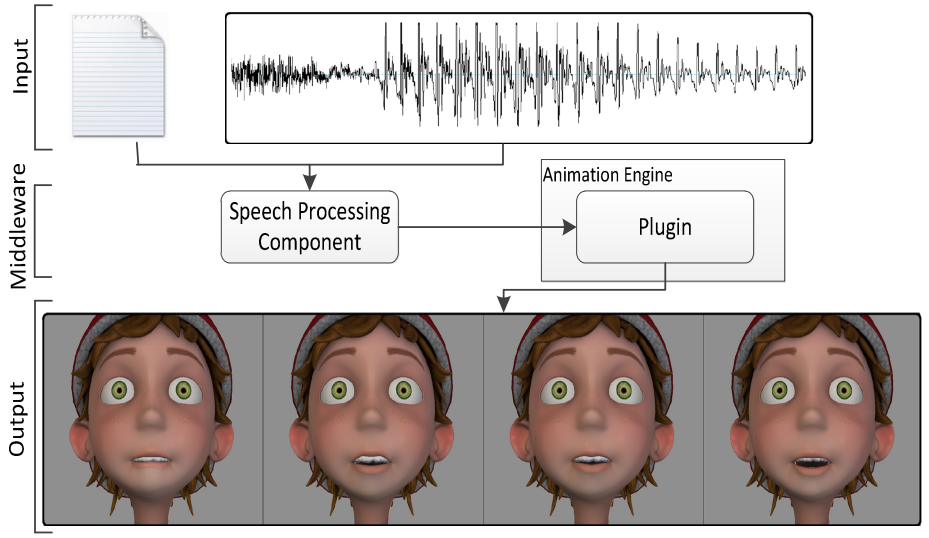


Fig. 1. Overall data pipeline of the framework. After receiving the input (audio, text or both), the speech processing tool generates the animation data and, through the plug-in, passes it on to the graphics engine that generates the animation.

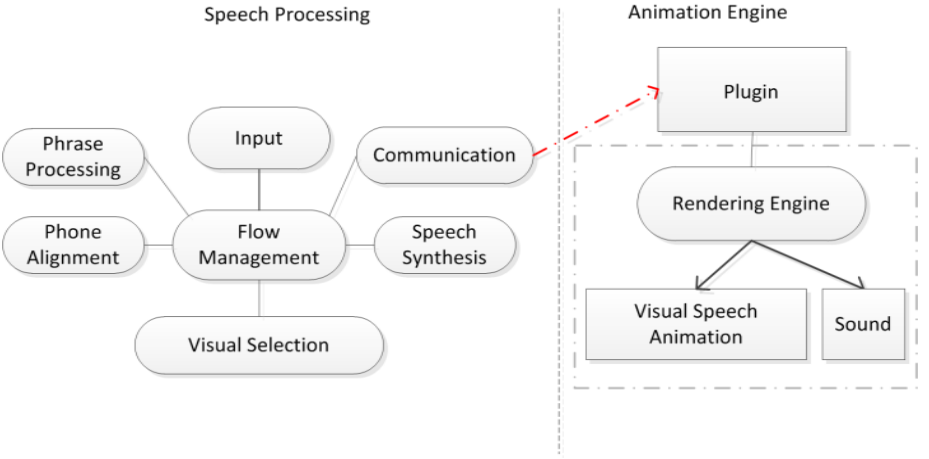


Fig. 2. Framework architecture overview. The framework is divided into the speech processing component (left) and the plug-in embedded in the animation engine (right).

The visual speech animation authoring process begins with a 3D face model in its neutral pose along with the 3D visemes, all created by a digital artist, and the utterances that we want the character to say (in the form of audio, text or both). The utterance-related information is used as direct input for animating the 3D face. The framework that makes this possible is divided into two main components: a speech

processing component and a plug-in embedded in a 3D animation engine. The speech processing component analyses the utterance-related information and generates the data that drives the animation, while the plug-in acts as an external interface to the 3D animation engine. This division allows a clear distinction between the processing of the speech data and the animation of the virtual character. Thus, integrating different animation engines in our system is possible through the adaption of the plug-in. Figure 2 illustrates the modules that constitute the conceptual framework architecture. It is important to note that the framework is independent of the system we created with it as a basis.

3.1 Speech Processing Component

The speech processing component deals with the creation of the animation data. Central to it is the flow management module that administrates the data interaction between the different modules. The input module gets the data (audio, text or both) that will drive the animation. Phrase processing uses automatic speech recognition (ASR) to obtain the phonetic transcriptions of the input utterances. The current version of our system uses Microsoft Speech API (SAPI 5.4) [13] together with an EP phonetic lexicon developed at Microsoft. The language model of the ASR engine is essentially based on unigrams, bigrams and trigrams of common words, as well as telephone numbers, person names, business names and addresses specific to the Portuguese market. The synchronization of the audio and the visemes is handled by the phone alignment module, which guarantees that the speech is matched correctly with the lip movements. If the visualization is not correct, the animated utterances may become less understandable [14]. Techniques used in ASR and speech synthesis (TTS) are commonly used for synchronization. With ASR, for instance, a time-aligned phonetic transcription can be obtained by means of a forced alignment. It aligns a speech signal with a predefined sequence of acoustic models associated with the phonemes in question. Our current approach, on the other hand, relies on the statistical duration of phonemes; the total estimated duration of the phonemes in an utterance is normalized to be the same as the duration of the corresponding speech signal. However, in the future we intent to improve this module by changing the current approach to force alignment. The EP phone durations were obtained from a database of 100 hours of Portuguese speech provided by Microsoft.

The speech synthesis module is necessary when the input is text-based. To generate the audio from text input data, the current version of our system uses the EP TTS engine that comes with SAPI.

The visual selection module plays a crucial role in the framework as it is responsible for choosing the animation curves and the visemes that the virtual character will employ. There are two possible techniques that can be applied: a sub-phonetic approach or a phoneme-to-viseme mapping. In the current system, we map the phonemes directly to the corresponding visemes. The visemes were created by an artist from directly observing the mouth movements of a person speaking each phone independently. If the sub-phonetic approach is desired, only this module needs to be changed.

Finally, the communication module sends the animation data (stored in an external file) to the plug-in so that it can be displayed by the animation engine.

3.2 Animation Engine

The animation engine is divided into the plug-in and the rendering engine. The plug-in encapsulates the animation data that is sent from the communication module. The data is later translated into the final animation by the 3D rendering engine. As an animation engine, the current version of our system uses Maya, a 3D modeling and animation authoring system [15]. A cartoon character was created in Maya that relies on a bone based rig. An artist changed the default pose to create all the visemes, wich are then concatenated based on the data given by the speech processing component

4 Preliminary Evaluation

In order to analyze the impact introduced by a new phoneme-to-viseme mapping, we carried out a preliminary subjective user evaluation. The following section describes the evaluation experiment and its results.

4.1 Experiment

The evaluation was carried out using a total of 38 subjects recruited at a student fair (20 subjects) and at a multimedia systems class at the University of Porto (18 subjects) in Porto, Portugal. The subjects did not have any problems with their vision or hearing, and only 3 of them had expertise in the area of visual speech animation. They were between 11 and 69 years of age, and 79% of them were male.

The evaluation was carried out using the following three phonetically rich sentences presented to all of the subjects:

S1: A fala é um importante meio de comunicação.
'Speech is an important means of communication.'
S2: Depois do Zé, o Ricardo joga xadrez com o Daniel.
'After Zé, Ricardo plays chess with Daniel.'
S3: O velho hoje não vê nenhum barco no mar.
'The old man does not see any ships at sea today.'

Each phonetically rich sentence was animated using the two phoneme-to-viseme mappings described in Section 3. A video with all the animations can be seen in http://youtu.be/0zZwoakx6LE. Each of the animation pairs were shown three times to the subjects, who then filled out a questionnaire according to their preferences. Corresponding to the first and second mapping, respectively, the subjects had to choose one of the following alternatives for each sentence:

L1: Strongly prefer the first animation
L2: Slightly prefer the first animation

L3: Neutral
L4: Slightly prefer the second animation
L5: Strongly prefer the second animation

4.2 Results and Discussion

Figure 3 summarizes the distribution of the preferences collected during the experiment.

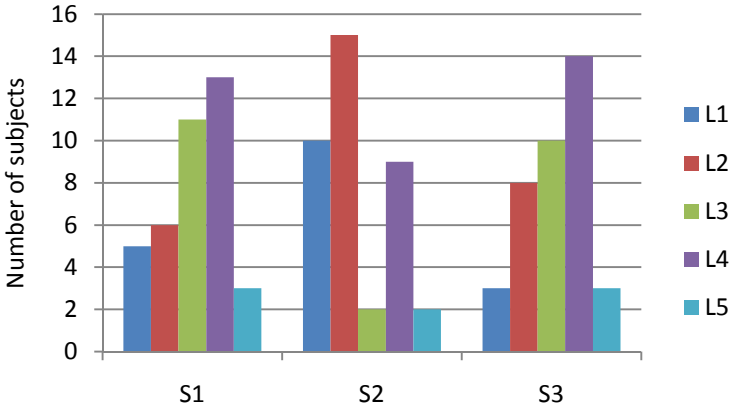


Fig. 3. The distribution of preferences for the three test sentences. L1 represents a strong preference for the first mapping and L5 a strong preference for the second mapping.

We can see from Figure 3 that the subjects slightly preferred the second mapping for S1 and S3 and that, in the case of S2, the first mapping was preferred. We can infer that the reduction in the number of viseme classes had little importance for the quality of S1 and S3, for which most users preferred the second mapping, but that it had a negative impact on the quality of S2.

The fact that the vast majority of the preferences are centered around the more neutral alternatives (L2-L4) shows that the differences in the phoneme-to-viseme mappings affected the animations less than one might expect The differences between the two animations were rather small and, hence, our results are not fully conclusive. We can, however, conclude that the use of different mappings does influence the quality of speech animation.

5 Conclusions and Future Work

It is challenging to accurately generate a talking 3D character based on speech input or text (or both) and obtain human-like facial movements. The main contribution of this paper is the creation of the first fully automatic – albeit technologically still requiring improvements – system capable of generating visual speech for European

Portuguese. The modular structure of a new visual speech animation framework makes it simple to integrate new tools into existing animation pipelines and can considerably speed up the overall visual speech animation process. Together with the framework, we also propose two different phoneme-to-viseme mappings for European Portuguese. Our preliminary evaluation experiments show that, during animation, the differences between the two mappings cause noticeable but still inconclusive changes to the quality of the animation.

In future work, we intend to improve the animation by modeling and finding a solution for the co-articulation problem, taking into account the specificity of EP. A clear starting point would be to understand the relationship between speech intensity and the visual weight distribution between visemes. As a weight model by itself is not enough to tackle the problem of co-articulation, we will also implement a co-articulation model, such as the Cohen-Massaro model [2]. We are also looking into the possibility of devising a sub-phonetic mapping method that would implicitly model co-articulation. As soon as the co-articulation problem is tackled for the case of EP, with sufficient and scientifically sound results, we will also design new objective and subjective user evaluation experiments, to validate our approach.

Acknowledgements. This work is partially supported by Instituto de Telecomunicações, Fundação para a Ciência e Tecnologia (SFRH/BD/79905/2011), the projects LIFEisGAME (ref: UTA-Est/MAI/0009/2009), VERE (ref: 257695), Marie Curie Golem (ref.251415, FP7-PEOPLE-2009-IAPP) and by FEDER through the Operational Program Competitiveness factors - COMPETE under the scope of QREN 5329 FalaGlobal. The authors would like to thank Xenxo Alvarez and Pedro Bastos for creating the visemes and the mel script in Maya.

References

1. Ostendorf, M.: Moving beyond the 'beads-on-a-string' model of speech. In: Proceedings of IEEE ASRU 1999, Keystone, CO, USA (December 1999)
2. Cohen, M., Massaro, D.: Modeling coarticulation in synthetic visual speech. Models and Techniques in Computer, 139–156 (1993)
3. Bregler, C., Covell, M., Slaney, M.: Video Rewrite. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997, New York, USA, pp. 353–360 (1997)
4. Zhou, Z., Zhao, G., Pietikäinen, M.: Synthesizing a talking mouth. In: Proc. of the 7th Indian Conf. on Computer Vision, Graphics and Image Processing, ICVGIP 2010, New York, USA, pp. 211–218 (2010)
5. Liu, K., Ostermann, J.: Optimization of an Image-Based Talking Head System. EURASIP Journal on Audio, Speech, and Music Processing, 1–13 (2009)
6. Gutierrez-Osuna, R., Kakumanu, P., Esposito, A., Garcia, O., Bojorquez, A., Castillo, J., Rudomin, I.: Speech-driven facial animation with realistic dynamics. IEEE Transactions on Multimedia 7(1), 33–42 (2005)
7. Liu, J., You, M., Chen, C., Song, M.: Real-time speechdriven animation of expressive talking faces. International Journal of General Systems 40(4), 439–455 (2009)

8. Hofer, G., Yamagishi, J., Shimodaira, H.: Speech-driven Lip Motion Generation with a Trajectory HMM. In: Proc. of Interspeech, pp. 2314–2317 (2008)
9. Demartino, J., Pinimagalhaes, L., Violaro, F.: Facial Animation based on context-dependent visemes. Computers & Graphics 30(6) (2006)
10. Berger, M., Hofer, G.: Carnival - Combining Speech Technology and Computer Animation. IEEE Computer Graphics and Applications 31, 80–89 (2011)
11. Sutton, S., Cole, R., Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, D., Cohen, M.: Universal Speech Tools: The CSLY toolkit. Language, 3221–3224 (1998)
12. Erber, N.: Auditory, Visual, and Auditory-visual Recognition of Consonants by Children with Normal and Impaired Hearing. Journal of Speech and Hearing Research 15, 413–422 (1972)
13. Microsoft Speech API (March 30, 2012)., http://msdn.microsoft.com/en-us/library/ee125663%28v=VS.85%29.aspx
14. Verwey, J., Blake, E.: The Influence of Lip Animation on the Perception of Speech in Virtual Environments. In: Proc. of the 8th Annual International Workshop on Presence, University College London, pp. 163–170 (2005)
15. Autodesk Maya (March 30, 2012), http://usa.autodesk.com/maya/

Online Learning of Log-Linear Weights in Interactive Machine Translation

Francisco Javier López-Salcedo, Germán Sanchis-Trilles,
and Francisco Casacuberta

Pattern Recognition and Human Language Technologies Group
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
{flopez,gsanchis,fcn}@dsic.upv.es

Abstract. Whenever the quality provided by a machine translation system is not enough, a human expert is required to correct the sentences provided by the machine translation system. In this environment, the human translator is generating bilingual data after each translation has been marked as correct, and expects the system to be able to learn from the errors made. In this paper, we analyse the appropriateness of discriminative ridge regression for adapting the scaling factors of a state-of-the-art machine translation system within a conventional post-editing scenario and also within an interactive machine translation setup. Results show that the strategies applied in the former setup cannot be directly applied in the latter framework. Hence, the discriminative ridge regression is revised and adapted for the interactive machine translation framework, with encouraging results.

Keywords: Machine translation, online learning, interactive machine translation.

1 Introduction

Machine translation is not only needed in fields where the amount of data is overwhelming, but also in fields where bilingual data is less abundant, but translation quality is critical. In these scenarios, machine translation systems need to collaborate closely with human experts, with the purpose of achieving high quality translations efficiently, giving rise to the popularisation of the *computer assisted translation* (CAT) [1] paradigm. In such paradigm, the *statistical machine translation* (SMT) [2] system proposes a hypothesis to a human translator, who amends the hypothesis to obtain an acceptable target sentence. Two different user interaction schemes will be considered in this paper. The first one, *post-editing* (PE), is being embraced by more and more human translators as an efficient way of generating high-quality translations. In PE, the SMT system provides an initial translation, and then the user modifies such translation so as to correct it. The second one, *interactive machine translation* (IMT) [3,4], is a more cutting-edge technology which has been receiving an increasing amount of attention. The IMT system attempts to predict the text the user is going to input. Whenever such prediction is wrong and the user provides feedback to the system, a new prediction is performed. Such process is repeated until the translation is considered correct.

One important problem which SMT systems need to tackle with when used for a CAT purpose is adaptability. In these scenarios, the user expects the system to learn

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 277–286, 2012.

dynamically from its own errors, so that errors corrected once do not need to be corrected over and over again. Hence, the models need to be adapted *online*, i.e. without a complete retraining of the model parameters, since such retraining would be too costly.

The grounds of modern SMT were established in [5], by formulating the SMT problem as follows: given an input sentence $\boldsymbol{x}$ in a certain source language, the best translation $\hat{\boldsymbol{y}}$ in a certain target language is to be found:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \ \Pr(\boldsymbol{y} \mid \boldsymbol{x}), \tag{1}$$

where $\Pr(\boldsymbol{y} \mid \boldsymbol{x})$ is modelled directly by the so-called log-linear models [6], yielding

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{x}, \boldsymbol{y}) = \underset{\boldsymbol{y}}{\operatorname{argmax}} \, \boldsymbol{\lambda} \cdot \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \underset{\boldsymbol{y}}{\operatorname{argmax}} \, g(\boldsymbol{x}, \boldsymbol{y}), \tag{2}$$

where $h_m(\boldsymbol{x}, \boldsymbol{y})$ represents an important feature for the translation of $\boldsymbol{x}$ into $\boldsymbol{y}$, M is the number of models (or features) and λ_m are the weights acting as scaling factors of the score functions. $g(\boldsymbol{x}, \boldsymbol{y})$ represents the score of a hypothesis $\boldsymbol{y}$ given an input sentence $\boldsymbol{x}$, and is not treated as a probability since the normalisation term has been omitted. Common feature functions $h_m(\boldsymbol{x}, \boldsymbol{y})$ include translation models, re-ordering models or the target language model. Typically, $\boldsymbol{h}$ and $\boldsymbol{\lambda}$ are estimated by means of training and development sets, respectively. However, the domain of such sets has an important impact on the final translation quality [7], and adaptation arises as an efficient way of alleviating this fact by using very limited amounts of in-domain data. In this paper, only $\boldsymbol{\lambda}$ will be adapted, although the same methods could also be applied to adapt $\boldsymbol{h}$ [8].

This paper is structured as follows. Next section briefly reviews related work. Then, in Sec. 3, a short introduction to IMT is presented. Sec. 4 reviews discriminative ridge regression and the modifications needed to apply it within IMT. Experiments are described in Sec. 5, and the last section is reserved for conclusions and future work.

2 Related Work

Batch adaptation (as opposed to online) is a very broad field that has been receiving a large amount of attention. In [9], adaptation in speech recognition is confronted by means of the maximum likelihood framework. In [10], the maximum likelihood framework is expanded so as to obtain maximum a posteriori estimators. In [11], adaptation is confronted as a classification problem, by extending the set of features by an additional tag. In [12], Bayesian predictive adaptation is applied for adapting $\boldsymbol{\lambda}$ in a batch setup.

However, there are also cases where there is no adaptation data at all available beforehand, and the system needs to adapt itself online without falling into an excessive time burden. Such problem led, among others, to the development of an incremental version of the Expectation-Maximisation algorithm [13]. This algorithm has been successfully applied in an IMT scenario in [14], where the models involved are incrementally updated as the user feedback is received.

In [8], an in-depth comparison of four online adaptation algorithms, i.e. passive-aggressive, perceptron, discriminative ridge regression and Bayesian predictive adaptation are studied for their application in a post-editing scenario. Both $\boldsymbol{\lambda}$ and $\boldsymbol{h}$ are

SOURCE ($\boldsymbol{x}$)**:** Para encender la impresora:
REFERENCE ($\boldsymbol{y}$)**:** To power on the printer:

ITER-0	($\boldsymbol{p}$)	( )
	($\hat{\boldsymbol{s}}$)	*To switch on:*
ITER-1	($\boldsymbol{p}$)	To
	(k)	power
	($\hat{\boldsymbol{s}}$)	*on the printer:*
ITER-2	($\boldsymbol{p}$)	To power on the printer:
	(k)	(#)
	($\hat{\boldsymbol{s}}$)	( )
FINAL	($\boldsymbol{p} \equiv \boldsymbol{y}$)	To power on the printer:

Fig. 1. IMT session to translate a Spanish sentence into English. Non-validated hypotheses are displayed in italics, whereas accepted prefixes are printed in normal font.

adapted, alternatively, presenting the most promising results when adapting the scaling factors $\boldsymbol{\lambda}$. In the present paper we study the application of the best-performing algorithm, namely discriminative ridge regression, within an IMT framework, showing that such algorithm may not be applied directly, and propose an alternative approach.

3 Interactive Machine Translation

In IMT, the purpose is to use fully-fledged SMT systems to produce full target sentence hypotheses, or portions thereof, which can be partially or completely accepted and amended by a human translator [3]. Fig. 1 illustrates a typical IMT session. Initially, the user is given an input sentence $\boldsymbol{x}$ to be translated. The reference $\boldsymbol{y}$ provided is the translation that the user would like to achieve. At iteration 0, the IMT system has to provide an initial complete translation $\hat{\boldsymbol{s}}$, as if it were a conventional SMT system. Next, the user validates a prefix $\boldsymbol{p}$ (word "*To*") and introduces a new word k. This being done, the system suggests a new suffix $\hat{\boldsymbol{s}}$. Again, the user validates a new prefix, introduces a new word and so forth. The process continues until the whole sentence is correct. In this example, a potential user of the IMT system would have typed only one word out of five, i.e., a potential effort reduction of 80% with respect to translating the whole sentence from scratch. If a PE environment is assumed as baseline, the user would have typed three words, versus only one in the case of IMT: an effort reduction of 66%.

Formally, IMT is specified as an evolution of the SMT framework. However, Eq. 1 needs to be modified according to the IMT scenario in order to take into account the part of the target sentence that is already translated, that is $\boldsymbol{p}$ and k:

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}}{\operatorname{argmax}}\, Pr(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{p}, k) \tag{3}$$

where the maximisation problem is defined over the suffix $\boldsymbol{s}$. This allows us to rewrite Eq. 3, by eliminating constant terms, achieving the equivalent criterion

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}}{\operatorname{argmax}}\, Pr(\boldsymbol{p}, k, \boldsymbol{s}|\boldsymbol{x}). \tag{4}$$

An example of the intuition behind these variables is shown in Fig. 1.

Note that, since $(\boldsymbol{p}\, k\, \boldsymbol{s}) = \boldsymbol{y}$, Eq. 4 is very similar to Eq. 1. The main difference is that the argmax search is now performed over the set of suffixes $\boldsymbol{s}$ that complete $(\boldsymbol{p}\, k)$, instead of complete sentences ($\boldsymbol{y}$ in Eq. 1). This implies that we can use the same models if the search procedures are adequately modified [3].

Typically, the IMT system makes use of the word graph generated for a given sentence in order to complete the validated prefixes [15]. Specifically, the system finds the best path in the word graph associated with a given prefix. A word graph is a weighted directed acyclic graph, where each node represents one or more partial translation hypotheses. The edges represent transitions between such nodes, and are labelled each with one word of the target sentence, and weighted by a score which evaluates how likely it is to emit such word after having already emitted the current prefix.

4 Discriminative Ridge Regression

The main purpose of discriminative Ridge regression [8] (DRR) is that *good* hypothesis within a given N-best list score *higher*, and *bad* hypotheses score *lower*. It implements the estimation of $\boldsymbol{\lambda}$ as a regression problem between $g(\boldsymbol{x}, \boldsymbol{y})$, with $\boldsymbol{y} \in nbest(\boldsymbol{x})$, and the translation quality of $\boldsymbol{y}$.

In an online learning framework, the learning algorithm processes observations sequentially. The purpose is then to modify the prediction mechanisms according to the user's feedback in order to improve the quality of future translations. Considering that the user's feedback is the reference translation $\boldsymbol{y}^{\tau}$, Eq. 2 is redefined as follows

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, \boldsymbol{\lambda}_t \cdot \boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{y}), \tag{5}$$

where the log-linear weights $\boldsymbol{\lambda}_t$ vary according to samples $(\boldsymbol{x}_1, \boldsymbol{y}_1^{\tau}), \ldots, (\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}^{\tau})$ seen before time t. To simplify notation, we will omit subindex t from input sentence $\boldsymbol{x}$ and output sentence $\hat{\boldsymbol{y}}$, although it is always assumed. Either $\boldsymbol{h}_t$ or $\boldsymbol{\lambda}_t$ can be adapted, or even both at the same time. However, in this paper, we focus on adapting only $\boldsymbol{\lambda}_t$.

The hypothesis $\hat{\boldsymbol{y}}$ that maximises the likelihood is not necessarily the hypothesis with the highest quality from a human perspective or in terms of a certain quality measure. Let $\boldsymbol{y}^*$ be the hypothesis with the highest quality, but which might have a lower likelihood[1]. Our purpose is to adapt the model parameters so that $\boldsymbol{y}^*$ is rewarded and achieves a higher score according to Eq. 2.

We define the *difference* in translation quality between the proposed hypothesis $\hat{\boldsymbol{y}}$ and the best hypothesis $\boldsymbol{y}^*$ in terms of a given quality measure $\mu(\cdot)$:

$$l(\hat{\boldsymbol{y}}) = |\mu(\hat{\boldsymbol{y}}) - \mu(\boldsymbol{y}^*)|, \tag{6}$$

where the absolute value has been introduced in order to preserve generality. The score difference between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}^*$ is related to $\phi(\hat{\boldsymbol{y}})$, which is defined as

$$\phi(\hat{\boldsymbol{y}}) = g(\boldsymbol{x}, \boldsymbol{y}^*) - g(\boldsymbol{x}, \hat{\boldsymbol{y}}). \tag{7}$$

[1] $\boldsymbol{y}^*$ does not necessarily match the reference translation $\boldsymbol{y}^{\tau}$ due to eventual coverage problems.

Ideally, we would like differences in $l(\cdot)$ to correspond to differences in $\phi(\cdot)$: if hypothesis $\boldsymbol{y}$ has a translation quality $\mu(\boldsymbol{y})$ that is very similar to the translation quality of $\mu(\boldsymbol{y}^*)$, we would like this to be reflected in translation score g, i.e., $g(\boldsymbol{x}, \boldsymbol{y})$ is very similar to $g(\boldsymbol{x}, \boldsymbol{y}^*)$. Hence, the purpose of our online procedure should be to promote this correspondence after each sample $(\boldsymbol{x}_t, \boldsymbol{y}_t^\tau)$.

For computing the new scaling factors $\boldsymbol{\lambda}_t$, the previously learnt $\boldsymbol{\lambda}_{t-1}$ is combined, for a certain learning rate α, with an appropriate update step $\check{\boldsymbol{\lambda}}_t$, yielding [8]:

$$\boldsymbol{\lambda}_t = (1-\alpha)\boldsymbol{\lambda}_{t-1} + \alpha\check{\boldsymbol{\lambda}}_t. \tag{8}$$

Although adapting $\boldsymbol{\lambda}$ is a coarse-grained strategy, its effect cannot be underestimated, since it implies adjusting the importance of every single model in Eq. 2.

4.1 Discriminative Ridge Regression in Post-editing

In a conventional post-editing scenario where the hypotheses are provided by a regular SMT system, the DRR algorithm requires an N-best list of hypotheses in decreasing order of likelihood. Let $nbest(\boldsymbol{x})$ be such a list computed by our models for sentence $\boldsymbol{x}$. For adapting $\boldsymbol{\lambda}$, we define an $N \times M$ matrix $\mathrm{H}_{\boldsymbol{x}}$, where M is the number of features in Eq. 2, containing the feature functions $\boldsymbol{h}$ of every hypothesis:

$$\mathrm{H}_{\boldsymbol{x}} = [\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}_1), \ldots, \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}_N)]'. \tag{9}$$

Additionally, let $\mathrm{H}^*_{\boldsymbol{x}}$ be a matrix such that

$$\mathrm{H}^*_{\boldsymbol{x}} = [\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}^*), \ldots, \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}^*)]', \tag{10}$$

where all rows are identical and equal to the feature vector of the best hypothesis $\boldsymbol{y}^*$ within the N-best list. Then, $\mathrm{R}_{\boldsymbol{x}}$ is defined as

$$\mathrm{R}_{\boldsymbol{x}} = \mathrm{H}^*_{\boldsymbol{x}} - \mathrm{H}_{\boldsymbol{x}}\ . \tag{11}$$

The key idea is to find a vector $\check{\boldsymbol{\lambda}}_t$ such that differences in scores are reflected as differences in the quality of the hypotheses. That is,

$$\mathrm{R}_{\boldsymbol{x}} \cdot \check{\boldsymbol{\lambda}}_t \propto \mathbf{l}_{\boldsymbol{x}}\ , \tag{12}$$

where $\mathbf{l}_{\boldsymbol{x}}$ is a column vector of N rows such that

$$\mathbf{l}_{\boldsymbol{x}} = [l(\boldsymbol{y}_1) \ldots l(\boldsymbol{y}_n) \ldots l(\boldsymbol{y}_N)]'\ ,\ \ \forall \boldsymbol{y}_i \in nbest(\boldsymbol{x}). \tag{13}$$

The objective is to find $\check{\boldsymbol{\lambda}}_t$ such that

$$\check{\boldsymbol{\lambda}}_t = \operatorname*{argmin}_{\boldsymbol{\lambda}} |\mathrm{R}_{\boldsymbol{x}} \cdot \boldsymbol{\lambda} - \mathbf{l}_{\boldsymbol{x}}| \tag{14}$$

$$= \operatorname*{argmin}_{\boldsymbol{\lambda}} ||\mathrm{R}_{\boldsymbol{x}} \cdot \boldsymbol{\lambda} - \mathbf{l}_{\boldsymbol{x}}||^2, \tag{15}$$

where $||\cdot||^2$ is the Euclidean norm. Although Eqs. 14 and 15 are equivalent (i.e. the $\hat{\boldsymbol{\lambda}}$ that minimises the first one also minimises the second one), Eq. 15 allows for a direct

implementation thanks to the ridge regression[2], such that $\check{\boldsymbol{\lambda}}_t$ can be computed as the solution to the overdetermined system $\mathrm{R}_{\boldsymbol{x}} \cdot \check{\boldsymbol{\lambda}}_t = \mathbf{l}_{\boldsymbol{x}}$, given by

$$\check{\boldsymbol{\lambda}}_t = \left(\mathrm{R}'_{\boldsymbol{x}} \cdot \mathrm{R}_{\boldsymbol{x}} + \beta \mathrm{I}\right)^{-1} \mathrm{R}'_{\boldsymbol{x}} \cdot \mathbf{l}_{\boldsymbol{x}} \,, \tag{16}$$

where a small β is used as a regularisation term to stabilise $\mathrm{R}'_{\boldsymbol{x}} \cdot \mathrm{R}_{\boldsymbol{x}}$. $\beta = 0.01$ was used in the experiments described in this paper.

4.2 Discriminative Ridge Regression in Interactive Machine Translation

When attempting to apply DRR within an IMT setting, the quality metric that is used in IMT is no longer inherent to a single hypothesis, but to a complete wordgraph. It is quite common to measure the quality of a given IMT system by computing the amount of interactions required in order to modify the system's hypothesis so that it matches the reference. Once a single word has been introduced, the IMT system modifies the suffix, which implies that the number of interactions cannot be computed as a function of the hypothesis, but must be computed by first simulating the interaction procedure and is a function of a given wordgraph. Hence, DRR, as described in previous section, cannot be directly applied within an IMT framework. One would think that optimising a certain translation quality metric would also optimise the amount of interactions required. However, experimental results detailed in Sec. 5 show that this assumption is not completely true. Hence, since the metric to be optimised by online learning does not depend on a single-best hypothesis, the formulation of DRR needs to be reviewed.

At this stage, it would be reasonable to consider instead of a list of N-best hypothesis a list of N-best wordgraphs. However, the concept of N-best wordgraph is somewhat fuzzy. For this reason, instead of computing a true list of N-best wordgraphs we will obtain a set of N scaling factors $\boldsymbol{\lambda}$ obtained at random, $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}^1, \ldots, \boldsymbol{\lambda}^n, \ldots, \boldsymbol{\lambda}^N\}$, and compute the wordgraph $W_{\boldsymbol{\lambda}^n}(\boldsymbol{x})$ associated to a given input sentence $\boldsymbol{x}$ and obtained for a certain set of scaling factors $\boldsymbol{\lambda}^n$. Of course, since the weights have been obtained at random, the resulting wordgraphs will not constitute a true list of possible N-best wordgraphs. However, since the purpose of DRR is to reward those hypotheses (in this case wordgraphs) that score well, and penalise those that score worse, what is really important is to have wordgraphs (i.e., samples of $\boldsymbol{\lambda}$) which score well, and wordgraphs (samples of $\boldsymbol{\lambda}$) which score bad. Hence, $\mathbf{l}_{\boldsymbol{y}}$ will be a column vector of N rows such that

$$\mathbf{l}_{\boldsymbol{y}} = [l(W_{\boldsymbol{\lambda}^1}(\boldsymbol{x})) \ldots l(W_{\boldsymbol{\lambda}^n}(\boldsymbol{x})) \ldots l(W_{\boldsymbol{\lambda}^N}(\boldsymbol{x}))] \tag{17}$$

Another aspect that needs to be taken care of when considering DRR within an IMT setting is that matrix $\mathrm{H}_{\boldsymbol{x}}$ also needs to be redefined, since the features that need to be considered in this case no longer correspond to those of the hypotheses in the N-best list, but to the wordgraphs generated with $\boldsymbol{\Lambda}$. Since a certain wordgraph $W_{\boldsymbol{\lambda}^n}(\boldsymbol{x})$ does not have a single set of features, but rather one feature vector for each one of the paths through the wordgraph, we will consider for building $\mathrm{H}_{\boldsymbol{x}}$ the feature vector $\boldsymbol{h}$ of the best path in $W_{\boldsymbol{\lambda}^n}(\boldsymbol{x})$, i.e., the feature vector of the best hypothesis in $W_{\boldsymbol{\lambda}^n}(\boldsymbol{x})$. Abusing

[2] Also known as Tikhonov regularisation.

Table 1. Characteristics of the Europarl corpus and NC11 test set. OoV stands for "Out of Vocabulary" words, k stands for thousands of elements and M for millions of elements.

		Spanish	English
Europarl Training	Sentences	1.4M	
	Run. words	29.9M	28.9M
	Vocabulary	129.8k	85.3k
Europarl Development	Sentences	2000	
	Run. words	60.6k	58.7k
	OoV. words	164	99
NC11 test	Sentences	3003	
	Run. words	79.4k	74.7k
	OoV. words	1549	1708

notation and with the purpose of keeping notation unclogged, let $\boldsymbol{h}_{\boldsymbol{\lambda}^n}$ be such feature vector. Then, $\mathrm{H}_{\boldsymbol{x}}$ is defined for the IMT case as

$$\mathrm{H}_{\boldsymbol{x}} = [\boldsymbol{h}_{\boldsymbol{\lambda}^1}, \ldots, \boldsymbol{h}_{\boldsymbol{\lambda}^N}]' . \tag{18}$$

Equivalently, $\mathrm{H}^*_{\boldsymbol{x}}$ is defined in this case as

$$\mathrm{H}^*_{\boldsymbol{x}} = [\boldsymbol{h}_{\boldsymbol{\lambda}^*}, \ldots, \boldsymbol{h}_{\boldsymbol{\lambda}^*}] , \tag{19}$$

with $\boldsymbol{h}_{\boldsymbol{\lambda}^*}$ being the feature vector of the best hypothesis of wordgraph $W_{\boldsymbol{\lambda}^*}(\boldsymbol{x})$, and $W_{\boldsymbol{\lambda}^*}(\boldsymbol{x})$ being that wordgraph with the best performance according to the IMT metric used, from among those computed using the random set of weights $\boldsymbol{\Lambda}$.

5 Experimental Results

Given that a true CAT scenario is very expensive, since it requires a human translator to correct every hypothesis, such scenario will be simulated by using the reference of the test set. Such reference will be fed one at a time, following an online setting.

Translation quality will be assessed by means of *Translation Edit Rate* (TER) [16] and *Word Stroke Ratio* (WSR) [4]. TER is an error metric that computes the minimum number of edits required to modify the system hypotheses so that they match the references. Possible edits include insertion, deletion, substitution of single words and shifts of word sequences. Hence, TER is an automatic metric which intends to measure the effort required to post-edit the hypotheses provided by a SMT system. WSR measures the amount of words (interactions in this case) a human translator would require to type within an IMT framework to correct the system's hypothesis. Both TER and WSR are measured in percentage, i.e., both are normalised by the total amount of words of the reference, multiplied by 100. Also in both cases, lower TER and WSR rates are better.

As baseline system, we trained a SMT system on the Europarl English–Spanish training data, in the partition of the Workshop on SMT of the EMNLP 2011 [7]. The Europarl corpus [17] (Table 1) is built from the transcription of European Parliament speeches published on the web. We used the open-source MT toolkit Moses [18][3] in

[3] Available from `http://www.statmt.org/moses/`

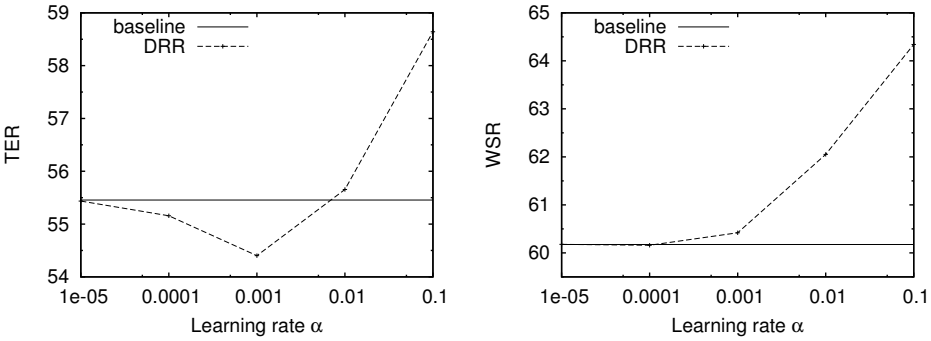


Fig. 2. Effect of the α learning rate on the effort within a PE and an IMT scenario, as measured by TER and WSR. DRR was implemented according to Sec. 4.1. N-best size was set to 1000.

its standard non-monotonic setup (including the `msd-reordering-fe` model [19]), and estimated $\boldsymbol{\lambda}$ using MERT [20] on the Europarl development set. The set of weights $\boldsymbol{\Lambda}$ described in Sec. 4.2 was obtained by sampling from a Gaussian distribution with mean vector the $\boldsymbol{\lambda}$ obtained by MERT and variance 0.01, following preliminary investigation. We also estimated a 5-gram LM with interpolation and Knesser-Ney smoothing [21]. Moses was also used for the purpose of building the required wordgraphs.

Since our purpose is to analyse the performance of an online adaptation strategy, in addition to Europarl we also considered a different test set that does not belong to the parliamentary domain, such as the News Commentary[4] (NC) 2011 test set. The News Commentary corpus was obtained from different news feeds and was used as test set for the 2011 EMNLP shared task on SMT [7]. See Table 1 for NC test set statistics.

As a first step, we carried out the experimentation according to Sec. 4.1, i.e., optimising a typical SMT evaluation metric which is ought to minimise post-editing effort. Such results can be seen in Fig. 2. The plot on the left displays TER, i.e., the amount of edits required in a PE scenario, whereas the plot on the right displays WSR, i.e., the amount of interactions required in an IMT setting. As shown, DRR achieves to provide improvements when the α learning rate is about 0.001 within the PE scenario, but fails to obtain the same results within the IMT setting. This is so because DRR, as described in Sec. 4.1, was implemented using TER as translation quality metric l. However, it would be quite risky to assume that minimising the number of edits within a PE setting would also lead to minimising the number of interactions within the IMT framework, and this fact is indeed reflected by the behaviour of WSR in the right plot of Fig. 2. It is important to point out that experiments using other translation quality metrics, such as BLEU [22], lead to similar results as the ones displayed here with TER.

After verifying that DRR, as described in Sec. 4.1, is not valid for its application in an IMT setting, we carried on implementing the version of DRR described in Sec. 4.2, and the results can be seen in Fig. 3. In this case, the approach proposed improves the amount of interactions required to correct a hypothesis, as measured by WSR. However, improvements obtained are not mirrored in the PE setting, where TER is only slightly improved for a very small α, and is actually higher with α values that do improve WSR.

[4] This corpus is available from `http://www.statmt.org/wmt11/`

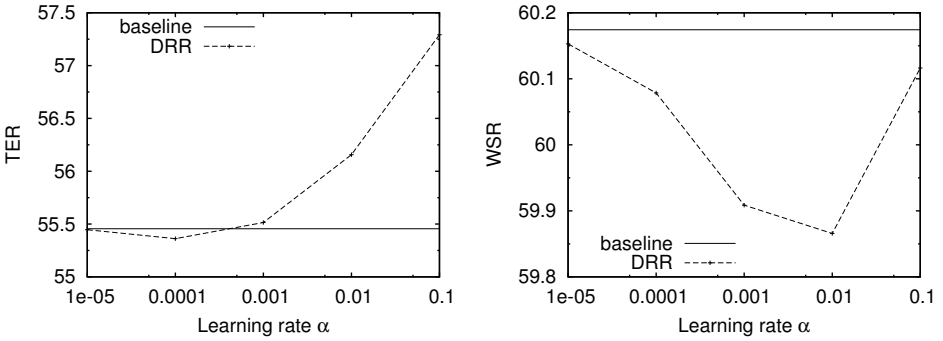


Fig. 3. Effect of α on the effort in PE and IMT, as measured by TER and WSR. DRR was implemented according to Sec. 4.2. The amount of random weights obtained was set to 500.

Table 2 sums up the results above, with the purpose of providing more precision.

Table 2. TER and WSR scores for the two optimisation methods described in Sec. 4

Optimisation method	α	TER	WSR
baseline	–	55.5	60.2
DRR (Sec. 4.1)	0.001	54.4	60.4
DRR (Sec. 4.2)	0.01	56.2	59.9

6 Conclusions and Future Work

In the present paper, we have analysed the applicability of discriminative Ridge regression within a simulated CAT environment. In the experiments reported, DRR was applied to update the log-linear weights of a state-of-the-art SMT system, both within a post-editing scenario and an interactive machine translation scenario. Results show that an implementation of DRR which optimises a traditional SMT evaluation metric and provides improvements within a PE scenario may fail to provide improvements in an IMT setting. Hence, a modification of DRR was carried out for its application in IMT, where the evaluation metric is not associated to a single hypothesis but to a complete wordgraph. Experiments with such modification present encouraging results.

As future work, we would like to study other possible ways of obtaining the set of random weights $\boldsymbol{\Lambda}$. An interesting possibility would be to obtain such weights by means of Markov chain Monte Carlo. In addition, the size of $\boldsymbol{\Lambda}$ might also be important, since the more weights sampled the higher the possibility of obtaining appropriate log-linear weights for a specific test sentence. We also intend to analyse this in future work.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement Nr. 287576 (CasMaCat). Also funded by the EC (FEDER/FSE) and the Spanish MICINN under projects MIPRCV "Consolider Ingenio 2010" (CSD2007-00018) and iTrans2 (TIN2009-14511), and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

1. Callison-Burch, C., Bannard, C., Schroeder, J.: Improving statistical translation through editing. In: Proc. of 9th EAMT Workshop "Broadening Horizons of Machine Translation and its Applications", April 26-27, pp. 26–32 (2004)
2. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)
3. Barrachina, S., et al.: Statistical approaches to computer-assisted translation. Computational Linguistics 35(1), 3–28 (2009)
4. Toselli, A.H., Vidal, E., Casacuberta, F. (eds.): Multimodal Interactive Pattern Recognition and Applications. Springer (2011)
5. Brown, P.F., Pietra, S.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19(2), 263–311 (1994)
6. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proc. of the 40th Annual Conf. of the ACL, July 8-10, pp. 295–302 (2002)
7. Callison-Burch, C., Koehn, P., Monz, C., Zaidan, O.F. (eds.): Proceedings of the Sixth Workshop on SMT. Association for Computational Linguistics, Edinburgh (2011)
8. Martínez-Gómez, P., Sanchis-Trilles, G., Casacuberta, F.: Online adaptation strategies for statistical machine translation in post-editing scenarios. Pattern Recognition 45(9), 3193–3203 (2012)
9. Christensen, H.: Speaker Adaptation of Hidden Markov Models using Maximum Likelihood Linear Regression. PhD thesis, Aalborg University, Denmark (1996)
10. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing 2, 291–298 (1994)
11. Daumé III, H.: Frustratingly easy domain adaptation. In: Proc. of the 45th Annual Conf. of the ACL, pp. 256–263 (June 2007)
12. Sanchis-Trilles, G., Casacuberta, F.: Log-linear weight optimisation via bayesian adaptation in statistical machine translation. In: Proc. of the Intl. Conf. on Computational Linguistics, August 23-27, pp. 1077–1085 (2010)
13. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1), 1–38 (1977)
14. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: Online learning for interactive statistical machine translation. In: Proc. of NAACL, June 2-4, pp. 546–554 (2010)
15. Och, F.: Minimum error rate training for statistical machine translation. In: Proc. of the 41st Annual Meeting of the ACL, July 7-12, pp. 160–167 (2003)
16. Snover, M., et al.: A study of translation edit rate with targeted human annotation. In: Proc. of the 7th Biennial Conf. of the AMTA, August 8-12, pp. 223–231 (2006)
17. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the 10th Machine Translation Summit, September 12-16, pp. 79–86 (2005)
18. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Proc. of the ACL Demo and Poster Sessions, June 25-27, pp. 177–180 (2007)
19. Koehn, P., Axelrod, A., Mayne, A.B., Callison-burch, C., Osborne, M., Talbot, D.: Edinburgh system description for the 2005 iwslt speech translation evaluation. In: Proc. of the International Workshop on SLT (October 2005)
20. Och, F.J.: Minimum error rate training for statistical machine translation. In: Proc. of the 41st Annual Conf. of the ACL, July 7-12, pp. 160–167 (2003)
21. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Proc. of ICASSP, May 9-12, pp. 181–184 (1995)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Conf. of the ACL, July 6-12, pp. 311–318 (2002)

Author Index