VTLN THROUGH FREQUENCY WARPING BASED ON PITCH

Carla Lopes and Fernando Perdigão

Abstract - This article describes a Vocal Tract Length Normalization (VTLN) procedure through frequency warping based on pitch estimates. This procedure aims to reduce the inter-speaker variability of speech signals in order to obtain a robust automatic speech recognition system.
Two additional methods are also described: one for reducing the environment variability and another for compensating the coarticulation effects on connected word pronunciation. Environment variability is compensated by explicitly modeling some frequent noise phenomena. Coarticulation phenomena compensation reduces speech signal variability by modeling events that result from coarticulation between adjacent models.
Inter-speaker variability removal is performed by a traditional speaker normalization method, which consists in expanding or compressing the *Mel* filterbank bandwidths, in order to normalize the Vocal Tract Length (VTL) of each speaker. Most of the existing methods for VTL estimation are based on formant estimation, but the difficulty of formant estimation is a known performance limitation. The proposed method overcomes such a problem since it estimates the warping factor through pitch. The recognition results, obtained for a telephone digit recognition task (with phones and sub words as units), prove that this procedure leads to similar improvements to those obtained with traditional methods based on formant estimates, actually outperforming them in some situations.

Keywords: Connected Digit Recognition, Coarticulation Models, Speaker Normalization, VTLN.

Resumo - Este artigo descreve um método de normalização de locutor através do tom que visa a redução da variabilidade de interlocutor presente nos sinais de fala de forma a conduzir a um robusto reconhecimento automático de fala. São adicionalmente descritos métodos para a redução dos efeitos de coarticulação e da variabilidade imposta pelo ambiente acústico envolvente ao reconhecedor.
A variabilidade ambiental é compensada modelando-se explicitamente fenômenos de ruído. A redução dos efeitos de coarticulação é conseguida atuando quer ao nível fonético, criando modelos acústicos que resultam da coarticulação entre modelos adjacentes, quer ao nível lingüístico, já que estas unidades obrigam à construção de um dicionário com pronúncias alternativas para algumas das palavras a reconhecer. A variabilidade interlocutor é compensada através de um método tradicional de normalização de locutor, que ao variar a largura de banda do banco de filtros em escala Mel (na obtenção dos coeficientes MFCC) pretende compensar a variação do tamanho do tracto vocal de cada locutor. Os métodos referidos em trabalhos anteriores que utilizam esta técnica baseiam-se nos formantes para obter indicadores do tamanho do tracto vocal. No entanto, estes métodos apontam como obstáculo a uma melhoria de resultados o fato de os formantes serem difíceis de estimar. O método proposto no presente trabalho ultrapassa este problema, na medida em que estima o fator de distorção, não a partir dos formantes, mas a partir do tom. Os resultados de reconhecimento apresentados, para uma tarefa de reconhecimento de cadeias de 9 dígitos, com um vocabulário de fones e subpalavras, mostram que o procedimento considerado não só alcança os resultados dos métodos tradicionais baseados em formantes como também os supera.

Palavras-chave: Reconhecimento de Dígitos Ligados, Coarticulação, Normalização de Locutor.

1. INTRODUCTION

The task of efficiently recognize spoken credit card numbers, telephone numbers or any other identification number is extremely important and requires an almost ideal recognition rate. Otherwise, it is not interesting and perhaps useless to potential users. The present work addresses the problem of an efficient recognition of nine connected digit strings, such as a telephone number. The process of speaker independent recognition of connected digits through the telephone is a special and an interesting case for Automatic Speech Recognition (ASR). It is a relatively simple task since the vocabulary involved is small but on the other hand, it must be extremely accurate since it needs only one wrong digit to result in an invalid string.
Despite the good performance of the presently available ASR systems, the recognition robustness which allow the system to operate with an unlimited vocabulary, has not been fully achieved yet; particularly, if speaker independency and environment changes are taken into account. Amongst other factors ASR systems limitations are related to speech signal variability. In addition to the adverse conditions of the environment, the effects of this variability represent the greatest challenge for present speaker independent ASR systems. In general, the sources of variability related to the speakers cannot be fully eliminated. Therefore, it is necessary that the ASR technology efficiently model this kind of obstacle. This work looks out for a valid contribution for solving the central problem of ASR system robustness. This is achieved by modulating some variability agents present on speech signals, namely inter-speaker variability, environment variability and the variability imposed by an overwhelming number of coar-

Carla Lopes is with Institute of Telecommunications, Pole of Coimbra, DEEC, Pinhal de Marrocos, 3030-290 Coimbra, Portugal and with Polytechnique Institute of Leiria ESTG, Morro do Lena, 2411-901 Leiria, Portugal. Fernando Perdigão is with Institute of Telecommunications, Pole of Coimbra and with Electrical and Computer Engineering Department (DEEC), University of Coimbra, Portugal. (E-mails: calopes@co.it.pt, fp@co.it.pt).

ticulation effects that appear in spontaneous speech.
By explicitly modeling the most frequent noise phenomena, such as background noise, clicks, labial noise and other speech production artifacts, a more robust ASR system can be achieved.
In order to improve further the system accuracy and the distinction between recognition models, other models were created to include explicitly the coarticulation present between some digits. This means that alternative pronunciations must exist for some digits in the recognition vocabulary.
In order to reduce the variances of the spectral distribution models due to speaker physiological differences, a speaker normalization method was implemented. Speaker normalization aims to reduce the speech variability resulting from differences between speakers - the so-called inter-speaker variability. These differences are essentially related to vocal and nasal tract shape and length, vocal chord physiology and also to gender and age. Speaker normalization is achieved by manipulating the acoustic parameters of the incoming speech signal in order to produce parameters similar to those of a reference speaker. If this is achieved, then the incoming speech features of distinct speakers will be similar, and the constructed recognition models will certainly be more robust. The main effect produced by speaker normalization is a shift of the speakers' formant frequencies, which are different according to their vocal tract lengths.
This paper is organized as follows. In Section 2 some speaker variability agents are presented. Section 3 characterizes the speech database used in the experiments and describes feature extraction, the model topology and the units used for recognition.
Section 4 and 5 describe the procedure for environment variability reduction and the method for coarticulation phenomena compensation, respectively. In this work it is also investigated the advantage of separating the models by speaker gender. In Section 6 the results from a gender dependent system are presented. Section 7 gives an outline of the traditional methods for speaker normalization. The pitch-based frequency warping normalization method implemented in this work is described in Section 8. Finally, Section 9 discusses the results obtained with the proposed normalization method and Section 10 concludes the paper.

2. SPEAKER VARIABILITY AGENTS

In the major speaker independent ASR systems, speech models are trained by making use of a great deal of speech, which are pronounced by a large variety of speakers. Each speaker has specific features, which are not only related to physiological characteristics like length and shape of vocal tract, but also to linguistic differences such as accent, dialect, stress and environment. Due to these speaker specific features, and to differences among all speakers, the speech signals arrive at the system with different acoustic properties. This usually originates spectral distributions with great variances and therefore with great overlapping among distinct models. These specific features severely limit the ASR systems performance.
In acoustic-phonetic context, the production of phonemes is highly dependent on the context where they appear. Moreover, they are affected by the impact that perturbations of the environment have on the speech production mechanism, and also by the physical and emotional condition of the speaker.
As previously pointed out [6], the speech of different speakers may be distinguished by changes in five classes of variability: intra-speaker variability; inter- speaker variability; environment variability; linguistic variability and context variability.
The intra-speaker variability reflects variations in the speech produced by a speaker, caused by physiological and psychological factors such as speech style, speaking rate, voice quality, environment context and stress. These factors are extremely important, since a little variation on an articulator may lead to important acoustic changes.
Inter-speaker variability is due to differences among individuals. These differences introduce specific characteristics of each person in the speech features, and are related to physical, behavioral and geographical aspects. The differences between speakers are related to intrinsic factors, essentially to the anatomy of the vocal apparatus. The variability arises due to factors related to length and shape of the vocal and nasal tract components. These factors give rise to vocal tract resonances formation, which are specific for each speaker. The formants are the peaks of the spectral signature and correspond to these resonances. Therefore, the formants will be affected by these intrinsic factors. The vocal tract length is also a function of age and gender. It increases with age up to a stationary level on adult age. Concerning gender, female speakers have shorter vocal tracts, on average, leading to formant frequencies higher then those of male speakers.
Noise is a well-known characteristic of all environments, and obviously affects the performance of ASR systems. A significant number of typical noises not only disturb the signal features but also affect the way the speech is produced, giving rise to the so-called *Lombard* effect. The appearance of a noise event in an utterance usually leads to an error, since the recognizer tends to detect an event by identifying it as the most probable alternative. Other extraneous speech events appear associated with speech pronunciation. These noises are labial noise, audible breathing noise, initial aspiration noise, coughs, hesitations and lip smacks, among others.
Linguistic variability is usually related to accent and dialect. In the former, variability is essentially linked to differences of pronunciation, while in the latter the differences are more scattered: it may result in syntax and vocabulary differences or differences in word morphology.
The contextual variability can be related to phonetic context, linguistic context (like syntax, semantics or pragmatics) or simply with social interaction.
These five classes of variability have a severe influence in the ASR system performance. Therefore, it is necessary to reduce the influence of some of these factors in such a way that allow the creation of speech models as robust as possible.
The proposed system attempts to reduce the speaker variability in three different ways. Firstly, by modeling coarticulation events, the intra-speaker variability is reduced. Secondly, the environment variability is reduced by explicitly modeling some frequent noise phenomena. Finally pitch based speaker

normalization is used to reduce inter-speaker variability and reduce the variances of the spectral distributions of the recognition models.

3. SPEECH DATA

In this work the recognition tests were performed using strings of nine connected digits from the TELEFALA database [11], a Portuguese speech database collected through the telephone network. The speech signals were recorded at a sampling frequency of 8kHz, with 8 bits/ sample, formatted with PCM-μ law. The training set has 1012 digit strings and the test set has 847. The number of female and male speakers is about 50% in both sets. The speakers are more or less uniformly distributed through all age ranges (including children) and all geographical areas of the country. Since each speaker utters a reduced number of utterances, there is a high inter-speaker variability, not only related to physiological differences but to geographic origins, rhythm and style of production. Both sets were labeled manually, based on phones and sub word units and all files were orthographically transcribed. In order to allow an initial estimation of the recognition units, 150 utterances were carefully segmented and labeled with these units.

3.1 FEATURE EXTRACTION AND MODEL TOPOLOGY

The speech signals were pre-emphasized with a pre- emphasis factor of 0.97. A 32ms Hamming analysis window with a frame rate of 10ms was used to calculate 26 filterbank energy values, for each frame. The 26 triangular shaped filters were uniformly distributed on a *Mel*-frequency scale, covering the range from 300 to 3400Hz. Finally, 12 Mel Frequency Cepstral Coefficients (MFCCs) were derived, from each frame. In addition to the 12 MFCCs, their first time derivatives (delta MFCCs), the log-energy (log E) and its first time derivative (delta log E) were also used. Each parameter vector is then composed with 26 features. Feature extraction was implemented by using the HTK toolkit [17].
The digit set of the Portuguese language was described using phones and sub word units. Furthermore, additional models were used to describe silence, general background noise and pauses between digits.
Continuous density Hidden Markov Models (HMMs), with a left-to-right topology were used to model the recognition units. On the HMMs used, only self-loops and transitions to next state are allowed except in the silence and noise models, where the last state may connect to the first one. The number of states for each recognition model was determined based on the average length of the examples in the hand-labeled set. The number of states ranges from 5 to 9 including the non-emitting states. A more detailed description of the recognition system can be found in [9].
The emission probability density functions are described as a mixture of 26-dimensional Gaussian probability density functions (with diagonal covariance matrices). In order to be able to study the recognition performance as a function of the acoustic resolution, mixtures containing from 1 to 20 Gaussians for the emission probability density function of each state were used.
Regarding the training, the recognition models were initialized starting from the hand-labeled set. Afterwards a Baum-Welch re-estimation was used to train further the models. Then the rest of the training set was realigned in order to choose the correct pronunciation of the digits. Starting with a single Gaussian emission probability density function for each state, several embedded Baum- Welch iterations were conducted. The increase of one more Gaussian mixture was made by duplicating the weightiest component followed by three embedded Baum-Welch iterations. This process was repeated until models with 20 Gaussians per state were obtained. It was observed that with more than 20 mixture components the performance of the system did not increase further significantly.
The recognition syntax allows either the existence or absence of one or more of non-digit models, before and after the digit string. Each digit may appear in any order but the string length is known. Nine digits compose each sentence and only pauses are allowed between digits. Both train and recognition were performed using HTK 2.1 software, [17].

3.2 RECOGNITION UNITS

Most of the current continuous speech recognition systems use phones as recognition units, usually with left and right context (triphones). In these systems, appending any new word to the dictionary is a simple task, but a large and phonetically balanced database is required in order to train the phones with context. Word- based models usually lead to more robust systems but only make sense when the recognition vocabulary is small because every new word model has to be trained from the scratch. Besides, these systems perform better in isolated form than in a connected word task due to coarticulation between words.
The recognition units used in this work were sub- words and phones of the digits. The choice of these units was based on an acoustic-phonetic study of Portuguese digits. The units used are acoustically well characterized, most of them corresponding to syllables or phones of the digits [9].
A set of 18 units (14 sub words and 4 phones) was defined to describe the 10 digits. It was found that a digit is not always pronounced in the same way; it is common that a vowel fall come out on the occurrence of some consonants (*elision* phenomenon). By using these recognition units, it is possible to build a dictionary with alternative pronunciations. Table 1 shows the system dictionary using the SAMPA notation for the units, [18]. Each digit model is composed of a concatenation of the selected units, with some of them shared. This is the case of the phone /S/ that is shared by the digits /dojS/, /treS/ and /s6jS/. This sharing of units corresponds to an effective increase of the available training data.

4. ENVIRONMENT VARIABILITY REDUCTION

A speech recognizer interprets each speech segment as a sequence of items of the recognition vocabulary. Even as-

Digits	System Dictionary		
/u~/	*u~*		
/dojS/	*doj*	*S*	
/treS/	*tre*	*S*	
/kwatru/	*kwa*	*tru*	
/si~ku/	*s*	*i~k*	*ku*
	s	*i~k*	
/s6jS/	*s*	*6j*	*S*
/sEt/	*s*	*Et*	*@*
	s	*Et*	
/ojtu/	*ojt*	*tu*	
	ojt		
/nOv/	*nOv*	*v@*	
	nOv		
/zEru/	*zEr*	*u*	

Table 1. System dictionary with the sequence of models for recognition.

Model Description	Adopted Labels
Silence	*sil*
Initial Aspiration	*asp*
Labial noise	*labial*
Respiration noise	*resp*
Short noises	*click*
Background talks	*fala*
Constant noise	*ruido*
Transient noises bars	*ruidov*
Narrow band noises bars	*ruidoh*

Table 2. Adopted labels for non-linguistic events.

	Mix	WRR (%)	SRR (%)	WER (%)	SER (%)	Improv.(%) WER	Improv.(%) SER
Baseline	2	90.8	52.1	9.2	47.9	–	–
Baseline	8	95.0	70.1	5.0	29.9	–	–
N/Ling Events	2	92.9	61.0	7.1	39.0	29.4	23.0
N/Ling Events	8	95.6	72.1	4.5	27.9	12.4	7.4

Table 3. Recognition rates and error rates for non-linguistic modeling. WRR - Word (digit) Recognition Rate; SRR - Sentence (string) Recognition Rate; WER - Word Error Rate; SER - Sentence Error Rate.

suming that no errors occur due to other causes, the recognizer will experience one word error if an out of vocabulary word (OOV) is produced [14]. Since this OOV does not exist in the recognition vocabulary, it will be replaced by the most probable alternative. If the expected number of words in the utterance is fixed, then for each insertion there is a corresponding word deletion. Therefore for each non-linguistic event present in the sentence, two errors may occur: an insertion and a deletion.

Analyzing the database it was found that beyond the digits there is silence, pauses and other occurrences of non-linguistic events that affect the recognition performance. These events come from the application environment and are caused by the speaker, most of the times before and after the speech production. These occurrences are labial noise, breathing noise, initial aspiration noise, speech from other persons, clicks, coughs, hesitations, background music and typical telephone line noises (frequently these line noises have vertical and horizontal bars on their spectrograms). Only 8% of the utterances do not have non-linguistic models before the digits.

Environment variability is then compensated by explicitly modeling some frequent noise phenomena. Its presence on the utterance cannot be ignored; otherwise the recognizer confuses it with other dictionary models. The non-linguistic models were specifically included in the recognition vocabulary. However, some of these occurrences do not exist in a sufficient number to allow an efficient estimation. These cases were not considered and hence they were not modeled. Table 2 lists the non- linguistic events used in the proposed recognizer and their associated labels. By using these new models, each utterance may have none, one or more of these events at the beginning and at the end. Nevertheless, there are three models for which this rule does not make sense. That is the case of the *asp*, *labial* and *resp* models. The first two situations only occur in the period that precedes pronunciation, so their existence at the end is not considered. The third situation (*resp* model) usually occurs after pronunciation of the last digit (usually in speech produced with a high speaking rate) so its instance is not considered at the beginning.

The recognition grammar considers the existence of any number of these events occurring more than once, before and after the digits. Additionally, it is assumed that only pauses may exist between adjacent digits. Speakers with a high speaking rate usually do not make pauses; on the contrary, they coarticulate between the end of one digit and the beginning of the next one. Those who make pauses, usually make short ones, not giving rise to any non-linguistic phenomenon.

The inclusion of these units requires annotation (segmentation and labeling) of the database utterances with the non-linguistic labels. As reported in [9], after this process 57% of all the labels in the entire database are non-linguistic events, 36% are silence and 21% are related to the other events defined in Table 2.

The process of explicitly modeling noise phenomena results in a significant increase of the system performance. The experimental results are presented in the next section.

4.1 ENVIRONMENT VARIABILITY REDUCTION EXPERIMENTS

The recognition results shown in Table 3 are gender independent and correspond to models with 2 and 8 components of the Gaussian mixture for each HMM state.

The experiments defined as *N/Ling Events* are those, which contain non-linguistic models. The performance is evaluated by comparing WER (Word Error Rate) and SER (Sentence Error Rate) of this system to the baseline system. The recognition results are only concerned with the recognition of the digits in the strings, not considering the non-linguistic events. The results shown in the table prove that the insertion of non-linguistic units leads to a significant improvement on WER and SER. When the number of mixtures varies between 2

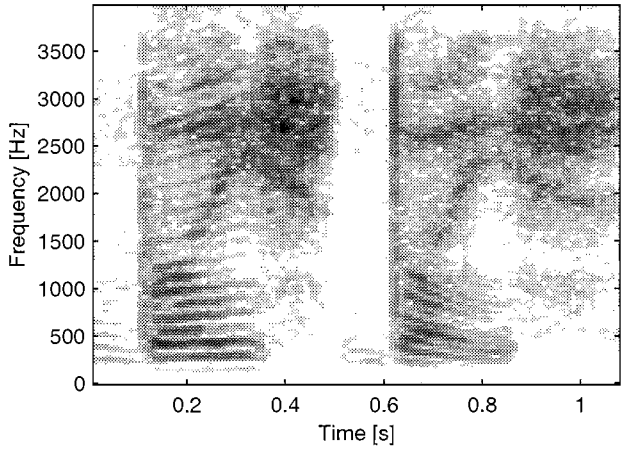

Figure 1. FFT sonogram of /dojS dojS/ when there is no coarticulation between the two digits.

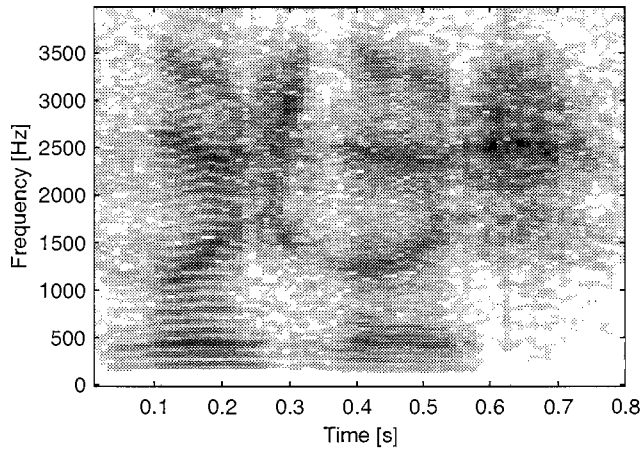

Figure 2. FFT sonogram of /dojS dojS/ when the end of the first /dojs/ coarticulates with the beginning of the second.

and 20, for WER there is an improvement between 29% and 3% while for SER this is 23% and 2%. The higher improvement occurs in models with 2 Gaussian mixtures. Although not shown in Table 3, the best results were obtained with 19 Gaussian mixtures. In this case the digit recognition rate was 96.7% and the sentence recognition rate (correct digit strings) was 77.7%.

5. COARTICULATION MODELING

During speech production, a spatial and temporal control as well as coordination of different articulators results in the sequence of phonemes. Sometimes for consecutive phonemes, these actions overlap, leading to coarticulation. In this case, the vocal tract is articulating a phoneme at the same time as it is preparing the articulation of the next one. Therefore, the degree of coarticulation is highly dependent on the pronunciation rate. The acoustic realization of a set of sounds is greatly related to the fact that our articulators cannot move instantaneously from one position to another, giving rise to phonemes that are only partially articulated.

The coarticulation is defined by Kirchhoff and Bilmes, [10] as "a change in the acoustic-phonetic content of a speech segment due to the anticipation or preservation of adjacent segments". According to these authors the degree of coarticulation varies with several factors: the speaking rate; the degree of syllabic stress and the quality of the vowel (central/peripheral and strong/weak). The authors conclude that a highly speaking rate associated with a low degree of stress leads to a strong coarticulation.

In this work, a study on sentence sonograms was performed to define the most frequent coarticulation phenomena. It was considered the existence of coarticulation when there is a change of a phone due to the presence of a neighbor one.

Figures 1 and 2 show an example of this case. Figure 1 shows a sonogram of an utterance of /dojS dojS/ ("two two"in Portuguese) pronounced with no coarticulation. In this case the sequence that better models the acoustic events is a sequence of the four following models: *doj*, *S*, *doj*, *S*. In the production of the diphthong /oj/ there is a decrease of the second formant followed by an increase. The phoneme /S/ is characterized by an energy cloud. Since /S/ is unvoiced there is no trajectory for the second formant ($F2$) in this period. The second digit is articulated in a similar way. When /dojS dojS/ is pronounced with coarticulation (Figure 2), the first /S/ coarticulates with the beginning of the adjacent /dojS/. The trajectory of $F2$ for the first /oj/ remains but there is a perturbation in the /S/ of the first /dojS/. In this period it is clear a continuity of $F2$ from de first /oj/ to the second one, which leads to the conclusion that the pronunciation of /S/ changed. The fact that following a /S/ there is a voiced consonant makes /S/ to sound like a /Z/ [8], [9].

In this study on sonograms, many coarticulation phenomena similar to those of figure 1 and 2 were found. Other example appears in the presence of two adjacent vowels, what result in the fall of one of them (elision).

When the speech production has an absence of a clear separation between the acoustic specific features of each phoneme or sub word, coarticulation is probably present. This situation leads to a poor estimation of the corresponding models and the performance of the system degrades. By explicitly modeling these pronunciation variations, some errors induced by speaker production variability could be corrected. It was decided to add four units of coarticulation to the models set. These units are not the only ones that appear in the database; however, the others do not occur often enough to obtain an accurate estimation. The most common coarticulation units were selected and presented in Table 4.

The phoneme /z/ appears when /S/ is followed by a vowel, while /Z/ appears when /S/ is followed by a voiced consonant. Other cases of common coarticulations are /u-u / and /u-ojt/ which correspond to two adjacent vowels. Coarticulation phenomena compensation reduce speech signal variability due to phonetic context. This mechanism is two fold: on the one hand, it operates at the phonetic level because coarticulation models were introduced; on the other hand, it operates at the phonological level because these new models originate alternative rules of pronunciation. The vocabulary becomes larger and there are some digits with more pronunciation alternatives. Regarding Table 1, the digit /dojS/ comprises the

Coarticulated Phones	Resulting Models
/S ... u~/	$\tilde{z}$
/S ... ojt/	$\tilde{z}$
/S ... zEr/	*Z*
/S ... doj/	*Z*
/u ... u~/	*u-u~*
/u ... ojt/	*u-oit*

Table 4. Selected coarticulation units.

	Mix	WRR (%)	SRR (%)	WER (%)	SER (%)	Improv.(%) WER	Improv.(%) SER
N/Ling Events	2	92.9	61,0	7.1	39,0	–	–
N/Ling Events	8	95.6	72.1	4.5	27.9	–	–
Coart. Events	2	93.3	62.3	6.7	37.7	6.9	3.5
Coart. Events	8	96.1	75.9	3.9	24.1	13.8	15.6

Table 5. Recognition rates for coarticulation modeling.

sequence of models *doj*, *S*. After inserting the coarticulation models, this digit has three alternative compositions: *doj S*, *doj Z* or *doj* $\tilde{z}$.

Despite the fact that the process of explicitly modeling coarticulation phenomena increases the number of models and alternative pronunciations, a significant improvement in the system performance is achieved. The experimental results are presented in the next section.

5.1 COARTICULATION MODELING EXPERIMENTS

In Table 5 the results obtained with *N/Ling Events* experiences are compared with the system with the coarticulation models, labeled as *Coart. Events*. Both experiments refer to models with 2 and 8 components of the Gaussian mixture for each HMM state.

The Table shows that better results are achieved by modeling coarticulation events, which validates the proposed approach. In the case of 2 Gaussian mixtures the improvement is 6.9% in WER and 3.5% in SER, while for 8 mixtures it is 13.8% in WER and 15.6% in SER.

Although it is not shown in the table, the best results were obtained with 20 Gaussian mixtures reaching 96.9% and 80.1% for word (digit) and sentence recognition rate, respectively.

6. GENDER DEPENDENT SYSTEM

One of the most important causes of variability in speech production is due to gender differences, especially VTL differences. Therefore, the automatic use of different acoustic models according to either speaker type or gender might help to increase the robustness of the ASR system.

Speaker dependent automatic systems are known to outperform speaker independent systems and nowadays gender dependent models are standard in speaker independent speech recognition systems. In these systems there is a pair of models for each unit. Each element of a pair corresponds to a different gender and is trained with male and female speech. In the decoding phase the model with higher likelihood is chosen.

	Mix	WRR (%)	SRR (%)	WER (%)	SER (%)
Coart. Events	2	93.3	62.3	6.7	37.7
Coart. Events	8	96.1	75.9	3.9	24.1
Female Models	2	96.4	76.2	3.6	23.8
Female Models	8	97.9	85.9	2.1	14.1
Male Models	2	96.4	79.6	3.6	20.4
Male Models	8	97.6	85.6	2.4	14.4
Female & Male	2	96.3	76.4	3.7	23.6
Female & Male	8	98.0	86.0	2.0	14.1

Table 6. Gender-dependent recognition results.

In order to introduce the gender dependent models, each model was duplicated and labeled accordingly. Regarding the train, the database was split up in two sets, according to gender. The models are separately estimated, with the male set estimating male models and the female set estimating female models. The noise models are generic in the sense that they are gender independent, because the speaker gender does not affect these models.

In this kind of systems, the available data for each set is less than in generic systems, where the models are independent of the gender. A total of 708 and 836 train utterances were obtained for female and male sets, respectively. In the set only half of the data is available, i.e., 382 female and male utterances.

The number of models increases, since the system uses two HMMs for each model: a female and male model. In the gender independent system, 18 recognition units and 4 coarticulation units describe the 10 digits, while in gender dependent system the recognition models are 36 and 8 coarticulation models.

The recognition grammar for the gender dependent system has the same format as the one for the gender independent system. However, in the former the number of possible words to recognize doubled because there is a male and a female version for each digit. Note that in this case the number of models to estimate is higher, which increases the computational cost.

Table 6 shows the experimental results obtained for three different recognizers: a female, a male and a gender dependent. *Coart. Events* label refers to the previous results obtained with non-linguistic and coarticulation models, while *Female models* and *Male models* refer to the training and testing sets separated by gender. *Female & Male* refers to the implemented system, which makes use of both gender models in the recognition process. As it can be seen in Table 6, the results show a significant improvement on the recognition rate of the gender dependent system. With 2 mixtures, the digit recognition rate (WRR) increases from 93.3% (*Coart. Events*) to 96.3% (*Female & Male*) while with 8 mixtures the same rate increases from 96.1% to 98%. This rising in recognition rate corresponds to an improvement of 81% and 95% on WER for models with 2 and 8 mixtures, respectively. In regard to the sentence recognition rate (SRR) the results are also significantly better. In the *Coart. Events* the rate is 62.3% and 75,9% for 2 and 8 mixtures, respectively. It rises to 76.4% and 86% if *Female & Male* models are used, cor-

responding to an improvement of 60% and 72% on SER for models with 2 and 8 mixtures, respectively.
The performance of female models in respect to a male test set was also tested. By analyzing the results one can realize that when data becomes more robust from the point of view of one set, it conducts to a decay in the recognition performance of the opposite gender set. This demonstrates the differences between genders, and motivates normalization.

7. SPEAKER NORMALIZATION PROCEDURE

In the past, different techniques have been investigated to normalize the parametric representation of speech signals through manipulation of its acoustic parameters. By reducing the speech signal variability due to inter-speaker differences, the ASR performance might be improved.
One of the techniques widely used in speaker normalization is the frequency warping technique. This technique attempts to normalize the vocal tract length of different speakers by reducing their influence on the spectral parameters. Using this method, the acoustic parameters are transformed by warping the speech signal in the frequency domain. This warping can be performed in two distinct ways. In the first one, warping is obtained by either compressing or expanding the speech signal in the spectral Fourier domain [1], [3], [13]. In the second one, warping is obtained either compressing or expanding the filter bank responses, used in the MFCCs estimation, [2], [7], [16]. Whether the warping is applied on the spectral signal or directly on the filter bank, the goal is the same: both of them attempt to map the spectrum of a phoneme pronounced by distinct speakers into a standard one.
Figure 3 illustrates this mapping process supposing that a female and a male speaker pronounce the same vowel. In the case of the female speaker all formants are above the reference vowels. In order to enable the filter bank to capture the same spectral information, this has to be expanded. In case of the male speaker the situation is almost the opposite. In order to capture the same information, the filter bank responses have to be compressed.
The mapping process of a speech signal spectrum pronounced by distinct speakers into a standard one is performed by a warping function that depends on a single warping factor. The warping function establishes the relationship between the frequency axis used to represent the speech produced by a reference speaker and the same axis in the case of a given speaker.
Both the selection of the warping factor and the shape of the function are vital for the success of the application. In regard to the shape, a wide variety of functions were proposed: linear as in Lee and Rose's work, [7]; piecewise linear as by Wegmann et all, [13]; non linear by Eide and Gish, [1], or bilinear as by Zhan and Waibel, [16] and Fukada and Sagisaka, [2]. Regarding the selection of the warping factor there are two main procedures: the selection based on maximization of likelihood (ML), [7],[13],[16] and the selection based on speaker specific acoustic parameters, [1],[4]. The first one uses a predefined set of warping factors and selects the best one for a specific speaker by following an iterative procedure

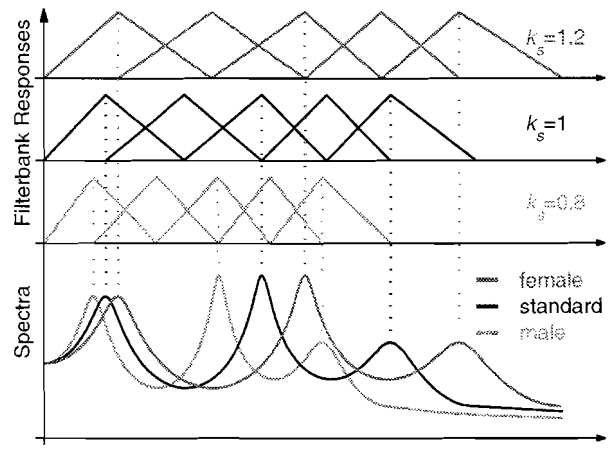

Figure 3. Mapping of a vowel spectrum into a standard spectrum.

based on ML. The warping factor is selected such that the probability of a set of acoustic features (of a given speaker) is maximized in regard to a reference acoustic model. The second one selects the warping factor by using an approach based on the measurement of the formant frequencies of the speaker. According to the authors, the position of these reflects the VTL.
Several authors obtained better results by using the maximization of likelihood criterion. However, the method based on speaker specific parameters has the important advantage of being computationally less expensive. Nevertheless, formants estimation is liable to errors, especially when the system works in adverse conditions. In the present work the warping factor was selected from pitch ($F0$) in order to overcome such problems.
Gouvêa, [3], in his work, uses the median of the three formants. He pointed out that the system performance only stabilizes when each speaker data reaches 12 seconds. On the proposed system, a reasonable estimation of pitch after a small set of voiced frames can be obtained.

8. PITCH BASED FREQUENCY WARPING

It is somehow intuitive that VTL estimation is based on acoustic studies. However, as already mentioned, this direct estimation from the speech signal is difficult because the relationship between formants and VTL is not simple.
In order to deal with this situation Eide and Gish, [1] proposed a method, which conducted to significant improvements in performance. Their method is based on the warping function given by the equation (1) and sketched in Figure 4, where k_s is the ratio between the median third formant ($F3$) of a given speaker and the median of $F3$ of all speakers of the train set.

$$f' = k_s^{\frac{3f}{8000}} \times f \quad (1)$$

The preference on $F3$ is due to the fact that this formant is the most stable, i.e., it is less dependent of linguistic information and therefore, from the statistic point of view, it is the most robust. In [15] Zhan and Westphal also use this criterion. They expand the work of Eide and Gish, [1] and make the

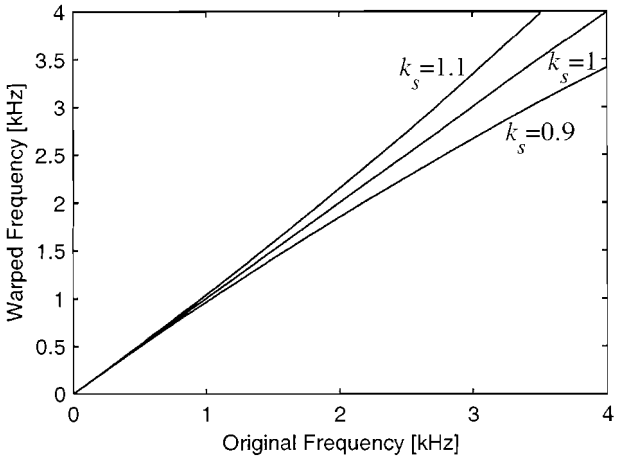

Figure 4. Warping function shape used for determining the low and high frequencies of the *Mel* scale filter bank given by (1).

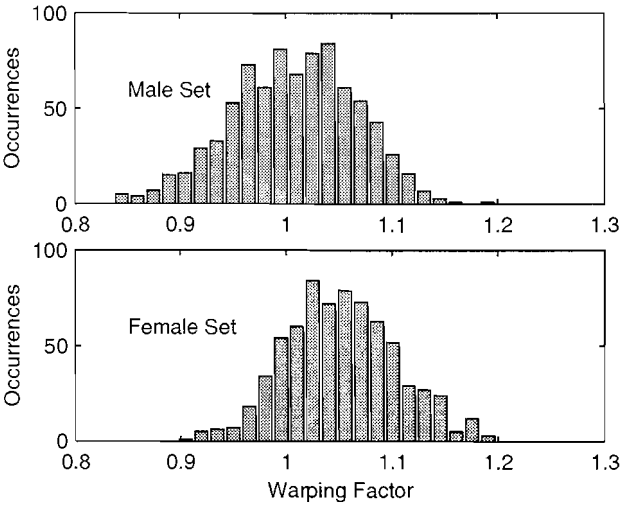

Figure 5. Warping factor distributions based on $F3$ and separated by gender.

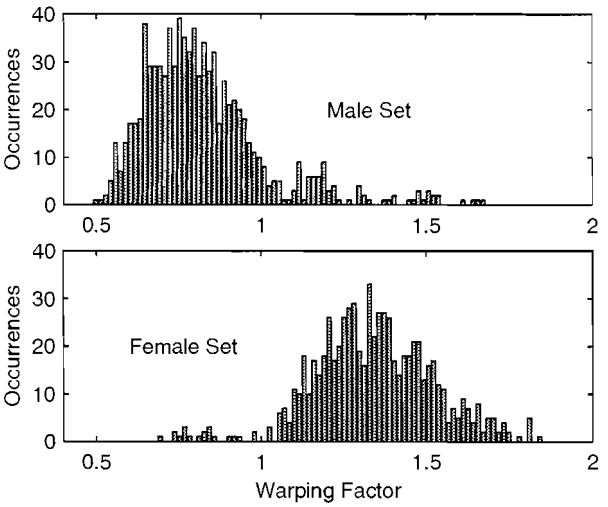

Figure 6. Warping factor distribution based on pitch.

normalization with the same warping function but estimating the warping factor also from the first and second formants. They did not achieve better results with this method. The proposed method makes use of each speaker's pitch to estimate his VTL and to perform normalization. This procedure seems worthwhile since pitch is more stable than $F3$ and its estimation is more reliable. Pitch determination obliges to a voiced/unvoiced separation, but this is also necessary on formant determination since it does not make sense to estimate formants in unvoiced frames. The motivation for choosing pitch comes from the fact that female speakers have shorter vocal tracts, higher formants and higher pitch frequencies, so pitch and vocal tract length are probably correlated.

As mentioned before, by doing normalization using $F3$ as an indicator leads to good results and, since the proposed method intends to perform normalization from $F0$, it is necessary to analyze if there is any relationship between these two features. $F0$ and $F3$ were calculated for each sentence frame. The pitch and the formants were estimated through the algorithms AMPEX [5] and SFS [12], respectively.

By observing a scatter plot of $F0$ and $F3$ means and medians a correlation factor of 0.45 and 0.3, respectively, were found. Since this last value is significantly below the first one, it was decided to use the means, $\overline{F0}$ and $\overline{F3}$, instead of medians, as a main factor for normalization.

The proposed method still uses Eide and Gish's function but k_s was defined using $F0$ instead of $F3$. Since the ratios $F3/\overline{F3}$ and $F0/\overline{F0}$ are different, k_s will be affected by a value according to expression (2), where $\overline{F0}$ is the mean of $F0$ among all speakers.

$$k_s = \alpha \frac{F0}{\overline{F0}} \qquad (2)$$

Each speaker has his own warping factor, which is independent of the speaker gender. However, since one of the major factors of variability is associated with gender differences, it was decided to analyze two sets of speakers, one of each gender, separately.

The warping factors (WF) for both sets were computed in two distinct ways. In the first case, WF was obtained as the ratio between the mean of the third formant ($F3$) of a speaker and the mean of $F3$ of all speakers in the train set ($\overline{F3}$=2300Hz). In the second case, WF was defined as the ratio between the mean pitch of the voiced set of a given speaker and the mean pitch of all speakers ($\overline{F0}$=160Hz). All means were estimated only by using voiced frames, detected by a pitch detector. The resulting warping factor distributions are depicted in Figures 5 and 6. In both Figures 5 and 6 the distributions of the warping factors for the female set are above those of the male one. This leads to the conclusion that the 3$^{\text{rd}}$ formant frequencies of male speakers are under the female frequencies and obviously the same happens with the pitch. However, this last distance is greater. It was found that the mean pitch of the female set is about 80Hz ($\approx$ 1 Bark) above the mean pitch of the male set.

By comparing both distributions, it can be seen that they have similar shapes and this was the motivation to find a mapping function between the two distributions. In order to obtain such a function, it was assumed that below the *standard* speaker's pitch (160Hz) the mapping function normal-

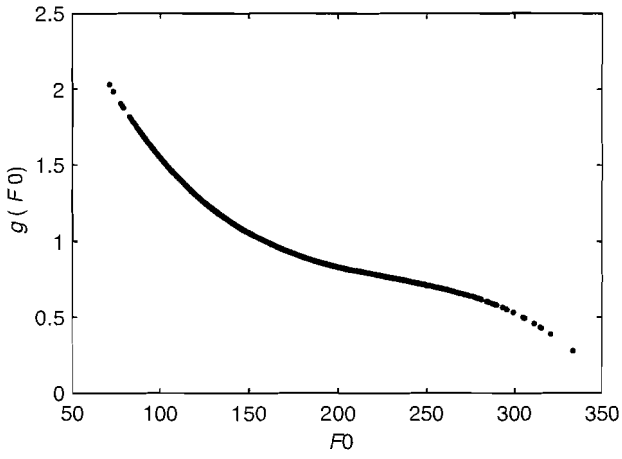

Figure 7. Third order parabolic function used for establishing α in equation (2).

izes male speakers and above this value it normalizes female speakers. Between 150Hz and 170Hz the warping factor distributions from female and male speakers overlap, hence no normalization was considered. After defining several mapping points, a third order fitting was done on them, resulting in the mapping function drawn in Figure 7. The ordinate $g(F0)$ establishes the value by which the pitch ($F0$) should be warped to agree with the third formant ($F3$) value. This function establishes α in equation (2), which becomes a function of $F0$, i.e., $\alpha = g(F0)$. The warping function is then given by:

$$f' = \left(g(F0) \frac{F0}{\overline{F0}} \right)^{\frac{3f}{8000}} \times f \qquad (3)$$

The normalization method proposed in this work is implemented in three distinct steps. First, the mean pitch in voiced frames is determined. Then the warping factor is calculated through the equation (3). Finally, the upper and lower frequencies of the *Mel* scale filter bank are changed accordingly to this equation.

In order not to exceed the *Nyquist* frequency, the warping factor was restricted to an interval ranging from 0.9 to 1.1. Values below 0.9 generate a signal with too much compression, which probably gives rise to a significant loss of information. This warping factor restriction makes equation (3) only a coarse estimation of $F3$.

8.1 PITCH BASED FREQUENCY WARPING EXPERIMENTS

The recognition results based on normalization with $F3$ (*Normaliz. w/ F3*) and $F0$ (*Normaliz. w/ F0*) are presented in Table 7. All experiments used training and testing sets with male and female speakers. The performance was determined by comparing the normalization experiences results with the non-linguistic events modulation method (*N/Ling Events*). Taking into account the normalization procedure, the proposed method not only reaches the results achieved by Eide and Gish [1] (*Normaliz. w/ F3*) as it outperforms them. The results of complete utterance with 2 mixtures show an improvement of 23.1% for the *Normaliz. w/ F0* method and

	Mix	WRR (%)	SRR (%)	WER (%)	SER (%)	Improv.(%) WER	SER
N/Ling Events	2	92.9	61.0	7.1	39,0	–	–
N/Ling Events	8	95.6	72.1	4.5	27.9	–	–
Coart. Models	2	93.3	62.3	6.7	37.7	6.9	3.5
Coart. Models	8	96.1	75.9	3.9	24.1	13.8	15.6
Normaliz. w/ F3	2	93,5	63,3	6,5	36,7	9,9	6,1
Normaliz. w/ F3	8	96,3	76,9	3,8	23,1	18,7	20,4
Normaliz. w/ F0	2	94,4	68,4	5,6	31,6	28,1	23,1
Normaliz. w/ F0	8	96,3	77,9	3,7	22,1	19,9	26,2

Table 7. Pitch based frequency warping recognition results.

6.1% in the Eide's method when compared to the baseline. With 8 mixtures the results are even better, 26.2% and 20.4% for normalization based on $F0$ and $F3$, respectively. However, the results with an increased number of mixtures did not accomplish the previous ones, leading to the conclusion that a fewer number of Gaussian mixtures will be necessary to model each sub word unit. Since the results of recognition were superior, it is expected that the sub words models become more compact.

The best results were obtained with 17 Gaussian mixtures, with normalization based on pitch. The digit recognition result was 96.9% and the sentence recognition result was 81.6%.

Another method was tested by considering models of entire word digits. The word recognition rate and sentence recognition rate obtained were 99.2% and 93.4%, with 20 mixtures. The normalization method applied to models of entire word digits does not evidence great improvements. It was obtained over 0.8% WER and 6.6% SER with 20 mixtures.

Although the recognition results of the normalization method are encouraging, they did not reach the same rate as entire word models. Nevertheless, taking into account the system versatility, where the vocabulary is based on sub words and phones, this method can be very useful in applications with large vocabularies.

9. CONCLUSIONS

The goal of the work described in this paper was to obtain a robust digit string recognizer for the Portuguese language. The most notorious and influent sources of variability related with the speaker were identified and studied. From this process it was concluded that some robustness improvement might be achieved by reducing environment variability and compensating coarticulation phenomena. Moreover, by removing inter-speaker variability through a normalization procedure, one can also obtain a more robust recognition system.

It was found that, by modeling noise and coarticulation events the system performance is significantly increased, hence validating the proposed method. The results prove that non-linguistic unit insertion leads to an improvement of 29% in WER and 23% in SER. By modeling coarticulation events the improvements were 13.8% in WER and 15.6% in SER.

A speaker normalization method based on pitch has been proposed. The method proved to be very useful and the recognition rate of a 9 connected digit string was increased. The method also overcomes the dependency of the system perfor-

mance on the reliability of formant estimation and a reasonable estimation of pitch can be achieved with only a small set of voiced frames.
The normalization based on pitch reached the same results as those obtained with formants. An improvement of about 26% on SER was achieved compared with the baseline performance, using 8 mixtures. These results stress the fact that a good correlation between pitch and vocal tract length should exist.
The proposed method did not reach the recognition results of the gender dependent system, presented on Table 6. The method that uses gender dependent models with 2 mixtures achieves recognition rates of 96% on digit recognition and 76% on complete utterance recognition, while the normalization based on pitch only reaches 94% and 68% for the same rates. Nevertheless, the proposed normalization method does not require model duplication and therefore its computational cost is about half.

REFERENCES

[1] Eide, E., Gish, H., A Parametric Approach to Vocal Tract Normalization, Proc. ICASSP' 96, vol. 1, pp. 346-348, May 1996.
[2] Fukada, T., Sagisaka, Y., Speaker Normalized Acoustic Modeling Based on 3-D Viterbi Decoding, Proc. ICASSP'98, vol.1. 1, pp. 437-440, Seattle, WA, May 1998.
[3] Gouvêa, E., Acoustic Feature-Based Frequency Warping for Speaker Normalization, Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, December 1998.
[4] Gouvêa, E., Stern, R., Speaker Normalization Through Formant-Based Warping of the Frequency Scale, Proc. Eurospeech, Rhodes, 1997.
[5] Immerseel, L., Martens, J., Pitch and Voiced/Unvoiced Determination with an Auditory Model, JASA 91(6), pp 3511-3526, June 1992.
[6] Junqua, J, Haton, J., Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, 1996.
[7] Lee, L. and Rose, R. C., Speaker Normalization using Efficient Frequency Warping Procedures, Proc. ICASSP'96, May 1996.
[8] Lopes, C., Perdigão, F., Modeling Intra and Inter Speaker Variability, 12th Conference on Pattern Recognition RECPAD2002, Aveiro, Portugal, June 2002.
[9] Lopes, C, Construção de Modelos Robustos de Reconhecimento de Fala Através da Modelação da Variabilidade Presente nos Sinais, Master Thesis, Coimbra University, Portugal, July 2002.
[10] Kirchhoff, K., Bilmes, J., Statistical Acoustic Indications of Coarticulation, Proceedings of ICPh99, pg. 1729-1732, San Francisco, 1999.
[11] Neves, F., Amaral, R., Plácido, P., Marta, E., Perdigão, F., Sá, L., A Portuguese Telephone Speech Database Collected Using an Automated System, 7ª Conf. da Associação Portuguesa de Reconhecimento de Padrões, Aveiro, 1995.
[12] Mark Huckvale, Speech Filing System, SFS Release 4.25, Version 1.25, University College of London. http://www.phon.ucl.ac.uk/resource/sfs/help/index.htm
[13] Wegmann, S., McAllaster, D., Orloff, J., Peskins, B, Speaker Normalization on Conversational Telephone Speech, Proc. ICASSP '96, vol. 1, pp. 339-341, May 1996.
[14] Woodland, P., Johnson, S., Joulin, P., Jones, K., Effects of Out of Vocabulary Words in spoken Document Retrieval, Proceedings of SIGIR 2000, Athens, Greece 2000.
[15] Zhan P., Westphal M., Speaker Normalization Based on Frequency Warping, Proc. ICASSP '97, pp. 1039-1042, Munich, Germany, April 1997.
[16] Zhan, P., Waibel, A., Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition, CMU-LTI-97-150, May 1997.
[17] Young, S., Jansen, J., Odell, J.,Ollasson D. and Woodland, P., The HTK Book (for HTK Version 2.1), Cambridge University, Cambridge, UK, 1995.
[18] John Wells, "SAMPA - Computer Readable Phonetic Alphabet", http://www.phon.ucl.ac.uk/home/sampa/home.htm, http://www.phon.ucl.ac.uk/home/sampa/portug.htm.

Carla Lopes graduated in Electrical Engineering in 1997, at the Department of Electrical Engineering, University of Coimbra, Portugal. She received the M.S.c degree in Systems and Automation in 2002 also from the Department of Electrical Engineering of the University of Coimbra, Portugal. Since 1998 she is with the Department of Electrical Engineering of the Polytechnic Institute of Leiria, Portugal where she is teaching Electronics and Telecommunications. C. Lopes is also a research member of the Institute of Telecommunications, Coimbra, Portugal where she works on Robust Speech Recognition. Her research interests are in the field of speech processing and recognition.

Fernando Perdigão graduated in Electrical Engineering in 1985 from the University of Coimbra, Portugal. In 1998 he received the Ph.D. degree from the same university, after which he became Assistant Professor in the Electrical and Computer Engineering Department. He is also a researcher of the Institute of Telecommunications at the pole of Coimbra. His research interests are in the area of signal processing, auditory modelling, speech recognition and neural networks.