# ON THE USE OF ACOUSTIC SEGMENTATION IN SPEAKER IDENTIFICATION[1]

*L. Rodríguez-Liñares and C. García-Mateo*
E.T.S.E. Telecomunicación
Dept. de Tecnoloxías das Comunicacións.
Universidade de Vigo,
36200-VIGO (Pontevedra), Spain.
Tel. +34 86 812664, FAX: +34 86 812116, E-mail: leandro@tsc.uvigo.es

## ABSTRACT

In this paper, we present a novel architecture for a Speaker Recognition system over the telephone. The proposed system introduces acoustic information into a HMM-based recognizer. This is achieved by using a phonetic classifier during the training phase. Three broad phonetic classes: voiced frames, unvoiced frames and transitions, are defined. We design speaker templates by the parallel connection of the outputs of the single state HMM´s and by the combination of the single state HMM's into a four state HMM after estimation of the transition probabilities. The results show that this architecture performs better than others without phonetic classification.

## 1. INTRODUCTION

Continuous HMM (Hidden Markov Model) based systems are presently the state of art for speaker recognition purposes. It is well known [1] that their performance relies on the total number of Gaussian mixtures of the model and not so much on how many states are used. Thus, one single-state multiple-Gaussian mixtures model called GMM is one of the preferred algorithms for this task [2]. After training it is supposed that the Gaussian components have learned the more relevant aspects regarding the distinctive phonetic features that characterize a person voice. The higher the number of Gaussian mixtures, the better. The important point is then to decide how many mixtures must be used to achieve a good compromise between representation accurateness and the amount of data required for training. Computational complexity must be kept as low as possible as well.

As the vocal tract exhibits widely articulatory configurations during the production of distinct sounds, an **average** set of features does not represents a speaker's voice characteristics accurately. To include acoustic discrimination helps to improve performance. The point is to select the best sound classes and how to perform a robust automatic speech classification.

In this paper, we investigate the role played by several phonetic characteristics regarding speaker recognition using HMM based systems. Namely, voiced part, unvoiced part and transitions are considered. Eventually, we propose an ergodic HMM model that combines these phonetic classes. In this way, we force the learning capability of the model and reduce the required number of Gaussian when compared with a complexity equivalent system that makes no use of a priori phonetic information.

The rest of the paper is organized as follows: Section 2 presents the database we used. In Section 3 we present the architectures of the models and in Section 4 the results obtained with these models. Finally, section 5 presents some conclusions and guidelines for future work.

## 2. EXPERIMENTAL CONDITIONS

The experiments were conducted using our own database called "TelVoice". It has been designed for Speaker Recognition purposes and its goal is to have at least 50 Spanish speakers with 10 sessions each, recorded over a period of one month an a half. Thus, time interval between sessions may vary from one speaker to another, but it is never less than three days. It consists of telephone speech sampled at 8 Khz.

In order to asses the performance of the proposed system, we have set up an experiment making some choices about recording conditions and speech parameterization. Mel-cepstrum and Δ-mel-cepstrum coefficients were computed using a frame length of 20ms, and a frame period of 10 ms. Energy and the first derivative of energy were appended to the parameters of each frame. The performance was evaluated for a speaker identification application using 12 mel-cepstral coefficients using an order for the cepstral coefficients of 14. The number of Gaussian mixtures varies from 1 to 32. We use covariance-tied models across all the experiments. We have done some tests using cepstral

[1] This work has been partially supported by the Spanish CICYT under the project TIC96-0956-C04-02

mean subtraction and preemphasis with a factor of 0.95 as well.

The recording work of TelVoice is still in progress, so in this experiment we use a subset of the database that contains 5 sessions uttered by 20 Spanish speakers (10 males and 10 females). The material we use from each session consists of 4 repetitions of the Spanish Identity Card number, made up of 8 digits (approximately 5 seconds each). The speakers were addressed to pronounce it naturally (digit by digit, grouping digits, or as a whole, as they usually do). The utterances recorded in one session were used for training and the other four were used for testing. The session used for training was rotated.

## 3. SYSTEM ARCHITECTURE

### 3.1. Acoustic segmentation

The phonetic classifier identifies the type of speech frame. In this implementation, we use a phonetic classifier that we have previously developed for speech coding purposes. It considers three distinct sound classes:

- Voiced sounds which have quasi-periodic waveforms and fairly harmonic spectra.
- Unvoiced frames which have aperiodic waveforms and irregular spectra; their energy is usually lower than that of voiced sounds.
- Transitions defined as the two first voiced frames after an unvoiced segment and the two last voiced frames before an unvoiced segment. This type of frames is characterized by a non stationary waveform.

One important point is that we also use a Voice Activity Detector (VAD) to identify the noise segments. The phonetic classifier uses an algorithm close to the one in [3] with some modifications to improve its behavior in noisy environments and to work in an Multiband Excitation Speech Coder [4].

The labeling of the training utterances is performed in a completely automatic way by this phonetic classifier. We train three HMMs per speaker with the frames corresponding to each phonetic class. All the non-voice material of the training session is used to train a noise HMM that is the same for all speakers.

### 3.2. The parallel system

When testing, we use a grammar for each speaker instead of using the phonetic classifier, as it can be seen in Fig. 2. This grammar allows any transition among the four HMMs (three for the voice and one for the noise) with equal probabilities, and the output probabilities are computed using the Viterbi algorithm for each type of voice segment. This means that the phonetic segmentation is embedded in the testing procedure. We accumulate all the output probabilities for the three possible phonetic classes. This way, it is possible to combine these output probabilities and to build up different decision rules.

In the experiments described in this section, we use two different configurations for the weighting factors:

- Equal factors. That is, we consider that importance across phonetic classes is the same.

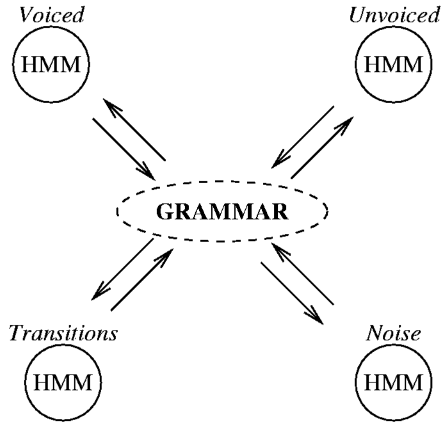

Fig. 1: The Parallel System

- Selecting factors. One factor is set to 1 and the other two are zero. This latter choice is very useful to study the relative importance of a particular phonetic class allowing also that different numbers of Gaussian mixtures in different states can be used.

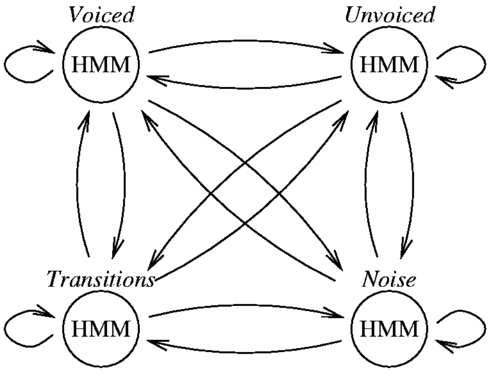

Fig. 2: The Ergodic Model

### 3.3. The Ergodic model

After doing some experiments with the Parallel System, we designed an Ergodic Model like the one shown in Fig. 2. As a starting point in the construction of this model, we take the four previous HMMs trained with the different acoustic segments. We combine these models into an ergodic one (all the transitions between states are allowed) and retrain this model with the

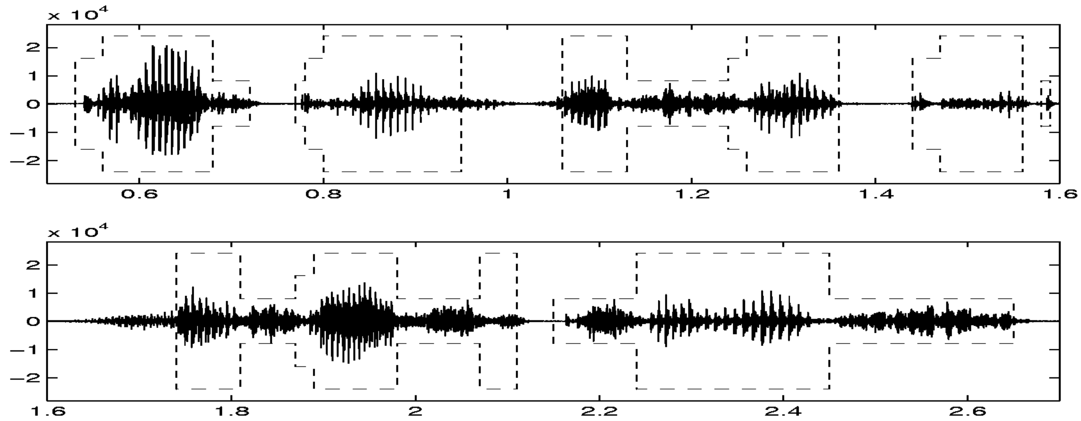

Fig. 4: Embedded segmentation in Parallel System (1 mixture per model)

restriction that only the transition probabilities can be modified.

With the purpose of studying the influence of the training of the transition probabilities in the final performance of the Ergodic Model, we also tested the same architecture with a totally free retraining. That means that, after building the Ergodic Model, we allow all the parameters of the model (transition probabilities, means and variances) to vary.

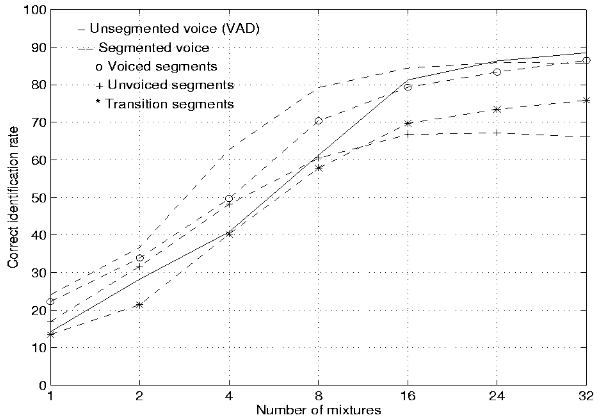

Fig. 3: Speaker Identification Rate - Parallel System

## 4. EXPERIMENTAL RESULTS

### 4.1. The parallel system

Fig. 3 shows the correct identification rate of our system using a cepstral calculation of 12. The solid line corresponds to the unsegmented voice (one HMM for the speech and another for the noise combined by a grammar) and the dashed lines to the segmented voice cases with equal factors and selecting factors. This figure shows that the relevant information regarding the identity of the speaker is mainly in the voiced part of the speech, particularly true when the number of Gaussian mixtures is low.

Fig. 4 presents the result of the embedded segmentation for a utterance during the testing phase with 1 Gaussian mixture per model. In this graphic, the height of the dashed line means the output of the phonetic segmentation: the maximum height corresponds to the voiced segments, the minimum to the unvoiced ones and the intermediate to the transitions; the chunks without dashed line are classified as noise. One important conclusion than can be derived from this figure is that the segmentation is quite correct with a number as low as one Gaussian mixture per model, in spite of the fact that the identification rate strongly depends on the number of mixtures.

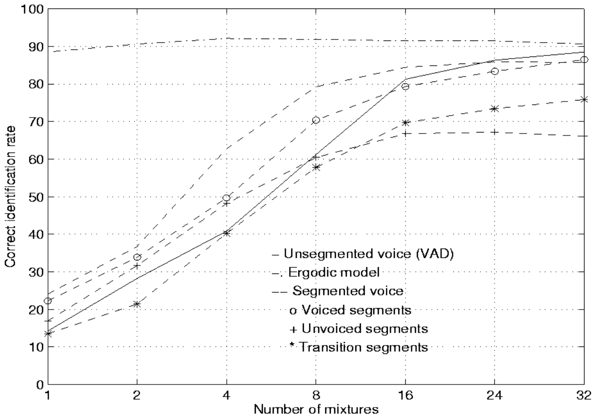

Fig. 5 : Speaker Identification Rate - Ergodic Model

### 4.2 The ergodic model

The resulting system experiments a dramatic improvement in the identification rate, as it can be seen in fig. 5. In this figure, the top line corresponds to the new model and the rest of the lines correspond to the Parallel System and to the system without Acoustic Segmentation.

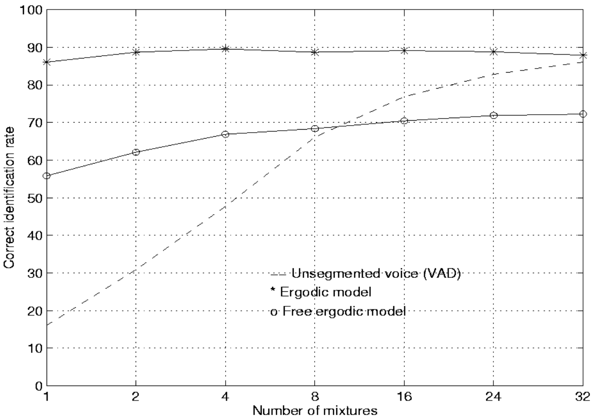

Fig. 6 : Speaker Identification Rate (Ergodic Models)

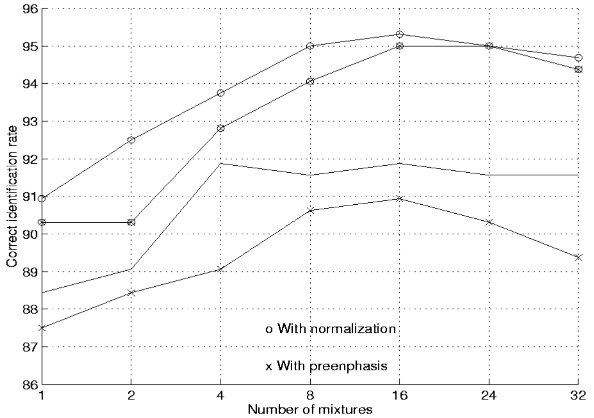

Fig. 7: Ergodic Model with Cepstral Mean Substraction and Pre-emphasis (k=0.95)

With the purpose of studying the influence of the training of the transition probabilities in the final performance of the Ergodic Model, we also tested the same architecture with a totally free retraining. That means that, after building the Ergodic Model, we allow all parameters of the model (transition probabilities, means and variances) to vary. The recognition rate obtained experiments a decrease between 15 and 30%, as it can be seen in fig. 6.

In fig. 7 the Identification Rate of the Ergodic Model applying cepstral mean substraction (CMS), pre-emphasis (k=0.95) and both is shown. In this case there is no rotation of the sessions and only the first session is used for training. It can be observed that the CMS improves the performance of the system, and that an additional preemphasis makes the results a little worse. Is important to have in mind that although the maximum difference in fig. 7 is about 4%, that means a reduction in the identification error about 40%.

## 5. CONCLUSIONS

The main conclusion that can be extracted from the Parallel System is that the voiced part of the speech plays a major role in the speaker identification task, although the identification score is far below requirements. The way of improving the recognition rate of this system would be to calculate weighting factors to be applied to the output probabilities of each of the models. Instead of this, we build up an Ergodic Model in which the transition probabilities are in charge of the weighting factors among classes. An advantage of this approach resides in its automatic procedure, along with its great flexibility.

In [1] it is stated that for HMM based systems, identification scores are highly correlated with the total number of mixtures independently of the number of states. Our results show that the performance of the Ergodic Model is very high even with a number of mixtures per state as low as one, an it maintains approximately constant independently of the number of mixtures. The explanation for this behavior can be derived from fig. 4: with one mixture per state the representation of the general characteristics of the phonetic classes is accurate enough. When we increase the number of mixtures, we are improving the representation of the boundaries between these phonetic classes and the overall effect is that the recognition rate increases. When we build an Ergodic Model and retrain the transition probabilities, we are introducing this information in the model in a different and much more efficient way. The high recognition rate of this architecture, along with its low computational load, encourages us to think that it can be a suitable choice for a real-world application as a speaker verification system over the telephone line.

## 6. REFERENCES

[1] T. Matsui, S. Furui, *Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's*, IEEE Trans. on Speech and Audio Processing, vol.2, No. 3, July. 1994, pp. 456-459.

[2] D. A. Reynolds. *Speaker Identification and Verification using Gaussian Mixture Speaker Models*, Speech Communication, vol. 17, August 1995. Pp. 91-108.

[3] R. Tucker, *Voice Activity Detection Using a Periodicity Measure,* IEEE Proceedings-I, vol.139, August 1992.

[4] C. García Mateo, D. Docampo Amoedo, *Modeling Techniques for Speech Coding: a Selected Survey*, Chapter in the book: "Digital Signal Processing in Telecommunications" edited by Professor Figueiras-Vidal. Published by Springer Verlag in Spring 1996.