Speaker Recognition Based on a Weighted Acoustic Discrimination

Carmen García-Mateo, Leandro Rodríguez Liñares
Departamento de Tecnologías de las Comunicaciones
Universidad de Vigo, Spain
Phone:34-86-812133, Fax:34-86-812116
e-mail:carmen@tsc.uvigo.es, leandro@tsc.uvigo.es

ABSTRACT

We combine multiple-mixture single-state Markov models with phonetic classification in order to improve the performance of a speaker recognition system. Three broad phonetic classes: voiced frames, unvoiced frames and transitions, are defined. We design speaker templates by the parallel connection of the weighted outputs of three single state HMM's. Each model corresponds with a distinct sound class and the output weights take into account the perceptual influences across phonetic classes. The preliminary results show that this novel architecture outperforms its counterpart without phonetic classification.

1 Introduction

Automatic speaker recognition consists in identifying one speaker from a population by processing a speech utterance. The speech is processed and compared against one or some stored speaker templates.The block diagram of the general architecture of such a system is shown in Figure 1.

Hidden Markov Models (HMM) are considered a suitable algorithmic choice for speaker recognition. In recent years, a number of experiments have been conducted aiming to explore their capability and performance to track the identity of the speakers [1] [2] [3]. The starting point of our work is the results presented in [1] and [2] and our previous work in designing phonetic-based voice classifiers for speech coders.

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. Thus a single state model with 64 mixtures performs equivalently as a four-state model with 16 mixtures each. The important point is to decide how many mixtures must be used to achieve a good compromise between, in one hand, representation accurateness and, in the other, the amount of data required for trainning (directly connected with enrollment duration time) and computational complexity.

In [2] they used the fact that since 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.

We use this latter idea in designing speaker templates by the parallel connection of the weighted outputs of three single state HMM's. Each model corresponds to a distinct sound class and the output weights take into account the perceptual differences across phonetic classes. Our main objective is to investigate how much the number of Gaussian mixtures can be reduced if phonetic classification is performed.

The rest of the paper is organized as follows: Section 2 presents the system architecture along with its main features. In Section 3 we describe the conditions of the conducted experiments. Section 2 shows the results we are currently obtaining. Finally, Section 5 presents some conclusions and guidelines for future work.

2 System Overview

The general architecture of our speaker recognition system is depicted in Figure 2. We distinguish three main parts: 1) the phonetic classifier, 2) the bank of HMM's, and 3) the output weighting block. The main features of each subsystem are:

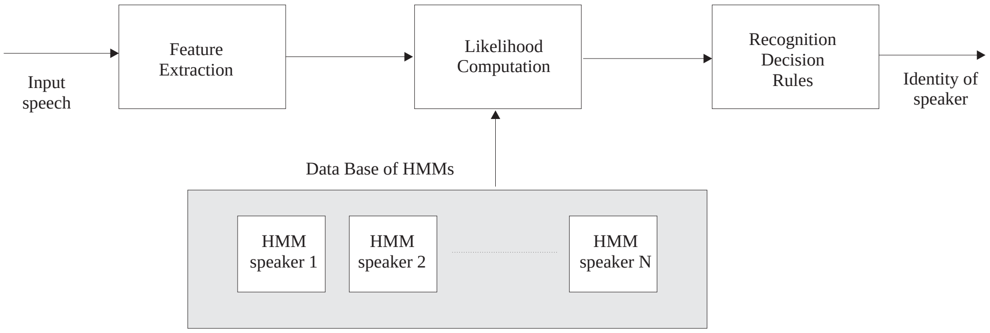

Figure 1: Speaker recognition general architecture

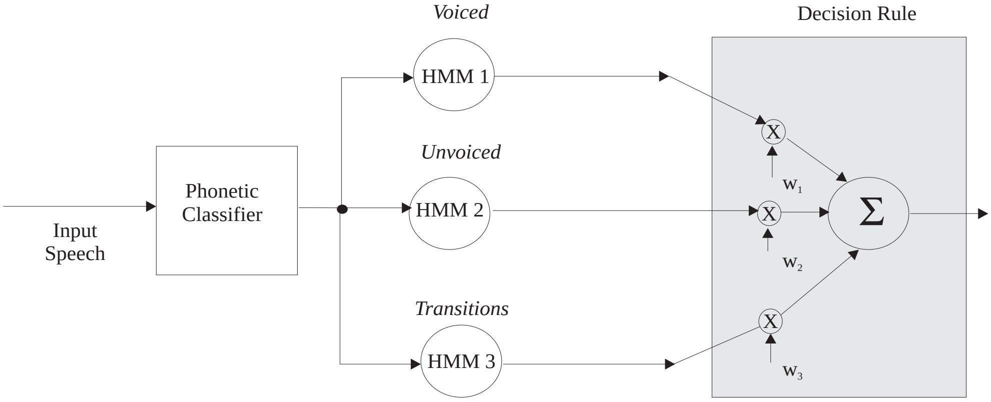

Figure 2: Acoustic Weighted HMM Configuration

1) The phonetic classifier identifies the type of speech frame. In this first 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 latter type of frames is characterized by a non stationary waveform.

One important point is that we also use a Voice Activity Detector (VAD) to eliminate the non-speech segments.

The phonetic classifier uses an algorithm close to the one in [4] with some modifications to improve its behavior in noisy environments and to work in an Multiband Excitation Speech Coder [5].

It is widely known that the performance of a phonetic classifier strongly relies on the number of classes and on the threshold policy. For coding purposes, a phonetic classifier is always forced to make a decision leading sometimes to miss-classifications. However, when applying a phonetic classification strategy to speaker recognition we can avoid to make a hard decision and instead to use a soft decision rule. Therefore, we discard those *difficult to classify* frames and we update the template and test the system just with the beyond any doubt correctly classified frames.

2) The bank of HMM's. For each type of sound, a multiple-mixture single-state HMM is built up. We use mel-cepstrum and Δ-mel-cepstrum coefficients, together with the energy and its first derivative. These models work in parallel, being fed by their corresponding types of frames.

3) The output weighting block. Its objective is to weight the output of each HMM before to compute the final score. This implies to have a more flexible structure compared with the mere addition of the HMM outputs. How to select the suitable weighting factors is an open problem and some heuristic

considerations would be taken into account.

3 Experimental Conditions

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.

Thus, having in mind future telephone applications, we use 3.4 KHz band limited speech sampled at 8 KHz. Mel-cepstrum and Δ-mel-cepstrum coefficients are computed using a frame length of 20 ms, and a frame period of 10 ms. Energy and the first derivative of energy are appended to the parameters of each of the frames.

The speakers Data Base consists of sessions of about 5 seconds each, uttered by 12 Spanish speakers (7 males and 5 females) recorded using a Shure SM-12 microphone under fairly good acoustic conditions. Up to now, we have recorded one session each two weeks up to a total of four sessions.

Therefore, this data base can be considered as a clean-speech band-limited data base. The collection of more sessions is still in progress being our target to acquire data from the speakers along a period of about six months.

The utterances pronounced by the speakers are the numbers of the Spanish Identity Card, that consist of eight digits. The utterance recorded in one session is used for training and the ones recorded in the three other sessions are used for testing. The session used for training is rotated.

The performance was evaluated for a speaker identification application using an order for the cepstral computation of 8 and 12. The number of Gaussian mixtures varies from 1 to 32. Provided that the utterances used for training and testing are rather short, we use covariance-tied models across all the experiments.

For the training utterances, the labeling of the different segments (voiced, unvoiced and transitions) are performed in a completely automatic way using the phonetic classifier. We train each HMM with the frames corresponding to each phonetic class.

When testing, we use a grammar for each of the speakers instead of using the phonetic classifier. This grammar allows all the transitions between the three HMMs with equal probabilities, and the output probabilities are computed using the Viterbi algorithm for each type of segments. This means that the phonetic segmentation is emmbedded in the testing procedure. We accumulate all the output probabilities for the three possible phonetic classes. This way, is possible to combine this output probabilities and to build up different decision rules, as can be seen in Figure 2.

In the experiments described in this paper, we used two different configurations for the weighting factors:

1. Equal factors. That is, to consider that the importance of the phonetic classes is the same.
2. Selecting factors. One of the factors is one and the other two are zero. This latter choice is very useful to study the relative importance of phonetic classes allowing also that distinct numbers of Gaussian mixtures are used.

4 Preliminary results

Figures 3 and 4 show respectively the correct identification rate of our system using cepstral calculation orders of 8 and 12. These results can also be seen in the table 1 for the unsegmented voice and the voiced segments.

These results show that the relevant information in order to identify the speaker of a certain utterance is mainly in the voiced part of the speech. This is particularly true when the number of Gaussian mixtures is low. The importance of the unvoiced segments and the transitions is lower than the one of the voiced segments and it seems that the transitions are more important than the unvoiced segments, especially when the number of mixtures is greater than 8.

5 Conclusions and Further Work

Our system is still under development, but in spite of that some facts must already be mentioned. The main conclusion is that the voiced part of the speech plays a major role for the speaker identification task. The results presented in the previous section encourage us to going on with this scheme trying to improve it.

Further experiments must be carried out in order to establish the best configuration for the weighting block. We plan to use an ergodic Hidden Markov Model with three states. Each phonetic class has its own state. The transition probabilities will be in charge of the weighting factors among classes. The

Num. of mixtures	1	2	4	8	16	24	32
8th order unsegmented	13.0	30.6	31.5	63.9	84.3	88.0	96.3
8th order voiced segments	23.1	45.4	56.5	68.5	87.3	90.3	90.7
12th order unsegmented	12.5	25.0	34.7	61.8	81.3	91.0	94.4
12th order voiced segments	24.3	37.5	50.0	68.1	85.1	89.1	91.1

Table 1: Speaker Identification Rates

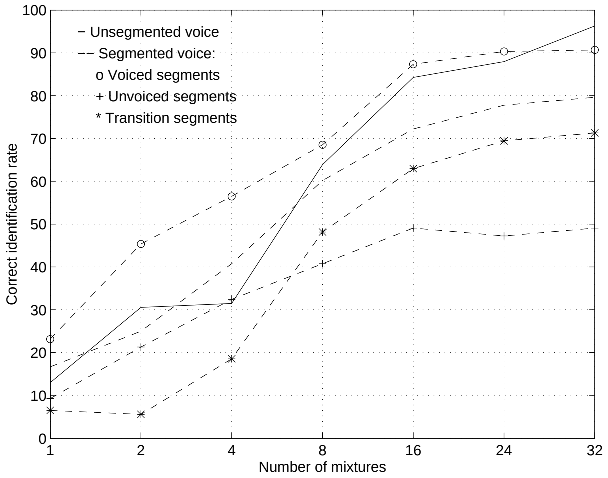

Figure 3: Speaker Identification Rate (order 8)

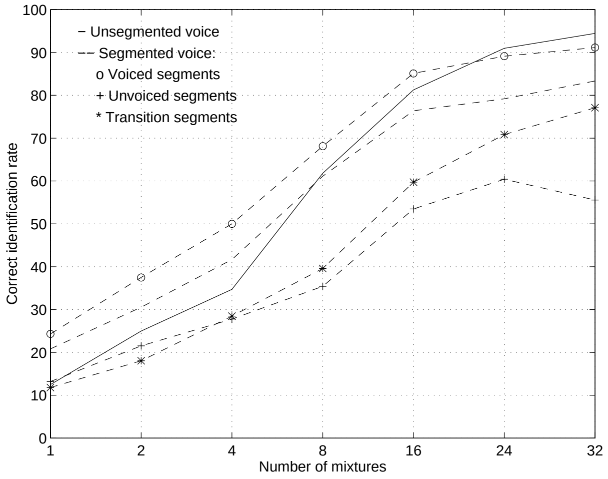

Figure 4: Speaker Identification Rate (order 12)

training procedure will be: first, each state will be trained separately as in the current system; second, only the transition probabilities in the ergodic model are computed using a Baum-Welch reestimation algorithm. In the testing phase, no phonetic classifier is used allowing that the hidden Markov system classifies the speech. An advantage of this approach resides in its automatic procedure for selecting the weighting factors, along with its greater flexibility.

References

[1] T. Matsui and 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] M. Savic and S. K. Gupta, "Variable Parameter Speaker Verification System Based on Hidden Markov Modeling"; *Proc. ICASSP*, 1990, pp. 281–284.

[3] R. C. Rose and D. A. Reynolds, "Coding of Wideband SpeechText Independent Speaker Identification Using Automatic Acoustic Segmentation"; *Proc. ICASSP*, 1990, pp. 293–296.

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

[5] C. Garcia-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.