APPLICATION OF SEVERAL CHANNEL AND NOISE COMPENSATION TECHNIQUES FOR ROBUST SPEAKER RECOGNITION*

L. Docío-Fernández and C. García-Mateo
E.T.S.I. Telecomunicación
Communication Technologies Dept.
University of Vigo, 36200 Vigo, Spain.
Tel. +34 86 812 664, FAX: +34 86 812 116, E-mail: ldocio@tsc.uvigo.es

ABSTRACT

This paper is concerned with the problem of Robust Speaker Recognition. An acoustical mismatch between training and testing conditions of hidden Markov model (HMM)-based speaker recognition systems often causes a severe degradation in the recognition performance. In telephone speaker recognition, for example, undesirable signal components due to ambient noise and channel distortion, as well as due to different variations of telephone handsets render the recognizer unusable for real-world applications. The purpose of this paper is to present several compensation techniques to decrease or to remove the mismatch between training and testing environment conditions. Some of the techniques described here have already been successfully applied in Robust Speech Recognition, and our preliminary results show that they are also very encouraging for Speaker Recognition.

1. INTRODUCTION

It is well known that Automatic Speaker and Speech Recognition systems trained in one environment often perform poorly in new environments due to mismatches between training and testing conditions. This is particularly true for Hidden Markov Models (HMM) recognizers. The mismatches could be due to different transducers, transmission channels, changing speaking styles and accents, the presence of varying ambient and channel noise, etc. The goal of robust speaker recognition is to remove the effect of this mismatch so as to bring the recognition performance as close as possible to that of the matched conditions.

Consider the situation of a speech signal transmitted over a telephone network where the distortion effect is assumed to be linear either in the spectral domain or in the cepstral domain. These effects constitute an additive component, **N**, which is representative of the ambient noise, and a multiplicative component, **H**, due to the filtering effect of the channel. This is simply expressed as:

$$\mathbf{Y} = \mathbf{H}[\mathbf{X} + \mathbf{N}]$$

where **X** is the original clean speech, **H** is the channel response and **Y** is the received signal. In the log domain

$$\log \mathbf{Y} = \log \mathbf{H} + \log[\mathbf{X} + \mathbf{N}]$$

Then the channel influence on the speech leads to an additive component on the cepstrum of the speech. That is, in the cepstral domain the relation can be expressed as:

$$\mathbf{c}_y[k] = \mathbf{c}_h[k] + \mathbf{c}_{xn}[k]$$

where $\mathbf{c}_y$[k] denotes the kth cepstral coefficient of **Y**, $\mathbf{c}_h$[k] denotes the kth cepstral coefficient of **H**, and $\mathbf{c}_{xn}$[k] is the kth cepstral coefficient of (**X+N**).

Thus the mismatch between the training and testing conditions can be represented as a linear transform in the cepstral domain:

$$\mathbf{c'} = \mathbf{A} * \mathbf{c} + \mathbf{b},$$

where **c'** represents the cepstral vector of the distorted signal, **c** the cepstral vector of the original signal, the matrix **A** and vector **b** model the transformation.

Long-term Cepstral Mean Subtraction (CMS) can be considered as a classical technique for reducing the channel influence both in speech and speaker recognition. Following the prior formulation, it assumes **A** as the identity matrix and estimates **b**. Recently, several maximum likelihood (ML) techniques for estimating the above transformation have been introduced to deal with the problem of robust speech recognition. Among them we can mention the techniques described in [1] and [2] as effective ways for handling the mismatch.

The purpose of this paper is to evaluate the performance of these techniques when applied to a speaker recognition system and compared them with CMS.

The remainder of the paper is organized as follows. Next, we describe our baseline speaker recognition system as well as the signal analysis and the database that we will use for the experiments. Section 3 describes the two stochastic methods for signal bias removal we have explored. In Section 4, experimental results and comparative performance among the different techniques are shown. A summary and conclusions are given in section 5.

2. BASELINE ERGODIC HMM RECOGNIZER SYSTEM

The baseline recognition system that we have used to explore the performance of the implemented techniques can be described as follow.

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

Speakers are represented by an ergodic four-state continuous density and multiple-Gaussian HMM trained by using an acoustic discriminative procedure [3]. Briefly, this training procedure consists of two stages: first, training frames are a priori segmented into four categories: voiced, unvoiced, transitions, and non-speech frames by means of a phonetic classifier. Then, all frames assigned to a particular category contribute to the estimation of the mixture Gaussian parameters of a particular state. Variances are tied across the mixtures due to the scarcity of training data. Second, transition probabilities are estimated holding fixed means and variances of the Gaussian mixtures.

In all experiments a database from 20 Spanish speakers (10 males and 10 females) recorded over the telephone network has been used. Each speaker provided five sessions collected over a period of about one month. The speakers were encouraged to use different handsets and telephone lines. Each session consists of four repetitions of the speaker Spanish Identity Card. One session is used for training and the remainder four sessions for testing. Training session is rotated across the available sessions in order to maximize the number of testing tokens. In this way, we achieve 1600 testing utterances and five assessment sets.

The speech input is sampled at 8 KHz. Then, a mel-scaled cepstral analysis is performed each 10 ms time interval with a 20 ms Hamming window. For each frame we extract a 18 element feature vector, which consists of 8 mel-scaled cepstral coefficients, 8 delta-cepstrum coefficients, the normalized log energy, and the delta normalized log energy.

3. STOCHASTIC MATCHING (SM) TECHNIQUES

We have tried three different techniques for handling the mismatch between training and testing conditions: long-term Cepstral mean subtraction (**CMS**) [4], Signal Bias Removal (**SBR**) by ML estimation as described in [1], we will call it Stochastic Method 1 (**SM1**) and the ML Stochastic Matching described in [2] that we will refer to as **SM2**.

3.1 CMS

CMS is a very simple method that subtracts from each frame of the observed utterance the average cepstrum over the entire utterance

$$\mathbf{c}' = \mathbf{c} - E[\mathbf{c}]$$

where the expectation, E[**c**], is approximated by time average.

It regards the average of speech cepstra as the channel multiplicative distortion, and has been proposed and applied as a simple and powerful adaptation technique to telephone speech recognition systems in that training is done on one channel condition while testing is done on another channel condition. CMS performance can be considered as the reference objective for any novel technique.

3.2 SM1

This method was originally applied to robust speech recognition by Rahim and Juang [1]. It is carried out as an independent process following feature analysis and proceeding with a HMM recognition, and has been integrated as part of a discrete density HMM recognition system. Thus, the approach use a vector quantization (VQ) codebook based model to estimate the bias process. As we have mentioned in section 2 our baseline recognition system uses a continuous density HMM architecture, then in order to apply this technique to our system, we compute the SBR VQ and store it separately from the model.

We have assumed a simple additive bias term **b** and the estimation is accomplished in the training phase by the following iterative procedure.

Given a set of centroids μ_i,

1. We compute an estimation of the bias b for each utterance as

   $$\mathbf{b} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_t - \mathbf{z}_t$$

   where $\mathbf{z}_t$ is the *nearest neighbor* to the distorted signal $\mathbf{y}_t$

   $$\mathbf{z}_t = \arg\min d(\mathbf{y}_t, \mu_i)$$

   After that **b** is subtracted from the signal

   $$\mathbf{x}'_t = \mathbf{y}_t - \mathbf{b}_t \qquad t = 1,\ldots,T$$

   This procedure is iterated several times, using $\mathbf{x}'_t$ rather $\mathbf{y}_t$, to reduce the bias and to obtain a maximization of the likelihood function.
2. We generate a new set of centroids with the processed data $\mathbf{x}'_t$.

The above procedure (steps 1 and 2) is repeated with the new improved set of centroids until the likelihood reaches a fixed point.

During the recognition phase only the step 1 is implemented.

We can see that there is a strong relationship between SM1 and CMS. The key issue of this method is the VQ required to compute the bias estimation.

3.3 SM2

It is also a maximum likelihood (ML) approach to decrease the acoustic mismatch between a test utterance and a given set of HMM's. This mismatch can be reduced in two domains: in the observation space mapping the observed utterance **Y** into an utterance **X** which matches better with the models, and in the model space mapping the original models to a transformed models that matched better with the observed utterance **Y**. The parameters of the functions that implement these

mappings are estimated using the expectation-maximization (EM) algorithm.

The algorithm works only on the given test data and the given set of speaker models, and no additional training data is required for the estimation of the mismatch prior to actual testing. The bias can be modeled as either a fixed bias or a state-dependent bias. In some situations a state-dependent bias is meaningful. As an example, an additive cepstral bias model for linear filtering is only valid for high signal to noise ratios (SNR's). When the SNR is low, the noise dominates, and the additive model for channel filtering is inaccurate. It is possible to estimate a separate bias for speech and non-speech segments. Recalling the architecture of our HMM's (it considers different states for different kinds of speech segments), a very simple way to deal with separate bias for each kind of sound is to use the approach that considers a state-dependent bias.

The iterative estimation procedure is based in the following two steps:

1. We first find the most likely state sequence using the Viterbi algorithm.
2. Then we find the bias to maximize the likelihood of the utterance **Y** conditioned on this state sequence using the EM algorithm.

It is important to note that the recognition hypothesis guides the algorithm and, hence, a very poor hypothesis can result in suboptimal performance. For a detailed description of the formulation of this approach in both, the observation space and the model space see [2].

In the observation space we obtain the new utterance as

$$\mathbf{x}_t = \mathbf{y}_t - \mathbf{b}_t$$

In the model domain we assume that the statistical model for the bias is a single Gaussian density with diagonal covariance matrix. Thus, the structure of the new model remains the same as that of the given model. Means and variances of the new model are derived as follows

$$\mu_y = \mu_x + \mu_b$$

$$\sigma_y^2 = \sigma_x^2 + \sigma_b^2$$

where μ_b and σ^2_b are the parameters of the bias model.

4. EXPERIMENTAL RESULTS

Closed-set identification experiments were conducted. The goal of the experiments was to examine the efficacy of the above approaches in improving the performance of the baseline recognition system in the presence of mismatch due to different transducers and channels.

We first carried out baseline experiments to study the effect of the HMM topology. The overall system performance strongly relies on the total number of Gaussian mixtures and so it does the computational complexity. Figure 1 shows how gracefully the performance of the baseline speaker recognition system, plotted with solid line, increases with the number of mixtures per state in a speaker identification experiment. It also shows the performance when CMS is applied. In all cases, performance improves. For example, we notice an improvement by 33% compared with the 4-mixtures baseline system.

As we said before, performance of SM1 techniques depends on the size and design procedure of the required VQ. We have tried two approaches for VQ design, always using the Generalized Lloyd Algorithm (GLA) to obtain the centroids. In the first approach a VQ is built up using training data from all speakers. This VQ is used both for training and testing. In the second approach, a VQ for each speaker is individually computed and used for training the corresponding HMM. In testing, a generic VQ formed by concatenation of the individuals VQ's is used. In both approaches three or four iterations are enough for the algorithm to converge. The performance of the first and second approaches are also shown in Figure 1 denoted as SM1-1 and SM1-2 respectively. In this example a VQ of 120 centroids has been used in the first approach. In the second a VQ of 4 centroids for each speaker plus a VQ of 6 centroids for the noise what makes a generic VQ in testing of 86 centroids. We can observe as the second approach has a better performance than the first one. It also slightly outperforms the CMS technique. These results were obtained using one of the five sessions as the training data and the others as test data and then rotating the order of them to came up with five assessment sets. We report the averages of the five assessments.

In the case of a *speaker verification* experiment the generic VQ is not required. Our preliminary results in this direction are also shown in Figure 1 (SM1-3). In this experiment each speaker VQ consists of 8 centroids. We see how big the improvement is.

As we said in section 3.3 the SM2 technique operates entirely on the test utterance and the speaker HMMs. In our experiments we have considered the two approaches: a single bias vector for the whole utterance and different bias vectors for different states of the HMM. These two approaches have been implemented both in the feature and model space. The bias parameters were estimated on a per-utterance basis. In the feature space the bias vector was initialized to zero. In the model space, the mean of the bias was initialized to zero, whereas the variance was initialized to a small positive number.

We first conducted an experiment in a hypothetical situation in which we suppose known the identity of the speaker. This allows to choose the model that we will use to estimate the bias. After removing the bias the identification is carried out using all speaker models. We can see this procedure as a mixture between *verification* and *identification* tasks. The Figure 2 shows the results for different HMMs topologies and

approaches to estimate the bias. The results correspond to train with a session and to recognize with the other four sessions. It can be observed that both a feature-space single bias (FS-1) and a state-dependent bias (FS-2) outperform the CMS technique. Furthermore, FS-2 is superior to FS-1 in this experiment.

In order to implement the technique on a *identification speaker task* we need to specify the model or models used to find the most likely state sequence. To handle with this problem we have explored three possible alternatives:

1. Given a test utterance the set of all the HMMs in the speaker population is used to segment it and to estimate the bias process (SOL1).
2. First, we find the speaker HMM that best matches the utterance. Second, we use his/her HMM and a set of close cohorts HMMs to compute the bias as in 1. (SOL2).
3. We estimate as many possible bias process as speakers are enrolled in the system. That is, for each speaker model, first we estimate the bias using the corresponding speaker model, second, we remove the bias and third we compute the likelihood of the utterance supposed it belongs to this speaker.. Thus, we obtain as many likelihood values as speakers enrolled in the system. Finally, the utterance will be identified as belonging to the speaker model that obtained the maximum likelihood. (SOL3).

In Table 1 some relevant results from these approaches are shown. These results are for a HMM architecture with 4 and 8 Gaussian mixtures per state (first and second row respectively). Also two different approaches for the bias are considered: 1) a *single* bias vector is estimated for the entire utterance (FS-1); and 2) a *separate* state-dependent vector is estimated (FS-2). As a reference, we also give both the baseline system performance (Base) and the CMS performance. The results show that these techniques significantly reduce the error rate compared with the baseline system, but they do not improve the CMS performance. Among them SOL3 seems to be the most effective technique.

5. SUMMARY AND CONCLUSIONS

We have presented our experiments with two stochastic techniques to increase performance in a robust speaker identification task by dealing with the mismatch between training and testing conditions. They have proved to be effective and, in some case, show better improvement when compared with CMS.

Regarding complexity, CMS is a fast and efficient technique. It is simpler and faster than the SM techniques proposed, since it requires only a small amount of computation in the front-end of the system. By the contrary the SM proposed techniques are based in iterative algorithms, what makes these methods heavy from the point of view of computational burden.

We believe that further improvement can be achieved in a verification task as our preliminary results show. This will be the next step to be conducted. Also we are studying the option of estimating a separate bias vector for speech and silence frames.

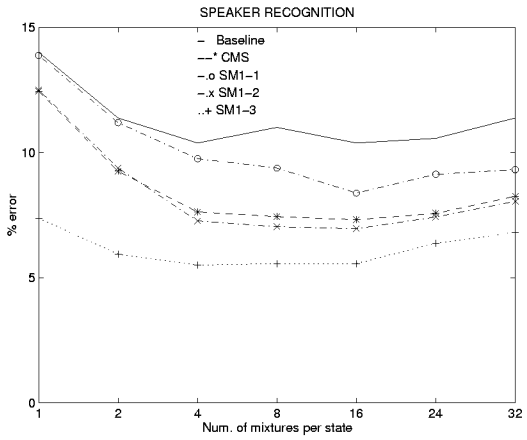

Figure 1. Speaker error rate with the method SM1

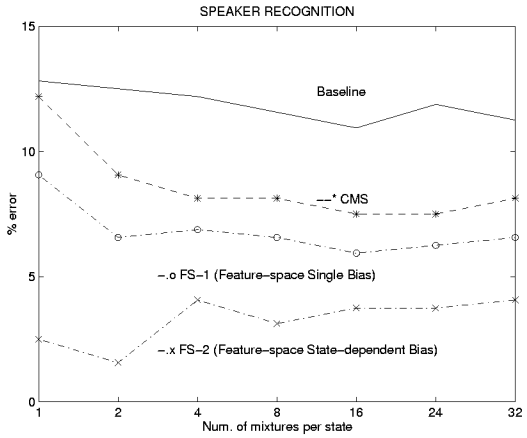

Figure 2. Speaker error rate with the method SM2

Table 1. Speaker error rate with SM2 for 4 and 8 mixtures

BASE	CMS	SOL1		SOL2		SOL3	
		FS1	FS2	FS1	FS2	FS1	FS2
12.2	8.1	10.9	11.6	11.56	11.6	9.4	8.7
11.6	8.1	8.8	10.0	9.37	10.9	8.3	7.5

6. REFERENCES

[1] M.G. Rahim and B.-H. Juang.; "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition."; *IEEE Trans. on ASP*, 1(4):19-30, January 1996.

[2] A. Sankar and C.-H. Lee.; "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition."; *IEEE Trans. on ASP*, Vol. 4, No. 3, pages 190-202, May 1996.

[3] L. Rodríguez-Liñares and C. García-Mateo; "On the use of acoustic segmentation in speaker identification."; In Eurospeech 97, Greece. September 1997.

[4] A.E. Rosenberg, C.H. Lee, F.K. Soong, "Cepstral Channel Normalization Techniques for HMM Based Speaker Verification", Proc. ICSLP'94, pp. 1835-1838, Yokohama (Japan), 1994.