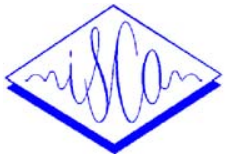

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Toward Parametric Representation of Speech for Speaker Recognition Systems

Rivarol Vergin, Douglas O'Shaughnessy and Pierre Dumouchel

Centre de Recherche Informatique de Montreal
1801, Ave. McGill, Montreal, Canada
INRS-Télécommunications, 16 Place du Commerce,
Île-des-Sœurs, H3E-1H6, Québec, Canada
email: vergin@inrs-telecom.uquebec.ca

10.21437/Eurospeech.1999-193

Abstract

The front-end used in many speaker recognition systems extracts, from the input speech signal, a set of coefficients based on a mel-cepstrum technique. This paper addresses the problem of efficiency of mel-cepstrum coefficients in a speaker recognition system and suggests a technique permitting an appropriate choice of these coefficients. It is shown, by the results obtained, that this technique can significantly increase the performance of a speaker recognition system.

1 INTRODUCTION

Automatic speaker recognition involves first a speech analysis process whose role is to extract from the input speech signal a set of feature vectors which reflects a person's vocal-tract structure. The classical approach used in many systems consists of evaluating a set of input coefficients over a 30-ms window. The coefficient vectors are updated each 10 ms and are generally calculated according to a mel-cepstrum technique [1].

An important application of automatic speaker recognition concerns the possibility of verifying a person's identity prior to admission to a secure facility or to a transaction over the telephone. It follows that the data used to train and test speaker recognition systems are collected through the telephone network. Because the channel involved in the communication can vary from call to call, there is often an acoustic mismatch between the data collected to train the speaker models and the data used during the testing process, which affects the performance of many speaker recognition systems.

Many algorithms, addressing the mismatch problem, have been proposed in the literature. Some, known as channel compensation methods [2], are based on variance transformation techniques to increase the acoustic coverage of the speaker models. The cepstral mean substraction technique [3] is often used to reduce the mismatch effects. This contribution deals with coefficient evaluation techniques that are less sensitive to channel effects.

Previous work has been conducted [4]in extracting speech features for speaker recognition. The use of formants or fundamental frequency as input parameters have been explored. This paper concentrates on a technique to evaluate the performance of input coefficients used in speaker recognition according to Gaussian Mixture Model (GMM) and suggests a method to calculate new coefficients that are more robust to channel effects.

The rest of this paper is organized as follows. Section 2 gives an overview of the Gaussian mixture models, section 3 presents the algorithm for coefficient performance measurement, section 4 explains the coefficient evaluation technique, section 5 describes the test procedures and presents some comparative results between different systems and section 6 summarizes this contribution.

2 GAUSSIAN MIXTURE MODEL

Unlike the clear correlation between phonemes and spectral resonances, there are no acoustic cues specifically or exclusively dealing with speaker identity. Most of the parameters and features used in speech analysis contain information useful for the identification of both the speaker and the spoken message. Indeed a mel-cepstral feature representation is often used as well in speech as in speaker recognition systems.

The two types of information, however, are coded quite differently. In a speech recognition system, decisions are made for every phone or word; a speaker recognition system requires only one decision, based

on parts or all of a test utterance. One of the most common methods used in text-independent cases, where training and testing involve different phrases, is GMM. According to this approach, each speaker is represented by a model

$$\lambda = \{p_m, \mu_m, \Sigma_m\}, \quad m = 1, \cdots, M, \tag{1}$$

where M is the number of component densities of the form:

$$b_m(x) = \frac{(2\pi)^{-D/2}}{|\Sigma_m|^{1/2}} exp(-\frac{1}{2}(x-\mu_m)^{'}\Sigma_m^{-1}(x-\mu_m)). \tag{2}$$

μ_m and Σ_m are respectively mean vector and covariance matrix and x is a feature vector of dimension D. The Gaussian mixture density is given by:

$$p(x|\lambda) = \sum_{m=1}^{M} p_m b_m(x). \tag{3}$$

p_m are the mixture weights satisfying the constraint that $\sum p_m = 1$.

The first motivation for using Gaussian mixture densities as a representation of speaker identity is the intuitive notion that the individual component densities of a multi-modal density may model some underlying set of acoustic classes.

Given a set, X, of training feature vectors for a speaker, the estimation of the model parameters, λ, is generally performed using the EM algorithm [4]. This algorithm can be summarized as follows. The process begins with an initial model λ; a new model $\lambda^{'}$ is estimated such that $p(X|\lambda^{'}) \geq p(X|\lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. Once the training step has been completed, the automatic speaker identification can take place.

The identification process requires choosing which of the N speakers known to the system best matches a given set of feature vectors, x_t, of dimension T. The objective is then to find the speaker model which has the maximum a posteriori probability for a given observation sequence, that is, speaker n will be identified if

$$p(\lambda_n|X) > p(\lambda_k|X), \quad \forall k \neq n. \tag{4}$$

Assuming that speakers are equally likely and observation vectors, x_t, are statistically independent, it can be shown that the rule of decision consists of associating speaker n to the test voice if:

$$\sum_{t=1}^{T} \log p(x_t|\lambda_n) > \sum_{t=1}^{T} \log p(x_t|\lambda_k), \quad \forall k \neq n. \tag{5}$$

In this equation, x_t is a vector whose elements are input coefficients evaluated at time t at the front-end of the system. Section 3 suggests a technique to evaluate the contribution of each coefficient in the speaker recognition process.

3 Coefficient Performance

The Gaussian Mixture Model is recognised as one of the most accurate models for automatic speaker recognition over telephone lines. According to this model the probability of observing x given a speaker model λ as defined by equation 3 is an arithmetic mean of a set of gaussian densities. Let us define $p^{'}(x|\lambda)$ as a geometric mean of the same gaussian densities, that is,

$$p^{'}(x|\lambda) = \prod_{m=1}^{M} (b_m(x))^{p_m}, \tag{6}$$

replacing b_m by its value, equation 2, in the previous equation and neglecting the term $(2\pi)^{D/2}$, $p^{'}(x|\lambda)$ is equal to

$$\prod_{m=1}^{M} (\frac{1}{|\Sigma_m|^{1/2}} exp(-\frac{1}{2}(x-\mu_m)^{'}\Sigma_m^{-1}(x-\mu_m)))^{p_m}. \tag{7}$$

Assuming that Σ_m is a diagonal covariance, we obtain for $p^{'}(x|\lambda)$ the following expression:

$$\prod_{m=1}^{M} exp(-\frac{1}{2} p_m (\sum_{i=1}^{D} \frac{(x_i - \mu_{m,i})^2}{\sigma_{i,m}^2} + 2\ln(\sigma_{i,m}))) \tag{8}$$

where x_i is the ith coefficient appearing in input vector parameter of dimension D. Taking the logarithm on both sides of this equation, we obtain:

$$\ln p^{'}(x|\lambda) = -\frac{1}{2} \sum_{i=1}^{D} \sum_{m=1}^{M} p_m (\frac{(x_i - \mu_{m,i})^2}{\sigma_{i,m}^2} + 2\ln(\sigma_{i,m})) \tag{9}$$

Defining α_i as

$$\alpha_i = \sum_{m=1}^{M} p_m (\frac{(x_i - \mu_{m,i})^2}{\sigma_{i,m}^2} + 2\ln(\sigma_{i,m})) \tag{10}$$

we obtain

$$\ln p^{'}(x|\lambda) = -\frac{1}{2} \sum_{i=1}^{D} \alpha_i. \tag{11}$$

It is important to realize that α_i is a function of the input coefficient x_i, calculated at a given time t, and

for a speaker model λ. Based on equation 5 we can redefine the rule of decision as follows and associate speaker n to the test voice if:

$$\sum_{i=1}^{D}\sum_{t=1}^{T}\alpha_{i,t,\lambda_n} < \sum_{i=1}^{D}\sum_{t=1}^{T}\alpha_{i,t,\lambda_k}, \quad \forall k \neq n. \tag{12}$$

Defining β_{i,λ_n} as

$$\beta_{i,\lambda_n} = \sum_{t=1}^{T}\alpha_{i,t,\lambda_n} \tag{13}$$

we obtain for the rule of decision in the following equation:

$$\sum_{i=1}^{D}\beta_{i,\lambda_n} < \sum_{i=1}^{D}\beta_{i,\lambda_k}, \quad \forall k \neq n. \tag{14}$$

The chief feature of this last equation is its emphasis on the contribution of each input coefficient i in the speaker recognition process. Indeed, the identified speaker is the one for which the sum of the elements over the entire set of coefficients is less than the term appearing on the right side of the equation. For some values of i the sense of the inequality, defined by equation 14, could not be respected.

In this contribution we have evaluated the efficiency of each coefficient in the recognition process according to the following procedure. For each speaker the same data used in the training phase is also used during the recognition phase, hence the correct speaker is always recognized. We consider a coefficient i to be inefficient if speaker n is identified while

$$\beta_{i,\lambda_n} > \beta_{i,\lambda_k} \tag{15}$$

for some values of k. The level of inefficacy of a particular coefficient depends on how many times equation xxx is true. If speaker n is recognized while

$$\beta_{i,\lambda_n} > \beta_{i,\lambda_k}, \quad \forall k \neq n, \tag{16}$$

then the coefficient x_i can be considered as irrelevant for the speaker recognition process. During the evaluation procedure of the input coefficients, this condition has never appeared, which let us to believe, that the coefficients we have used are not totally inefficient for speaker recognition according to our method. Some seem better than others. The next section describes the procedure used in this paper to evaluate an appropriate set of input coefficients according to their efficiency.

4 Coefficient Evaluation

The most common set of coefficients generally used for speaker recognition are Mel Frequency Cepstral Coefficients whose evaluation involves the use of a set of band-pass filters. Changing the value of these coefficients requires a modification in the bandwith of each filter or a modification in their position. In [5] a new algorithm was presented to evaluate input coefficients based on the Mel scale, which is more flexible than the classical MFCC technique. It is this algorithm that is used in this paper to evaluate input parameters.

Assuming that M static coefficients are needed, the main steps of this procedure can be summarised as follows, for each value of m:

1. Evaluate a vector of position P_m, whose elements, $P_{m,l}$, are given by:

$$P_{m,l} = \gamma(10^{\frac{l}{m}\frac{\theta}{2595}} - 1); \quad 0 \leq l \leq m. \tag{17}$$

The right side of this equation is based on the inverse of the equation, defining the mel scale. The value of γ is generally equal to 700; the value of θ is chosen in such a way that the last element, $P_{m,m}$, is equal to 4000 Hz for telephone speech.

2. Define between each pair of elements, $P_{m,l}$ and $P_{m,l+1}$, a subset of vectors, $\widetilde{w}_{m,l}$, given by:

$$\widetilde{w}_{m,l} = \{(-1)^l \cos(\frac{\pi}{I_l}(i-0.5)) \quad | \quad 1 \leq i \leq I_l\}, \tag{18}$$

where I_l is the total number of energy elements x_k between $P_{m,l}$ and $P_{m,l+1}$.

3. Concatenate all the vectors $\widetilde{w}_{m,l}$ for $0 \leq l \leq m$ to obtain the final vector $\widetilde{W}_m$, that is,

$$\widetilde{W}_m = \bigcup_{l\in[0,m-1]} \widetilde{w}_{m,l}. \tag{19}$$

The set of vectors $\widetilde{W}_m$ with $m \in [1, M]$ acts as a base of projection for the spectral energy vector X, obtained after the fast Fourier Transform. The set of coefficients, c_m, obtained using $\widetilde{W}_m$ is given by:

$$c_m = \beta \sum_{k=1}^{K} \widetilde{W}_{m,k} \log_{10}(x_k). \tag{20}$$

To modify the values of the coefficients c_m, we modify the base of projection $\widetilde{W}_m$ by changing the value of γ defined in the previous algorithm.

Experiments conducted in this paper are based on three value of γ, 350, 700 and 1400; hence three differents set of coefficients. In order to determine the most efficient coefficients of each set, the following test has been conducted, according to the algorithm

defined in section 3. For each speaker n in the training set, we calculate, y_i, corresponding to the number of times

$$\beta_{i,\lambda_n} > \beta_{i,\lambda_k} \tag{21}$$

that the input utterance belongs to speaker n. For each set of coefficients defined by a given value of γ, we retain the 5 coefficients for which y_i is minimum, allowing us to obtain 15 static coefficients. The dynamic coefficients are deduced accordingly. The following section described results obtained when this technique is applied.

5 TESTS AND RESULTS

The algorithms presented in this paper are used in the text-independent speaker identification system of INRS-Telecommunications. The evaluation of the system was conducted using a subset of 14 speakers of the Spidre database. The vectors x_t, containing 15 static and dynamic coefficients, are evaluated following the algorithms presented in section 3 and 4. The gaussian mixture models, λ containing 20 component densities, are evaluated following the expectation maximisation (EM) algorithm [6]. For each speaker the models are evaluated using approximately 60 seconds of speech from one channel, channel A. The identification is performed using approximately 10 seconds of speech. Table 1 shows comparative results obtained when only one model is used and when two models, as suggested in this paper, are used.

Type of coefficients	A	B
$\gamma = 350$	95.5%	35.6%
$\gamma = 700$	97.8%	36.7%
$\gamma = 1400$	99.1%	38.1%
mix	99.1%	41.7%

Table 1: Comparative results between different set of coefficients.

Results in column A are obtained when the training and the recognition are performed using the same transmission channel. Results in column B are obtained when different transmission channels are used for training and recognition. The performance of the speaker recognition system seems to increase with γ. The mix coefficients are those obtained when the 5 best static coefficients of each group are retained, according to the procedure described in section 3, and grouped together to form a specific set of coefficients.

6 SUMMARY

In this paper we have presented an algorithm to evaluate the performance of input coefficients. It is based on a modification of the Gaussian Mixture Model; changing the arithmetic mean for a geometric mean with a diagonal covariance matrix.

We have explored the possibility of using this algorithm from among differents set of input coefficients in order to retain those more adequate for speaker recognition according to the criteria defined in section 3. result obtained with this technique give a confidence interval about its use in speaker recognition process.

REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, pp. 357-366, Aug. 1980.

[2] F. Beaufays and M. Wientraub, "Model Transformation For Robust Speaker Recognition From Telephone Data", IEEE Trans. Acoust., Speech, Signal Processing, 1997

[3] A. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Acoust., Speech, Signal Processing, vol. 3, No. 1, pp. 72-83, January 1995.

[4] C.R. Janowski, T.F. Quatieri, D.A. Reynolds, "Measuring fine structure in speech: Application to Speaker Identification", in Proc. ICASSP-95, pp. 325-328.

[5] R. Vergin, "An Algorithm For Robust Signal Modelling in Speech Recognition", in Proc. ICASSP-98.

[6 A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via EM algorithm", J. Royal Stat. Soc., vol. 39, pp. 1-38, 1977.