



A Double Gaussian Mixture Modeling Approach to Speaker Recognition

Rivarol Vergin and Douglas O'Shaughnessy

CML Technologies, 75 Blvd de la technologie
Hull, J8Z-3G4, Québec, Canada

INRS-Télécommunications, 16 Place du Commerce,
Île-des-Sœurs, H3E-1H6, Québec, Canada
email: vergin@inrs-telecom.quebec.ca

ABSTRACT

The first motivation for using Gaussian mixture models for text-independent speaker identification is based on the observation that a linear combination of gaussian basis functions is capable of representing a large class of sample distributions. While this technique gives generally good results, little is known about which specific part of a speech signal best identifies a speaker. This contribution suggests a procedure, based on the Jensen divergence measure, to automatically extract from the input speech signal the part that best contribute to identify a speaker. It is shown, by results obtained, that this technique can significantly increase the performance of a speaker recognition system.

1 INTRODUCTION

An important application of speech analysis, automatic speaker recognition, is a subject of many recent studies. One application concerns the possibility of verifying a person's identity prior to admission to a secure facility or to a transaction over the telephone. To attain this goal many algorithms, based on some measures of speaker variability, have been proposed in the literature. One of the most popular is the Gaussian Mixture Model (GMM), often used in text-independent speaker identification [1]. This technique involves first a speech analysis process whose role is to extract from the input speech signal a set of feature vectors which reflect a person's vocal tract structure. These vectors are used in a second step, during the training phase, to evaluate the model, $\lambda = \{p_m, \mu_m, \Sigma_m\}$, characterizing each speaker. Generally, each individual component gaussian, corresponding to a fixed value of m , is interpreted to represent some broad acoustic classes.

Because the whole utterance is used during the training and the identification process, it is difficult to identify which set of acoustic classes representing

some broad phonetic events, such as vowels, nasals or fricatives, contribute or do not contribute to identify a speaker.

Since speaker recognition, especially in text-independent cases, depends primarily on accurate model estimation, special attention must be directed toward efficient modeling of each speaker. This paper suggests a procedure, based on the Jensen divergence measure [2], to automatically extract from the input speech signal the part that best contributes to identify a speaker. The results obtained with this technique give a confidence interval about its use in the speaker recognition process.

The rest of this paper is organized as follows. Section 2 gives an overview of the Gaussian mixture models, section 3 presents the Jensen difference measure, section 4 explains the double Gaussian mixture modeling approach suggested in this paper, section 5 describes the test procedures and presents some comparative results between two 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 [3] is often used; this is also the case in this paper, as well in speech as in speaker [1] 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 the Gaussian mixture model (GMM). According to this approach, each speaker is represented by a model λ ,

$$\lambda = \{p_m, \mu_m, \Sigma_m\}, \quad m = 1, \dots, M, \quad (1)$$

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

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

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:

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

p_m are the mixture weights satisfying the constraint that

$$\sum_{m=1}^M p_m = 1. \quad (4)$$

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 λ' 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

$$p(\lambda_n|X) > p(\lambda_k|X), \quad \forall k \neq n. \quad (5)$$

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:

$$\sum_{t=1}^T \log p(x_t|\lambda_n) > \sum_{t=1}^T \log p(x_t|\lambda_k), \quad \forall k \neq n. \quad (6)$$

It is important to realise that the whole utterance is used during the training and the testing procedures. Accordingly it is difficult to say which specific part of a speech signal, representing some broad phonetic events, best identifies a speaker.

Section 3 explains briefly the *Jensen difference* measure [2] which will be used in section 4 to automatically extract from the input speech signal the part that best contributes to identify a speaker.

3 DIVERGENCE MEASURE

The Shannon entropy, defined by:

$$H_n(x) = - \sum_{i=1}^n x_i \log x_i, \quad (7)$$

is one of the most widely used indices of diversity of a multinomial distribution, $x = (x_1, \dots, x_n)$ where $x_i \geq 0$ and $\sum_i x_i = 1$. The concavity of $H_n(x)$ permits defining the diversity of a mixed distribution, $\frac{x+y}{2}$, as

$$H_n\left(\frac{x+y}{2}\right) = \frac{1}{2}[H_n(x) + H_n(y)] + J_n(x, y). \quad (8)$$

The first term of the second part of the equation, $\frac{1}{2}[H_n(x) + H_n(y)]$, is interpreted as the average diversity within the distributions. The second term given by:

$$J_n(x, y) = H_n\left(\frac{x+y}{2}\right) - \frac{1}{2}[H_n(x) + H_n(y)], \quad (9)$$

which is often called the *Jensen difference* [2], is non-negative and vanishes if and only if $x = y$. $J_n(x, y)$ can then be used as a natural measure of divergence between two vectors in a convex set of n -dimensional real vector space. If x is similar to y , the value of $J_n(x, y)$ will be relatively small. Inversely if x is quite different from y the value of $J_n(x, y)$ will be relatively high.

It is interesting to note that the *Jensen difference* can also be used to make a selection from among a set of vectors. Indeed assuming that x is a fixed vector and Z a set of vectors, $\{z_1, \dots, z_m\}$, the *Jensen difference* can be used to find a subset of vectors of Z that are close to x . It is in the sense that the *Jensen difference* measure is used in the next section to automatically extract from the input speech signal the part that best contributes to identify a speaker.

4 DOUBLE GAUSSIAN MIXTURE MODELS

The rule of decision, in the speaker identification process, defined by equation 5 corresponds to the

maximum a posteriori probability for a given observation sequence T . According to the rule of decision, defined by equation 6, the identified speaker, n , is the one for which the sum of the elements, $\log p(x_t|\lambda_n)$, over all of the input utterance, T , is greater than the term appearing on the right side of the equation and for all values of $k \neq n$. Clearly they must have a subset of vectors x_t for which the sense of the inequality, $>$, defined in the equation is not respected. That is, there exist some values of t for which $\log p(x_t|\lambda_n) \leq \log p(x_t|\lambda_k)$. The ones meeting this condition do not really contribute identifying a speaker.

To quantify the contribution of each feature vector, x_t , in the decision scheme, we suggest in this paper the following procedure: a vector w^t of dimension N is evaluated for each input feature x_t . The elements, w_n^t , of w^t are given by:

$$w_n^t = \frac{\log p(x_t|\lambda_n)}{\sum_i \log p(x_t|\lambda_i)}, \quad n = 1, \dots, N. \quad (10)$$

This equation assumes that there are N known speakers, represented by a model λ_n , in the system. Each element, w_n^t , shows the accuracy of a given model, λ_n , producing an observation vector x_t . Clearly if w_n^t are similar for all values of n , the feature vector, x_t , does not really contribute to the identification process. Inversely if the elements, w_n^t are not similar, that is, if for some values of n , w_n^t are low and for some other values of n w_n^t are high, it is reasonable to conclude that this particular vector, x_t contribute to the identification process.

To compute the similarity measure between the elements w_n^t , we evaluate the *Jensen difference* between the vector w^t and a fixed reference vector, r , whose elements r_n ,

$$r_n = \frac{1}{N}, \quad n = 1, \dots, N, \quad (11)$$

are all equal. The resulting difference measure,

$$J_N(w^t, r) = H_N\left(\frac{w^t + r}{2}\right) - \frac{1}{2}[H_N(w^t) + H_N(r)], \quad (12)$$

is a good indicator of the contribution of x_t to the identification process. Indeed if $J_N(w^t, r) = 0$, the probability that a model λ_n produces an observation vector, x_t , is the same for all n because $w^t = r$, and accordingly x_t does not contribute to the identification process, which is not the case if $J_N(w^t, r)$ is quite different from zero.

The principle described above is used in this paper to create a double gaussian mixture model as follows: In a first step, the whole utterance for each speaker is used to estimate a first set of models λ^1

corresponding to the classical GMM approach. In a second step, using the same utterance, an evaluation of $J_N(w^t, r)$ is made for each vector x_t and those for which $J_N(w^t, r)$ is below a certain threshold, α , are discarded; the remaining ones are used to evaluate a second set of models λ^2 . The training phase then produces two sets of models, λ^1 and λ^2 . The recognition phase is implemented according to the same principle: in a first step, using the set of models λ^1 , some input vectors x_t are discarded following the divergence measure $J_N(w^t, r)$. The remaining ones are used in a second step, using the second set of models λ^2 , to identify the speaker. An input signal characterized by a set of vectors x_t will be associated to a particular speaker, n , during the identification process if:

$$\sum_{t=1}^{T'} \log p(x'_t|\lambda_n) \geq \sum_{t=1}^{T'} \log p(x'_t|\lambda_k), \quad \forall k \neq n. \quad (13)$$

In this equation x'_t is an input vector x_t for which $J_N(w^t, r)$ is greater than the threshold, α , and T' is the total number of vectors x_t meeting this condition. There are no good theoretical means to guide one choosing α . Its value have been fixed experimentally. The next section shows results obtained when this technique is applied.

5 TESTS AND RESULTS

The algorithm presented in this paper is 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 10 static and dynamic coefficients, are evaluated following the MFCC algorithm [3]. The gaussian mixture models, λ^1 and λ^2 , containing 20 component densities, are evaluated following the expectation maximization (EM) algorithm [4]. 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 system	Channel A	Channel B
One GMM model	83.8%	41.6%
Two GMM model	88.9%	51.7%

Table 1: Comparative results between two different systems.

These results give a confidence interval about the use of two GMM models in the speaker identification process. It can be observed that results obtained

with the system using two GMM models is greater than those obtained with the system using only one GMM model and this is true even if the channel used during the recognition phase is the same (channel A) or different (channel B) from the channel used during the training phase.

6 SUMMARY

In this paper we have examined one of the most common methods in a speaker identification system, that is, the Gaussian mixture model. Because this technique uses the whole input speech signal during the training and the testing procedures, it becomes difficult to say which parts of the speech signal best contribute to identify a speaker.

We have explored the possibility of using the *Jensen difference* measure to automatically extract from the input speech the parts that best contribute to the identification process. The new proposed algorithm uses two sets of Gaussian mixture models for speaker recognition. The first set of models is evaluated using the whole input utterance of each speaker; the second set of models is evaluated using the subset of feature vectors for which the corresponding *Jensen difference* measure is above a predefined threshold. Results obtained with this technique give a good confidence interval about the use of two GMM models in the speaker identification process.

REFERENCES

- [1] D. 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.
- [2] J. Burbea and C. R. Rao, "On the Convexity of Some Divergence Measures Based on Entropy Functions", *IEEE Trans. Information Theory*, vol. IT-28, No. 3, pp. 489-495, May 1982.
- [3] 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.
- [4] 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.