ON THE USE OF SOME DIVERGENCE MEASURES IN SPEAKER RECOGNITION

Rivarol Vergin *Douglas O'Shaughnessy*†

Université de Moncton,
Moncton, E1A-3E9, New-Brunswick, Canada
†INRS-Télécommunications, 16 Place du Commerce,
Île-des-Sœurs, H3E-1H6, Québec, Canada
e-mail: vergin@inrs-telecom.uquebec.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 contributes to identify a speaker. Experiments conducted using the Spidre database indicate a significant improvement in the performance of the 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 for 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, sections 4 and 5 explain how this measure is used in this paper, section 6 describes the test procedures and presents some comparative results between two different systems and section 7 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,

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

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

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

μ_m and Σ_m are respectively the 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 $\lambda^{'}$ 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

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

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 realize 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, $x = (x_1, \cdots, x_n)$ 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

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

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:

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

which has been 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.

Our intention in this paper is to use the *Jensen difference* measure to make a selection from among a set of vectors. Let us assume that x is a fixed vector and $Z = \{z_1, \cdots, z_m\}$ is a set of vectors; by evaluating the *Jensen difference* between x and each element, z_i, of Z, we can find a subset of vectors of Z that are closer to x according to some similarity criteria. In the next section, we describe how this measure is used to select from the whole input speech signal the parts that best identify a particular speaker.

4. INPUT-SPEECH CLASSIFICATION

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 there must exist some values of t for which

$$\log p(x_t|\lambda_n) \leq \log p(x_t|\lambda_k), \quad \forall k \neq n. \quad (10)$$

The subset of vectors x_t for which the preceding equation is true does not really contribute to identifying a speaker. In this paper, we suggest a procedure to quantify the contribution of each feature vector, x_t, in the decision scheme, according to the following algorithm: let us assume that there are N known speakers represented by a model λ_n in the system; we evaluate for each input feature x_t, a second vector w^t whose elements, w_n^t, are given by:

$$w_n^t = \frac{\log p(x_t|\lambda_n)}{\sum_i \log p(x_t|\lambda_i)}, \quad n = 1, \cdots, N. \tag{11}$$

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, contributes to the identification process.

To quantify the similarity between the elements w_n^t of w^t, we evaluate the *Jensen difference* between the vector w^t and a reference vector, r, whose elements r_n are given by:

$$r_n = \frac{1}{N}, \quad n = 1, \cdots, N. \tag{12}$$

Since the elements of r are all equal, it results that the *Jensen difference* between w^t and r:

$$J_N(w^t, r) = H_N(\frac{w^t + r}{2}) - \frac{1}{2}[H_N(w^t) + H_N(r)], \tag{13}$$

is close to zero if and only if the elements of w^t are all similar. 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 example shown in figure 1 is intended as preliminary evidence about the capacity of the *Jensen difference* measure in helping discriminate between subsets of input vectors, x_t, that are more relevant than others for the speaker recognition task. The waveform shown in figure 1 (a) corresponds to the sentence: "Hi, Carolyn, dear, are you reading the papers". If our intuitive feeling is that unvoiced consonants like /s/ or /ch/ are not really relevant for speaker recognition, the curve shown in figure 1 (c) also shows some other preliminary information. The value of the *Jensen difference* measure is greater in the transitions between phonemes than in the stable parts of phonemes, which suggests that transitions between phonemes are more relevant than stable parts of phonemes in the speaker recognition process. Obviously these results are closely related to the type of input features, MFCC, used in this analysis. Futher investigations using other kinds of input parameters have to be conducted before concluding about the contribution of different classes of phonemes in the speaker recognition process.

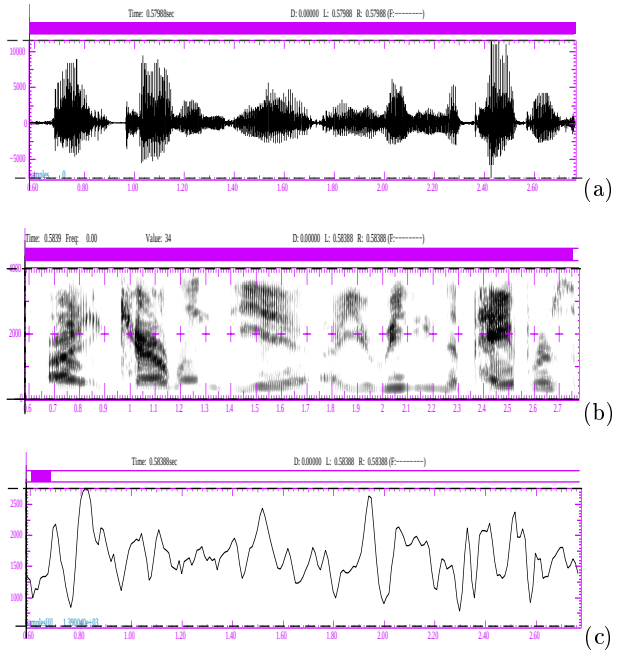

Figure 1. a) waveform, b) wideband spectrogram, c) Jensen Difference values, all for a typical sentence.

5. ENHANCEMENT OF GAUSSIAN MIXTURE MODELS

The principle described above is used in this paper to enhance the gaussian mixture models [5] 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. Those for which $J_N(w^t, r)$ is below a certain thresold, 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. The rule of decision is now dictated as follows: an input signal characterized by a set of vectors x_t will be associated to a particular speaker, n, during the identification process if:

$$\sum_{t=1}^{T'} \log p(x_t'|\lambda_n) \geq \sum_{t=1}^{T'} \log p(x_t'|\lambda_k), \quad \forall k \neq n. \qquad (14)$$

In this equation x_t' is an input vector x_t for which $J_N(w^t, r)$ is greater than the thresold, 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 has been fixed experimentally. The next section shows results obtained when this technique is applied.

6. TESTS AND RESULTS

The evaluation of the system was conducted using a subset of 20 speakers of the Spidre database. The vectors x_t, containing 15 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 maximisation (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 for different speech durations. Table 1 shows comparative results obtained when the whole utterance is used and when only some specific parts of the input utterance are used.

Duration	Enhancement	A	B
1 second	Not Applied	84.3%	40.2%
	Applied	86.5%	42.6%
3 seconds	Not Applied	92.02%	41.6%
	Applied	93.9%	44.5%
5 seconds	Not Applied	93.7%	43.3%
	Applied	95.8%	45.3%
10 seconds	Not Applied	99.1%	45.6%
	Applied	99.3%	47.7%

Table 1. Comparative results between two different systems.

Results (A) are obtained when the same channel is used for training and recognition. Results (B) are obtained when the channel used for training is different from the channel used for recognition. It can be observed that the speaker recognition system performs better when the suggested enhancement technique is applied. This is true for both cases (A & B) and for different durations as shown in Table 1.

7. 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 thresold. Results obtained with this technique give a good confidence interval for its use 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.

[5] R. Vergin and D. O'Shaughnessy, "A Double Gaussian Mixture Modeling Approach to Speaker Recognition", European Conference on Speech Communication and Technology, pp. 2287-2290, Sep. 1997.