GMM BASED SPEAKER IDENTIFICATION USING TRAINING-TIME-DEPENDENT NUMBER OF MIXTURES

Chakib Tadj[†], Pierre Dumouchel[†‡] and Pierre Ouellet[†]

[†]École de Technologie Supérieure - Electrical Engineering, 1100 rue Notre Dame Ouest
Montréal (Qc) - H3C 1K3 - Canada - e-mail: {ctadj,pouellet}@ele.etsmtl.ca

[†‡]Centre de Recherche Informatique de Montréal, 1801, avenue McGill College, bureau 800
Montréal (Qc) - H3A 2N4 - Canada - e-mail: Pierre.Dumouchel@crim.ca

ABSTRACT

In this paper, we present the study of the performance of our standard *GMM* speaker identification system in "a limited amount of training data" context. We explore the use of different mixture components for different speakers/models. Different approaches are presented: (a) a non-linear transformation of speech duration vs. number of mixtures is proposed in order to set correctly the appropriate number of model mixtures for each speaker according to the available training data. (b) From exhaustive experiments, the appropriate linear transformation is deduced. The resulting transformation offers several advantages: (a) each speaker is well modelized (b) the performance is improved by more than 6% on the SPIDRE corpus and finally (c) the number of mixtures is reduced and thus leads to a faster system response.

1. INTRODUCTION

Speaker identification systems are usually trained on a large amount of speech data collected in a specific environment. During identification, these systems require the same environment to achieve good accuracy. Discrepancies between training and testing environments result in a degradation of performance. Different environments could mean:

- different (mismatched) microphones: microphones distort the speech signal mainly into two distinct ways. First they allow different levels of ambient noise that account for an additive effect in the recording speech and second they act as unknown linear filters, causing a variable spectral tilt that depends on the specific microphone characteristics.
- voice changes (voice's modification over time, cold, cough).
- different psychological states of the speaker (such as stressed, anxious, relaxed).
- different amount of training data available per speaker.

In this paper we will study the performance of a *GMM* speaker identification system for a limited amount of training data. We will explore the use of different mixture components for different speakers/models.

2. AMOUNT OF TRAINING VS. MODELIZATION

2.1. Gaussian Mixture Models: Review

In the Gaussian Mixture Model (*GMM*) [7], the distribution of the parametrization speech vector of a speaker is modeled by a weighted sum of Gaussian densities:

$$p(\vec{x}\,|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \quad with \sum_{i=1}^{M} p_i = 1 \qquad (1)$$

where $\vec{x}$ is a D-dimensional cepstral vector, λ is the speaker model, $b_i(\vec{x})$, $i = 1, \cdots, M$, are the component densities characterized by the mean $\vec{\mu}_i$ and the covariance matrix Σ_i and p_i, $i = 1, \cdots, M$, are the mixture weights. Each component density is a D-variate Gaussian Mixture function of the form:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\} \qquad (2)$$

The model parameters $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}$ are estimated by an EM algorithm[2]. In the next section, we will show how the number of mixtures in our models is set, according to the amount of training data available.

2.2. Description of the Problem

As with almost all speaker identification systems [3], each speaker in our standard system is modelized with the same number of mixtures. Our experiments have already shown that, after performing silence detection, some speakers have a smaller amount of training data than others. In spite of this discrepancy in the training data, we still use the same fixed number of mixtures per speaker [10]. This leads us to ask this question: is it appropriate to fix the same number of mixtures for all speakers independently of the amount of training data?

2.3. Solution: Use of Different Mixture Components for Different Speakers

In this section we propose to focus on the importance of the number of mixtures on our system's accuracy and explore the use of different mixture components for different speakers. The purpose of this work is to confirm that there is a strong relation between the amount of training data and the number of mixtures. Intuitively, we think that this number is proportional to the available amount of training data. We propose to determine from experiments this proportionality coefficient.

In order to find a relationship between the amount of training data and the number of mixtures, we propose two different approaches: (a) determine the number of mixtures vs. speech duration according to a nonlinear transformation and (b) determine the number of mixtures by exhaustive experiments. From the exhaustive experiments, we will deduce an appropriate transformation speech signal duration vs. number of mixtures.

2.3.1. Nonlinear Transformation

A nonlinear transformation (speech duration vs. number of mixtures) with different parameters is proposed. This transformation $\mathcal{F}_1$ is defined by[1]:

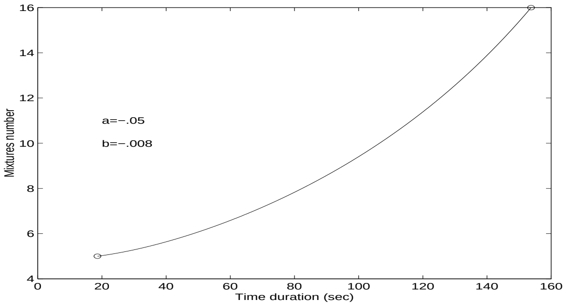

Figure 1: Concave curve

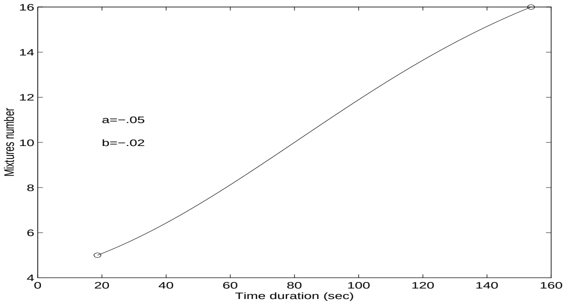

Figure 2: Flat curve

$$\mathcal{F}_1 : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$$
$$t \longmapsto n_{mixtures} = \mathcal{F}_1(t) = \frac{e^{at} - c}{e^{bt} - d} \qquad (3)$$

where coefficients a and b were set to different values, as shown in Figures 1, 2 and 3. The coefficients c and d are determined according to the following constraints (cf. Figure 4):

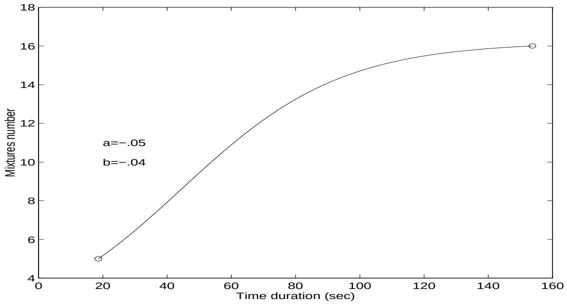

Figure 3: Convex curve

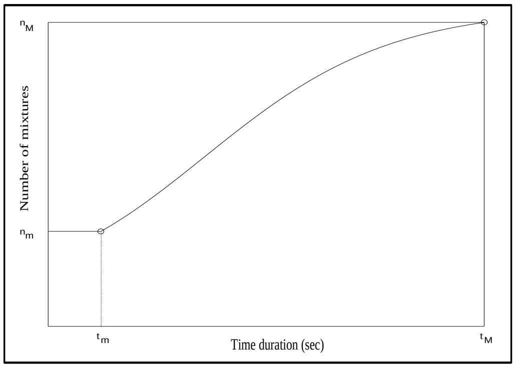

Figure 4: Constraints to determine constants c and d

$$\mathcal{F}_1(t_m) = n_m = \frac{e^{at_m} - c}{e^{bt_m} - d} \qquad (4)$$

$$\mathcal{F}_1(t_M) = n_M = \frac{e^{at_M} - c}{e^{bt_M} - d} \qquad (5)$$

where t_m and t_M are the minimum and the maximum length duration of the training data for a given speaker respectively, and n_m and n_M the corresponding number of mixtures. From Equations (4) and (5) we determine c and d:

$$c = \frac{n_M}{(n_M - n_m)}[e^{at_m} - n_m e^{bt_m}] - \frac{n_m}{(n_M - n_m)}[e^{at_M} - n_M e^{bt_M}]$$

$$d = \frac{-(e^{at_M} - e^{at_m}) + (n_M e^{bt_M} - n_m e^{bt_m})}{(n_M - n_m)} \qquad (6)$$

The experiment results on the nonlinear transformation are presented in section 3.2.

2.3.2. Exhaustive Experiments

The main idea of the exhaustive experiments consists of running complete training and test processes at each time t for different number of mixtures (incremented reasonably), and then the number of mixtures corresponding to the best recognition performance is selected. Figure 5 shows an example at $t = 4$ seconds. In this case, the best number of mixtures selected is six, which corresponds to a 47% recognition rate.

[1] Since the number of mixtures is an integer, the function is really $\mathbb{R}^+ \longrightarrow \mathbb{N}^+$, but we keep the above notation for generality, and we truncate the result at the end of the arithmetic manipulations.

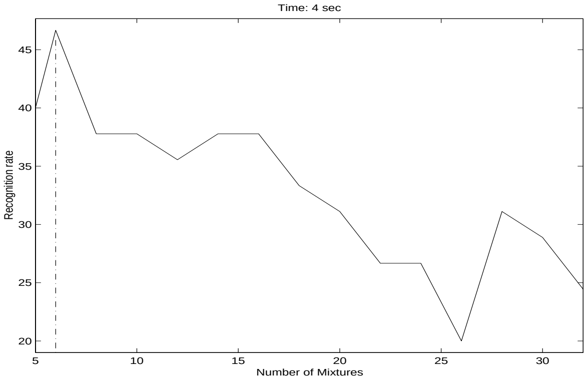

Figure 5: Number of mixtures for four seconds of speech

The time duration t is then incremented by a predefined step and the same processes are repeated. In order to make these experiments run faster, some constraints are added to avoid some useless experiments as described in Figure 6.

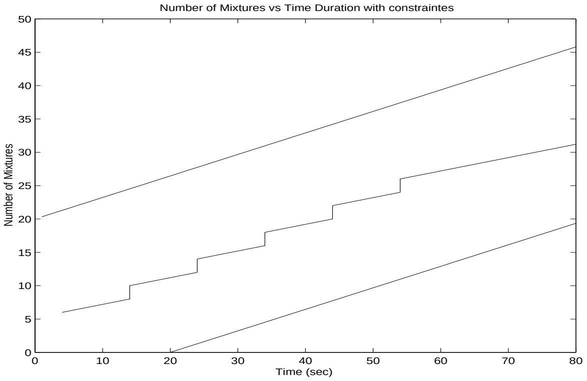

Figure 6: Optimizing computation by adding time warping constraints

However, since the speech signals do not have the same duration, the score results can be biased for a given speaker if she/he does not have a sufficient amount of training data when the time variable t grows. In order to remedy this problem, the process is applied as long as all the speakers have enough data for a given time duration t, and by extrapolation the number of mixtures can be found for a longer speech signal, as described in the next section.

2.3.3. Linear Transformation by Extrapolation

For up to 20 seconds of speech, all speakers have a speech signal of the same length. The transformation is found to be linear as shown in Figure 7. This transformation can be defined by:

$$\begin{aligned} \mathcal{F}_2 :\ & \mathbb{R}^+ \longrightarrow \mathbb{R}^+ \\ & t \longmapsto n_{mixtures} = \mathcal{F}_2(t) = 0.4 * t + 4 \end{aligned} \tag{7}$$

For all signals longer than 20 seconds, the number of mixtures is computed by extrapolation with respect to Equation (7). The transformation given by Equation (7) will produce the best results for a given amount of training data, but is not necessarily the optimal one according to the minimum number of mixtures vs. amount of training data, as will be shown in section 3.3.

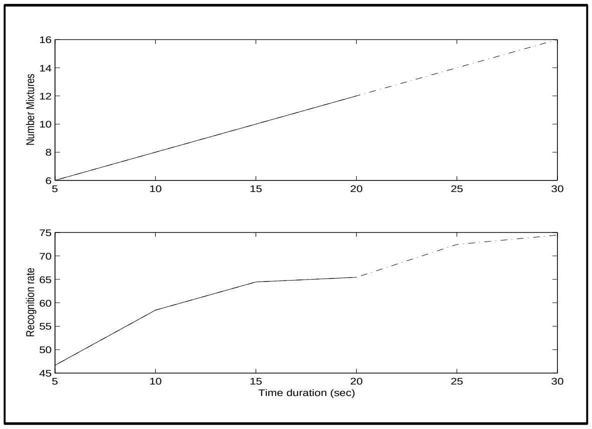

Figure 7: Determination of the number of mixtures by extrapolation

3. EXPERIMENTS ON SPIDRE

3.1. The SPIDRE Corpus

The original conversations for the 45 SPIDRE target speakers were subdivided in training and test as follows: three conversations (two from match conditions and one from mismatch conditions) for training, and one conversation (from mismatch conditions) for the test. In all the experiments shown, the identification test has been done on 10 seconds of speech signal. The features are a 26 dimensional vectors consisting of 12 cepstral coefficients, 12 Δ coefficients, logarithmic power and Δ logarithmic power. Analysis conditions are listed in table 1.

Pre-emphasis	1-0.97z^{-1}
Window length	25.0 ms
Window shift	10.0 ms
MFCC cepstrum order	24
Cepstral coefficient liftering	22
Cepstral mean normalization	yes
Hamming window	yes

Table 1: Analysis conditions used for different experiments

3.2. Results from the Nonlinear Transformation

Table 2 shows the results obtained from the nonlinear transformation obtained from Figures 1, 2 and 3 corresponding to $b = -0.008$, -0.02 and -0.04 respectively.

Parameter b	-0.008	-0.02	-0.04
Recognition (%)	66.67	71.11	68.89

Table 2: Identification results for nonlinear transformation

These results show that the best result is obtained with $b = -0.02$, which corresponds to the less drastic transformation. This information motivated us to suggest a more linear transformation as shown in the following results.

3.3. Results from the Linear Transformation

Table 3 shows the results obtained from a linear transformation obtained by Equation (7). As this transformation seems not to be optimal, we present some experiments with slopes tuned around the slope obtained from Figure 7.

Slope a*t + 4	0.30	0.35	*0.36	0.37	0.38
Recognition	75.56	75.56	77.78	77.78	75.56

Slope a*t + 4	0.39	*0.40	0.41	0.42
Recognition	77.78	77.78	75.56	73.33

Table 3: Identification results for linear transformation

As expected, the best recognition rate result is the one obtained with slope 0.4. However, it is not the optimal slope according to the number of mixtures, as we can obtain the same performance with a smaller slope 0.36 which implies a reduced number of mixtures and thus a faster system response.

3.4. Comparison With a Fixed Number of Mixtures

The use of the slope 0.36 for all 45 speakers has shown that:

- the average number of mixtures used is 27
- the minimum number of mixtures used is 10
- the maximum number of mixtures used is 59

In order to show the efficiency of our transformation, we have done some other experiments independently with a fixed number of mixtures. We have performed an exhaustive experiments to find out the best number of mixtures which corresponds to the best system's performance. For brevity, we present the results for 10, 16, 27, 32 and 59 mixtures respectively.

Number of mixtures	10	16	27	32	59
Recognition (%)	66.67	66.67	73.33	71.11	68.89

Table 4: Identification results with a fixed number of mixtures

Table 4 shows that the best result is obtained with a number of mixtures equal to 27, which is the average number of mixtures used by the best slope's transformation result obtained in section 3.3. However the corresponding recognition accuracy is less than the result obtained with the linear transformation with an optimal slope equal to 0.36. These results allows us to enumerate several advantages of this transformation: (a) the performance is improved by more than 6% on SPIDRE corpus with respect to the best fixed mixture's number (b) each speaker can be modelized with respect to the available training data (c) the number of mixtures is reduced and thus leads to a faster system response and finally (d) there is no need to make several exhaustive experiments in order to find the appropriate number of mixtures to create speaker's models for the use of new databases.

4. CONCLUSION

In this paper, we have proposed two different methods in order to determine the relationship between the amount of training data and the number of mixtures. In the first one, we have proposed a nonlinear transformation with different parameters. Even if this technique has shown good results, the linear transformation with an appropriate slope has shown better performance with an improvement of more than 6% on the recognition rate. This last transformation has also shown its optimality in the sense that it uses a minimum number of mixtures and leads to a faster speaker identification's system.

5. ACKNOWLEDGEMENTS

We wish to acknowledge the financial support of Canada's Ministry of Defense, and to thank Luc Gagnon and Karl Boutin for several good ideas. We also wish to acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada.

6. REFERENCES

[1] B. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", JASA 72, Vol. 55, pp. 1304-1312, 1972.

[2] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm", J. Roy. Stat. Soc, Vol. 39, pp. 1-38., 1977.

[3] S. Furui, "An Overview of Speaker Technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[4] J. P. Haton, J. M. Pierrel, G. Caelen and J. L. Gauvain, "Reconnaissance Automatique de la Parole", ed. Bordas, Paris, 1991.

[5] H. Hermansky, "RASTA-PLP Speech Analysis Technique", IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, pp. 121-124, 1992.

[6] R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust Speaker Recognition - A Feature Based Approach", IEEE Signal Processing, pp. 58-71, 1996.

[7] D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", PhD Thesis, Georgia Institute of Technology, 1992.

[8] D. A. Reynolds, "The Effect of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard Corpus", IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, pp. 113-116, 1996.

[9] F. Soong and A. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", Acoustics, Speech and Signal Processing, Vol. 36, pp. 871-879, 1988.

[10] C. Tadj, P. Dumouchel and Y. Fang, "N-best GMM's for Speaker Identification", Proceeding of Eurospeech Conference, Vol. 5, pp. 2295 - 2298, 1997.