GMM AND ARVM COOPERATION AND COMPETITION FOR TEXT-INDEPENDENT SPEAKER RECOGNITION ON TELEPHONE SPEECH

J.-L. Le Floch, C. Montacié & M.-J. Caraty

LAFORIA-IBP, Université Paris 6, CNRS-URA 1095
4 place Jussieu, 75252 Paris Cedex 5, FRANCE

ABSTRACT

In order to improve the performances of speaker recognition on telephone speech, we investigate the ability to cooperate of two different natures modelizations: the GMM and the ARVM.

For the cooperation and competition of the GMM and ARVM modelizations, we used normalized measures. We develop two approaches for these cooperation and competition : a global approach and an analytical approach. We investigate experiments on whole sentences or selected phonetic segments.

Theses approaches allow us to obtain performances improvements for both cooperation and competition, and good results on 168 speakers of the NTIMIT database (GMM : 61.7 %, ARVM : 78.1 %, cooperation : 79.9 % and competition : 82.6 %).

1. INTRODUCTION

Speaker identification, claimed identity verification and talker separation during conversations are the different tasks of the speaker recognition. In this paper, we are interested in text-independent speaker identification in phone quality speech, but the presented approaches could be used on the other tasks.

We develop a cooperation and a competition of two different natures modelizations. The first one, the GMM [1], is a modelization of the parametrisation distribution of the speaker speech. The second, the ARVM [2, 3], is a modelization of the speaker speech spectral evolution. To allow cooperation and competition between different modelizations we use a classical measure normalization. We investigate the cooperation/competition of the GMM and ARVM on two levels : global and analytic. In order to improve the performances, we used results of previous study [4] and repeat the experiments on selected phonetic segments.

The experiments are carried out on the test corpus of the NTIMIT database (168 speakers) [5]. The NTIMIT database is made of TIMIT utterances which are transmitted over a variety of telephone lines conditions (250 various short and long distances lines). Each sentence of a speaker is transmitted over a different telephone line in order to have realistic conditions. The first eight sentences of each speaker are used for the speaker models training (GMM and ARVM). The two remaining sentences are separately used as tests.

2. GMM AND ARVM MODELIZATIONS

In this section we present two modelizations which proved to be robust in speaker recognition on telephone speech [6, 7] and have different approaches.

2.1. Gaussian Mixture Models

In the Gaussian Mixture Model, the distribution of the parametrisation speech vector of a speaker is modelled by a weighted sum of Gaussian densities :

$$p(y \mid \lambda) = \sum_{i=1}^{m} p_i \, b_i(y) \quad \text{with} \sum_{i=1}^{m} p_i = 1$$

and

$$b_i(y) = \frac{1}{(2\pi)^{p/2} \, |\Sigma_i|^{1/2}} \, e^{\left[-\frac{1}{2} (y-\mu_i)^T \, \Sigma_i^{-1} \, (y-\mu_i) \right]}$$

where y is a p-dimensional cepstral vector,

λ is the speaker model,

p_i, (i = 1,..,m) is the Gaussian densities weights,

b_i, (i = 1,..,m) is the Gaussian densities characterised by the mean vector 7_i and the covariance matrice $™_i$.

Diagonal covariance matrices are used in this study. The model parameters $6 = \{p_i, 7_i, ™_i\}$ (i = 1, .., m) are estimated by an EM algorithm [8]. A component density of the GMM can be view like a modelization of an acoustic class, without utilisation of explicit speech segmentation and label data during the model estimation.

2.2. AR-Vector Models

The AR-Vector modelling is a classical tool used to process a signal with various components. It is used here to describe the trajectories of the cepstral vectors of speech.

Let $\{y_n\}$, (n = 1, .., N) be a succession of N p-dimensional cepstral vectors. Their evolution is described by an auto-regressive vectorial model of order q.

$$y_n = \sum_{i=1}^{q} A_i \, y_{n-i} + \varepsilon_n$$

where $\{A_i\}$, (i = 1, .., q) are p x p matrices,

$\{\varepsilon_n\}$, (n = 1, .., N) is a vectorial white noise which has a covariance matrice D.

The coefficients of matrices A_i are estimated by the algorithm of Levinson-Whittle-Robinson. The criterion to minimise is the trace of the covariance matrice D. Applied to speaker recognition [2, 3, 9, 10], they are interpreted as a representation of articulatory capacities of the speaker. The chosen inter-speakers measure IS_I [2] is based on the symetrized Itakura measure [11].

In order to use together two different modelizations, we have to normalise their measures.

3. GMM AND ARVM MEASURE NORMALIZATION

The measure used, for the GMM modelization, depend on the training models and the test sentence. The formulas are the following :

$$Mes_{GMM}(L) = \sum_{n=1}^{N} Mes_{GMM}(y_n, L)$$

$$Mes_{GMM}(y_n, L) = -\log\left(p(y_n | \lambda_L)\right)$$

where N : number of cepstral vectors,
$\{y_i\}$ (i = 1,..,N) : succession of N cepstral vectors of the test sentence,
λ_L : the GMM speaker L model,
Mes_{GMM} (L) : measure on the whole test sentence with the speaker L GMM model,
Mes_{GMM} (y_n, L) : measure on the vector y_n with the speaker L GMM model,
Mes_{GMM} (L) and Mes_{GMM} (y_n, L) 0 [0; +T[.

Mes_{GMM} (L) is used for the global approach, and the Mes_{GMM} (y_n, L) for the analytical approach.

For the ARVM in addition to the training model, a model is estimated on the test sentence. The formulas for the ARVM are the following :

$$Mes_{ARVM}(L) = IS_I(L)$$

$$Mes_{ARVM}(y_n, L) = \sum_{i=1}^{q} A_i^{test} y_{n-i} - \sum_{i=1}^{q} A_i^{L} y_{n-i}$$

where $\{A_i^L\}$ (i = 1,..,q) : ARVM model of the speaker L,
$\{A_i^{test}\}$ (i = 1,..,q) : ARVM model of the test sentence,
Mes_{ARVM} (L) : measure on the whole test sentence with the speaker L ARVM model and the test sentence ARVM model,
Mes_{ARVM} (y_n, L) : measure on the vector y_n with the speaker L ARVM model and the test sentence ARVM model,
Mes_{ARVM} (L) and Mes_{ARVM} (y_n, L) 0 [0; +T[.

In order to have the same measure dynamic, we normalise both GMM and ARVM measures. The same normalization is used on each modelization.

The normalized measures are defined by the following formula:

$$MesN_{model}(L) = Mes_{model}(L) \Big/ \sum_{l \in K_{NN}(test)} Mes_{model}(l)$$

$$MesN_{model}(y_n, L) = Mes_{model}(y_n, L) \Big/ \sum_{l \in K_{NN}(test)} Mes_{model}(y_n, l)$$

where model : the GMM or the ARVM modelization,
K_{NN}(test) : a set of the K nearest neighbour speakers on the test sentence of the speaker which have the lower measure Mes_{model}.

This normalization directly depend on the proximity of the K_{NN} speakers of the nearest speaker on the test sentence and can be used with different modelizations. In order to avoid abnormal high measures, we don't sum all measures for the normalization. We use this normalization in all cooperation/competition experiments presented in this paper.

4. COOPERATION/COMPETITION

In this section, the experiments are carried out on whole sentences for both training and test steps. A separation speech/noised on the signal energy allow us to used only speech material.

The first two experiments are a cooperation and a competition of the GMM and ARVM on a global approach.

4.1. Global Approach

In this approach we calculate a global normalized measure on the test sentence for both GMM and ARVM.

For the cooperation of GMM and ARVM, the final decision D_{coop} is based on a sum of both GMM and ARVM normalized measure on the test sentence. D_{coop} is defined by the following formula:

$$D_{coop} = \underset{L=(1,..,N_L)}{Argmin} \left[MesN_{ARVM}(L) + O\left(\alpha, MesN_{GMM}(L)\right)\right]$$

with N_L : number of speaker,
$MesN_{ARVM}$ (L) : normalized measure on the test sentence with the speaker L ARVM model,
$MesN_{GMM}$(L) : normalized measure on the test sentence with the speaker L GMM model,
α : threshold on the GMM normalized measure,
$O(\alpha, MesN_{GMM}(L)) = MesN_{GMM}(L)$ if $MesN_{GMM}(L) < \alpha$, otherwise 0.

For the competition between GMM and ARVM, the final decision D_{comp} is based on the best normalized measure on the test sentence between GMM and ARVM measure. D_{comp} is defined by the following formula :

$$D_{comp} = \underset{L=(1,..,N_L)}{Argmin} \left[\min\left(MesN_{ARVM}(L), \beta \times MesN_{GMM}(L)\right)\right]$$

with β : weighting coefficient on the GMM normalized measure.
The threshold α and the weighting coefficient β are both used to adapt the GMM measure contribution to the ARVM measure.

For their determinations we separate the training set in two parts, 6 sentences of each speaker for the speaker models training and the two remaining sentences for the estimation of α and β.

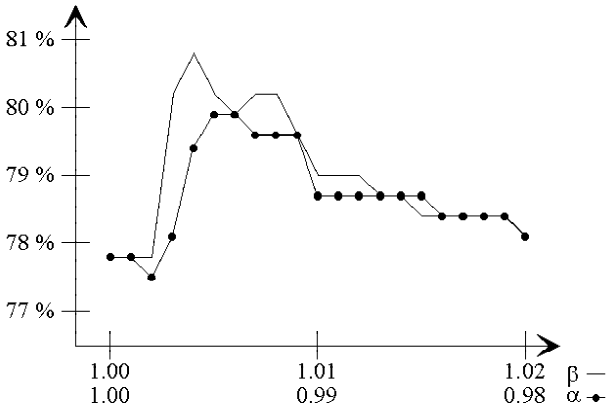

Figure 1 : Identification percentages of cooperation and competition of GMM and ARVM with the variation of the threshold α and the weighting coefficient β.

The α and β variations effects on the identification results in the cooperation (α) and the competition (β) of the GMM and ARVM show the control importance on the different modelizations contributions (figure 1).

Modelizations	Identification
GMM	61.7 %
ARVM	78.1 %
Cooperation	79.6 %
Competition	80.8 %

Table 1 : Identification percentages of GMM, ARVM, cooperation and competition of GMM and ARVM on 168 speakers of the NTIMIT database for the global approach.

The best result is obtain by the models competition (80.8 % vs. 79.6 % for models cooperation). These results improvements in models cooperation (79.6 % vs. 78.1 %) and in models competition (80.8 % vs. 78.1 %) show that the utilisation of different modelizations as GMM and ARVM is promising in speaker identification.

In the following section we present an analytical approach which has a higher computing cost but could have a better exploitation of the differences between the GMM and the ARVM modelization.

4.2. Analytical Approach

In the analytical approach, the final speaker identification decision is made with all the local identification speaker tests realised for each cepstral vectors.
For the cooperation of GMM and ARVM, the final decision D_{an_coop} is based on a sum of both GMM and ARVM normalized measures on each vector of the test sentence. D_{an_coop} is defined as :

$$D_{an\_coop} = \underset{L=(1,..,N_L)}{\text{Argmin}} \left[\sum_{n=1}^{N} \begin{pmatrix} MesN_{ARVM}(y_n, L) \\ + O(\alpha,\ MesN_{GMM}(y_n, L)) \end{pmatrix} \right]$$

where $MesN_{GMM}(y_n, L)$: normalized measure on the vector y_n with the speaker L GMM model,
$MesN_{ARVM}(y_n, L)$: normalized measure on the vector y_n with the speaker L ARVM model,
α : threshold on the GMM measure,
$O(\alpha, MesN_{GMM}(y_n, L)) = Mes_{GMM}(y_n, L)$ if $MesN_{GMM}(y_n, L) < \alpha$, otherwise 0.

For the competition of GMM and ARVM, the final decision D_{an_comp} is based on the sum of the best GMM and ARVM normalized measures on each vector of the test sentence. D_{an_comp} is defined as :

$$D_{an\_comp} = \underset{L=(1,..,N_L)}{\text{Argmin}} \left[\sum_{n=1}^{N} \min \begin{pmatrix} MesN_{ARVM}(y_n, L), \\ \beta \times MesN_{GMM}(y_n, L) \end{pmatrix} \right]$$

where β is a weighting coefficient on the GMM measure.

For the analytical approach, the determination of α and β show us that the best value for both is 1.0. In this matter, it's like there are no threshold (α) for the models cooperation and no weighting coefficient (β) for the models competition.

Modelizations	Identification
GMM	58.7 %
ARVM	74.8 %
Cooperation	79.9 %
Competition	82.6 %

Table 2 : Identification percentages of GMM, ARVM, cooperation and competition of GMM and ARVM on 168 speakers of the NTIMIT database for the analytical approach.

The decrease in performances of the GMM and ARVM due to the normalization of the measures on each cepstral vectors before the measures sum don't seem to have an effect on the cooperation/competition of the GMM and ARVM. The results of the analytical approach are better than those of the global approach (79.9 % vs. 79.6 % in the models cooperation and 82.6 % vs. 80.8 % in the models competition). As in the global approach, the models competition obtain the best results. The difference between global and analytical approaches results could be due to a best exploitation of the different speaker characteristic informations extracted by the GMM and the ARVM in the analytical approach.

In previous study, we shown that speaker identification on selected phonetic segments allows an improvement performance on the ARVM. In the following section, we present cooperation/competition of GMM and ARVM on selected phonetic segments.

5. EXPERIMENTS ON SELECTED PHONETIC SEGMENTS

A speaker verification system could be mislead by an speech record of the claimed person. To solve this problem, we suggest a verification system which impose on the speaker a different sentence chosen randomly for each access. The sentence verification module would be based on a HMM modelization of the impose sentence. This module would allow to obtain a realisation probability of the impose sentence and a phonetic segmentation of the pronounced sentence.

Some phonetic segments contain more speaker related informations [4]. In previous study [7], the bests results were obtained with the vowels, the diphtongues, the nasals, the liquids and the glides. In the NTIMIT database, these phonetic segments correspond on average to 57 % of the sentence.

For these experiments we used the phonetic transcription of the NTIMIT database.

Modelizations	Identification	
	Global	Analytical
GMM	60.5 %	58.7 %
ARVM	80.2 %	73.7 %
Cooperation	81.7 %	79.6 %
Competition	81.4 %	80.5 %

Table 3 : Identification percentages of GMM, ARVM, cooperation and competition of GMM and ARVM on 168 speakers of NTIMIT database for both global and analytical approaches.

For the global approach, selected phonetic segments allow to improve the results of the ARVM, but bring about a decrease in the GMM performances. After all, the cooperation and competition results of the GMM and ARVM modelizations are better than on the wholes sentences (81.7 % vs. 79.6 % and 81.4 % vs. 80.8 %).

For the analytical approach, we can see a decrease in the performances of the ARVM which involves a lesser improvement of GMM and ARVM cooperation and competition the on the whole sentences.

In conclusion, selected phonetic segments allow to improve the GMM and ARVM cooperation and competition results, compared to results on the whole sentences, only with the global approach.

6. CONCLUSIONS

We show that good results text-independent speaker recognition on speech phone quality can be obtain. We describe cooperations and competitions of two different natures modelizations, the GMM and ARVM modelizations, through two approaches : global and analytical. Both models cooperation and competition allows results improvements in global and analytical approaches. The good results of the analytical approach, 82.6 % on 168 speakers of the NTIMIT database, show that there are different instantaneous characteristic informations related to speakers extracted by different modelizations. Experiments using only vowels, diphtongues, nasals, liquids and glides (on average 57 % of each sentence) give best results than on the whole sentence, but only with the global approach. These results show that hybrid systems based on the cooperation/competition of different nature modelizations are promising solutions for the search of a robust speaker recognition. Now we investigate the integration of a text verification and phonetic segmentation based on phonetic HMM modelization in a verification system, in order to reject speech record of the claimed identity speaker.

REFERENCES

1. Reynolds A., "Speaker Identification and Verification using Gaussian Mixture Speaker Models", Workshop on autom. speaker recog. ident. verif. Proc., Martigny, pp. 27-30, 1994.

2. Montacié C., and Le Floch J.-L., "AR-Vector Models for Free-Text Recognition", ICSLP 92, Banff, vol. 1, pp. 611-614, 1992.

3. Montacié C., Le Floch J.-L., and Caraty M.-J., "Procédé et dispositif d'un contrôle d'accès par la vérification de la voix", European patent application, 1996.

4. Le Floch J.-L., Montacié C., and Caraty M.-J., "Investigations on speaker characterisation from ORPHÉE system technics", IEEE-ICASSP Adelaïde, vol. S_1, pp. 149-152, 1994.

5. Fisher W., Zue V., Bernstein J., and Pallet D., "An Acoustic-Phonetic Data Base", J. Acoust. Soc. Amer. Suppl. (A), 81, S92, 1986.

6. Reynolds A., and Rose C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. on Speech and Audio Processing, vol. 3, pp. 72-83, 1995.

7. Le Floch J.-L., Montacié C., and Caraty M.-J., "Coopération et Compétition de Modèles en reconnaissance du locuteur", XXI^{ème} JEP, 1996.

8. Dempster A., Laird N., and Rubin D., "Maximum likelihood from incomplete data via the EM algorithm", J. Roy. Statist. Soc., vol. 39, pp. 1-38, 1977.

9. Artières T.; Bennani Y., Gallinari P., and Montacié C., "Connectionist and Conventional Models for Free-Text Talker Identification Tasks", Neuronimes, Nimes, 1991.

10. Montacié C., and Le Floch J.-L., "Discriminant AR-Vector Models for Free-Text Speaker Verification", Eurospeech 93, vol. 3, pp. 161-164, 1993.

11. Itakura F., "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. ASSP, vol. 23, pp. 67-72, 1975.