SPEAKER MODELING FROM SELECTED NEIGHBORS APPLIED TO SPEAKER RECOGNITION

Yassine Mami & Delphine Charlet

France Telecom R&D
DIH/IPS, 2 av. Pierre Marzin
22307 Lannion Cedex, FRANCE
{yassine.mami, delphine.charlet}@rd.francetelecom.com

ABSTRACT

This paper addresses the estimation of a speaker GMM through the selection and merging of a set of neighbors models for that speaker. The selection of the neighbors models is based on the likelihood score for the training data on a set of potential neighbor GMM. Once the neighbors models are selected, they are merged to give a model of the speaker, which can also be used as an a priori model for an adaptation phase. Experiments show that merging neighborhood models captures significant information about the speaker but doesn't improve significantly compared to classical UBM-adapted GMM.

1. INTRODUCTION

Recently, it has been proposed to model a speaker relatively to a set of others speakers. This concept was first developed in the case of speaker adaptation for speech recognition, where eigenvoices [1] or speaker clustering [2] constitute promising ways to perform fast speaker adaptation. In speaker recognition, we have explored how the relative position of a speaker with respect to a set of reference speakers can directly be used for speaker recognition, through the Anchors Models concept [3].
In this paper, the relative position of a speaker with respect to a set of reference speakers is used to estimate a GMM for that speaker, thus in approaches that are closer to fast speaker adaptation based on speaker clustering. These latter approaches can be divided roughly into two major families. The first approach [2] [4] consists in building offline speaker clusters and during the adaptation phase, the system assigns the speaker to a cluster (or a weighted set of clusters) and derives a speaker-adapted model from the clusters. The second approach consists in building the cluster during the adaptation phase itself, e.g. by selecting the nearest neighbors [5]. This latter approach is investigated here for GMM-based speaker recognition.
This paper is structured as follows: in the next sections, we recall the GMM speaker recognition system and the state-of-the-art UBM-adapted GMM. In section 4, the description of the estimation of GMM from speaker neighborhood models is described. Then, this system is evaluated and compared with classical GMM in section 5, followed with a discussion on the results and a conclusion.

2. GMM SPEAKER RECOGNITION

In GMM speaker recognition, a speaker λ is modeled with the mixture weights, mean vectors and covariance matrices:

$$\lambda = \{p^i, \mu^i, \Sigma^i\}$$

where $i = 1, \cdots, M$ are the component densities.
To evaluate GMM, we use two basic decision processes: closed-set identification and verification.
The identified speaker $\widehat{s}$ (which pronounced the utterance X) is the one that corresponds to a maximum likelihood score:

$$\widehat{s} = \arg \max_{1 \leq s \leq \mathcal{E}} \log p(X|\lambda_s) \quad (1)$$

where $\mathcal{E}$ is the set of speakers to be identified.
In speaker verification, we evaluate a normalized likelihood score which will be compared to a threshold to make a verification decision:

$$\Lambda = \log p(X|\lambda) - \log p(X|\lambda_{UBM}) \quad (2)$$

where λ_{UBM} is the Universal Background Model.
In the following sections, we focus on the estimation of GMM parameters for each speaker.

3. UBM-ADAPTED GMM

State-of-the-art text independent speaker recognition is based on UBM-adapted GMM [6]. It can be seen as a simplified bayesian adaptation with a particular choice of the density a priori parameters. With a fixed adaptation speed for all parameters (weights, means and variances), the estimation formulae, for gaussian i are:

- Weights:

$$p^i = \frac{n^i_{UBM} + n^i_\lambda}{\sum_j (n^j_{UBM} + n^j_\lambda)} \quad (3)$$

- Means:

$$\mu^i = \frac{n^i_{UBM} \overline{X}^i_{UBM} + n^i_\lambda \overline{X}^i_\lambda}{n^i_{UBM} + n^i_\lambda} \quad (4)$$

- Variances:

$$\sigma^{2^i} = \frac{n^i_{UBM} \overline{X^i_{UBM} X^{i\,T}_{UBM}} + n^i_\lambda \overline{X^i_\lambda X^{i\,T}_\lambda}}{n^i_{UBM} + n^i_\lambda} - \mu^i \mu^{i\,T} \quad (5)$$

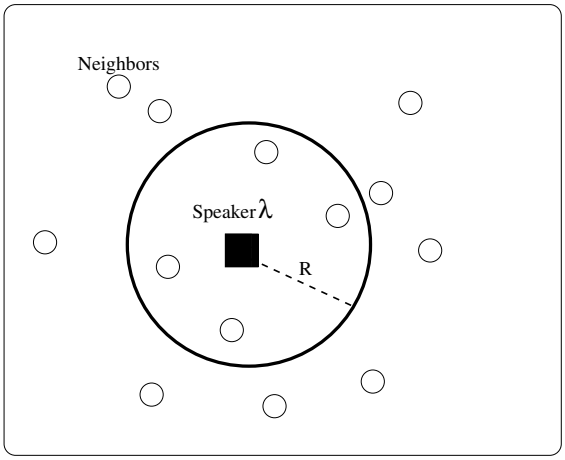

Fig. 1. *Determination of the optimal neighbors in a fixed radius.*

where n^i_{UBM} is the weight assigned to the UBM model in the adaptation. Thus it controls the adaptation speed. It is an empirical factor, fixed to 10 in our experiments (which is consistent to the range [8-20] mentioned in [6]).

4. NEIGHBORHOOD-ADAPTED GMM

The idea consists in determining, during the training phase, a set of nearest neighbors for the speaker (among a set of offline well-trained speakers), and merging their models to get the model of the speaker.
Doing this, we hope to inherit some information for the modeling that we weren't able to estimate from few training data. This is based on the assumption that we can robustly determine the neighborhood with few training data.
In the following, we describe how to determine the neighborhood and then, how to obtain a model for the speaker by merging the models of the neighbors.

4.1. Selection of the neighbors

During the training phase, the neighborhood is determined for each speaker λ using the likelihood score of the training data of speaker λ computed for each well-trained speaker model γ_i:

$$d = -\log p(X|\gamma_{i=1,\cdots,\mathcal{E}})$$

where d is compatible with a distance.
Then, we can select either:

- a fixed number of neighbors and we retain only N first neighbors having the highest score (or the smallest distance).
- or neighbors within a fixed radius R (see figure 1). The optimal neighbors are those which have a distance $\leq$ than R.

4.2. Merging the neighbors models

Once the set of N neighbors models $\{\gamma_k\}_{k=1,\cdots,N}$ is determined for speaker λ, for each gaussian i, the parameters can be estimated in the following way:

- Weights:

$$p^i = \frac{1}{N}\sum_{k=1}^{N} p^i_k \tag{6}$$

- Means:

$$\mu^i = \frac{1}{N}\sum_{k=1}^{N} p^i_k \mu^i_k \tag{7}$$

- Variances:

$$\sigma^{2\,i} = \frac{1}{N}\sum_{k=1}^{N} p^i_k \left[\sigma^{i\,2}_k + \mu^{i\,2}_k\right] - (\mu^i)^2 \tag{8}$$

This merging makes sense if only the gaussian i of the neighbor model A can be associated to the gaussian i of the neighbor model B, that is to say if only they correspond roughly to the same location in the acoustical space. We will question this point in the discussion later.
In the experiments, we call this model "neighborhood-merged model".

4.3. Adapting the merged-model

Once the model of speaker λ is obtained by merging the models of his neighbors, it is still possible to adapt this model just like in the UBM-adapted procedure, except that the initial parameters of the UBM are replaced with the parameters of the merged model.
In the experiments, we call this model "neighborhood-merged and adapted model".

5. EXPERIMENTS AND RESULTS

This section presents the experimental evaluation of the text-independent speaker identification and verification using the neighborhood-merged model and neighborhood-merged and adapted model.
In our experiments, we have used a France Telecom R&D telephone speech database organized in the following way:

- Subset $\mathcal{E}_1$ of 50 speakers to be recognized (composed of 33 female and 17 male speakers).
- Subset $\mathcal{E}_{UBM}$ of 500 speakers used to train the UBM (about 75 seconds of speech per speaker)
- Subset $\mathcal{E}_2$ used to select the neighborhood models: 200 speakers among $\mathcal{E}_{UBM}$

The acoustic space vectors are composed of 27 coefficients (energy and the first 8 MFCC, plus their first and second derivatives). The number of gaussian functions in the GMM is fixed to 64. The speaker models for $\mathcal{E}_2$ are adapted from the UBM. The sentences of this database are read and were extracted from the french newspaper "*Le Monde*".
For closed-set identification we make one test per sentence (its approximate duration is about 5 seconds) that makes more than 6000

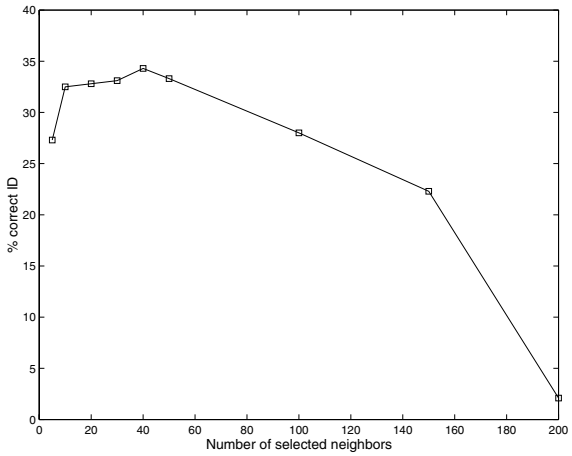

Fig. 2. *Neighborhood-merged model: speaker identification performance versus number of selected neighbors.*

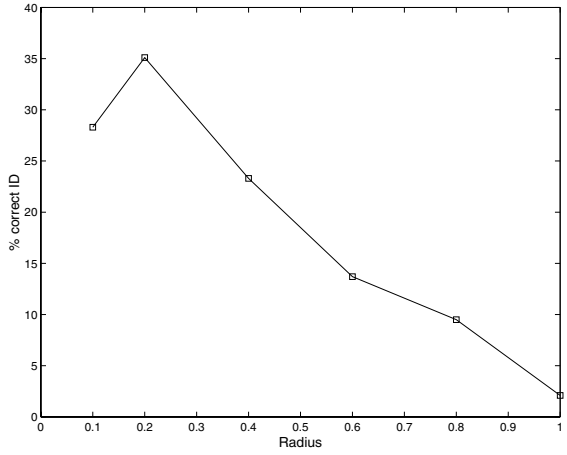

Fig. 3. *Neighborhood-merged model: speaker identification performance versus neighbors radius.*

tests. For verification, we make 6000 claimant tests and 6000 imposter tests.
As we expected the localization of the neighborhood to be robust to sparse training data, we focus the study on the case where only one sentence (roughly 4 seconds of speech) is available to train the model for each speaker of $\mathcal{E}_1$.

5.1. Neighborhood-merged model

Figure 2 shows a plot of the correct identification performances versus the number of selected neighbor models. The 50 speakers of $\mathcal{E}_1$ were trained on about only 4 seconds of speech. It can be seen that speaker identification system reaches its best performances when we have 40 selected speakers.
Of course, when all the 200 speakers are selected, their merged model is the same for every speaker so the performances obtained are those of a random classifier: 2% of correct identification for 50 speakers.
If the optimal neighbors are selected within a fixed radius R, the scores of each speaker are translated and normalized so that the closest neighbor has a distance equal to 0 and that the furthest neighbor has a distance equal to 1.
In figure 3, we have plotted identification rates versus the neighbors radius. This figure shows that the identification system improves slightly the performances if the radius is equal to 0.2. For a given value of the radius, the neighbors number is different for each speaker. For this value ($R = 0.2$), there is on average 20 neighbors (the minimum is 2 and the maximum is 44 neighbors).
Although these results are not as good as those of UBM-adapted identification (58.5% of correct identification), they show that merging neighbors models captures significant information about speaker. This result is also observed in speaker verification, we obtain an EER of 12.9% with the neighborhood-merged model against an EER of 11.9% with the UBM-adapted GMM.

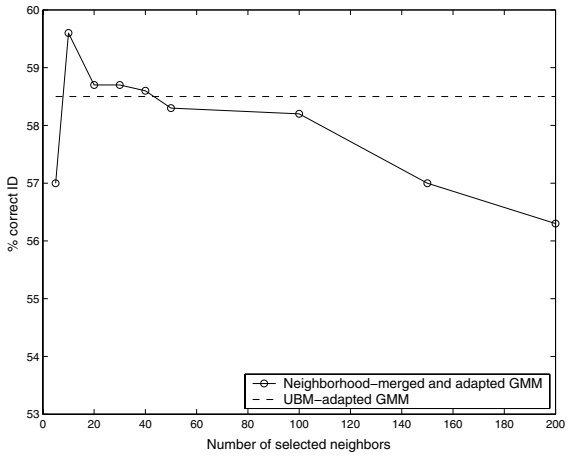

Fig. 4. *Neighborhood-merged and adapted model: speaker identification performance versus selected neighbors.*

5.2. Neighborhood-merged and adapted model

In these experiments, we adapt the merged model according to procedures of paragraph 4.3. The speaker models are still trained with about 4 seconds of speech.
The figure 4 shows that the optimal neighbors number is 10. At this value, the correct identification rate (59.6%) is slightly better than that of the UBM-adapted GMM (58.5%). We obtain also these performances if we select speaker neighborhood in a radius of 0.2 (see figure 5).
In addition, the neighborhood-merged and adapted model procedure gives an EER of 13.7% against 13.9% with the UBM-adapted GMM.
As can be observed in Figures 4 and 5, this approach does not seem to give great improvements compared to the GMM approach. An interesting point is the one when we select all the 200 neighbors. In this case, the merged model is the same for every speaker, and it is equivalent to an alternative UBM. We observe a slight degradation of the performances for this 200-speaker-merged model compared with the UBM. That means that by merging models, we have de-

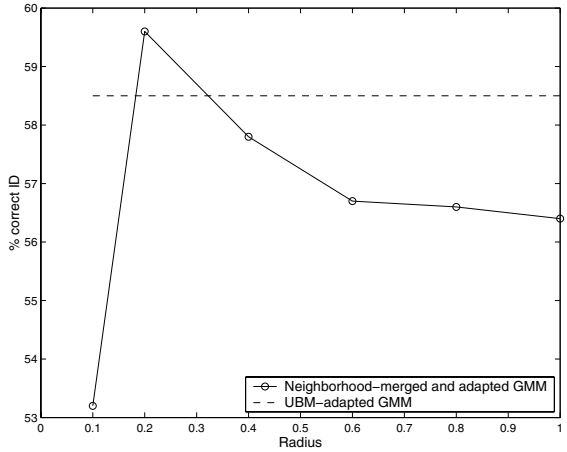

Fig. 5. *Neighborhood-merged and adapted model: speaker identification performance versus neighbors radius.*

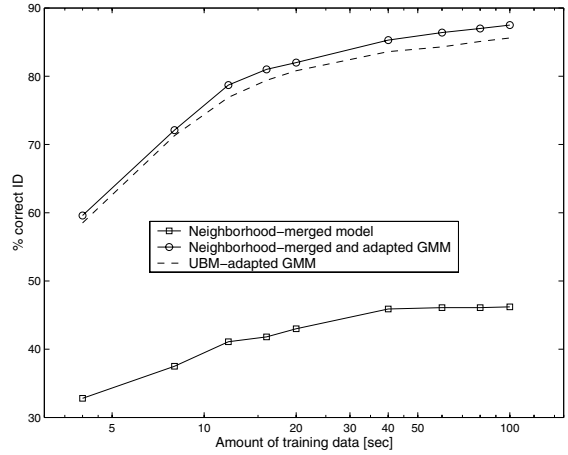

Fig. 6. *Speaker identification performance versus amount of training data.*

graded the acoustic representation of the model. One possible explanation should be that the correspondence between the merged gaussians is not guaranteed anymore.

5.3. Varying amount of training data

In figure 6, we plot the identification performances for different amounts of training data (recorded during a single call) for the UBM-adapted GMM and the neighborhood-merged and adapted GMM. We can see that the benefit of the localization increases significantly when the amount of training data increases (until about 40 seconds of speech), even with the merged models.
Differences in the merged models are only due to a different neighborhood according to the training data. So, contrary to what we expected, the selection of the neighborhood requires a certain amount of training data to be precise.

6. CONCLUSION

In this paper, we have presented a speaker modeling technique which consists in selecting the nearest neighbors and merging their models to get a new model of the speaker.
Experiments show that merging neighborhood models captures significant information about the speaker, but this information doesn't help much to estimate better model when it was coupled to a classical adaptation. One possible explanation is that we have on the one hand improved speaker representation and on the other hand degraded acoustical representation by merging gaussians. That's why further study should focus on the correspondence between the gaussians to be merged or by using a more constraint GMM, e.g. a phonetic-class GMM.

7. REFERENCES

[1] R. Kuhn, J-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, 2000.

[2] M. Padmanabhan, L.R. Bahl, D. Nahamoo, and M.A. Picheny, "Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems," in *International Conference on Acoustics, Speech and Signal Processing*, 1996, p. 701.

[3] Yassine Mami and Delphine Charlet, "Speaker identification by location in an optimal space of anchor models," in *International Conference on Spoken Language Processing*, 2002, vol. 2, p. 1333.

[4] E.J. Pusateri and T.J. Hazen, "Rapid speaker adaptation using speaker clustering," in *International Conference on Spoken Language Processing*, 2002, pp. 61–64.

[5] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, A. Lee, and K. Shikano, "Evaluation on unsupervised speaker adaptation based on sufficient hmm statistics of selected speakers," in *Eurospeech*, 2001, p. 1219.

[6] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.