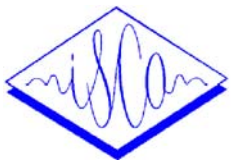
Neighborhood-adapted GMM for speaker recognition



Delphine Charlet

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

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Abstract

In previous work [1], it was investigated how the neighborhood can be used to estimate a better model for a speaker when few training data is avalaible. In this paper, this work is completed by investigating another way to merge models from the neighbors and by introducing a weight on the neighbor models to be merged. Experiments on a telephone speech database show that using the neighborhood-merged model to initialize the training phase provides improvement compared to the UBM approach, when few training data is available.

1. Introduction

Speaker clustering is an interesting approach for fast speaker adaptation. It is based on the assumption that a speaker can inherit, in the acoustic modeling, some characteristics of the cluster of speakers he belongs to. Fast adaptation from speaker clustering can be divided roughly into two families. The first approach, e.g. [2], consists in building off-line the clusters, and during the adaptation phase, the system determines the cluster (or the weighted set of clusters) that corresponds to the speaker and then performs fast adaptation. The second approach consists in building the cluster during the adaptation phase itself, e.g. by selecting the nearest neighbors [3]. This is the approach which was explored for speaker recognition in [1]. This work is completed in the present paper.

The paper is structured as follows: in section 2, the principles of GMM speaker recognition which will be the baseline system are briefly recalled. Then , section 3 presents the approaches which have been explored for deriving a speaker model from a set of neighbors models. In section 4, a weighting function on the neighbors in the merging process is introduced. Finally, experiments are presented in section 5 before a conclusion.

2. GMM-based speaker recognition

In this section, the basis of GMM-based speaker recognition are briefly recalled. In GMM speaker recognition, a speaker λ is modeled with the mixture prior probabilities, mean vectors and covariance matrices:

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

where $i = 1, \cdots, M$ are the component densities.
One basic decision process for speaker recognition is closed-set identification. The identified speaker $\widehat{s}$ (which pronounced the utterance X) is the one that corresponds to a maximum likelihood score (assuming equiprobabilities of speakers):

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

where $\mathcal{E}$ is the set of speakers to be identified.

Within this framework, the important point is the way the parameters $\{p^i, \mu^i, \Sigma^i\}$ are estimated. State-of-the-art systems use UBM-adapted GMM [4] which relies on a simplified bayesian adaptation procedure with a particular choice of the density a priori parameters. With a fixed adaptation speed for all parameters (prior probability, means and variances), the estimation formulae, for gaussian i are:

- Prior probabilities: $p^i = \frac{n^i_{UBM} + n^i_\lambda}{\sum_j (n^j_{UBM} + n^j_\lambda)}$
- Means: $\mu^i = \frac{n^i_{UBM} x^i_{UBM} + n^i_\lambda x^i_\lambda}{n^i_{UBM} + n^i_\lambda}$
- Variances: $\sigma^{2i} = \frac{n^i_{UBM} {x^i_{UBM}}^2 + n^i_\lambda {x^i_\lambda}^2}{n^i_{UBM} + n^i_\lambda} - {\mu^i}^2$

where n^i_{UBM} is the weight assigned to the UBM model parameters in the adaptation process. Thus it controls the adaptation speed. It is an empirical factor, fixed to 10 in our experiments (which is consistent with the range [8-20] mentioned in [4]). Usually, the number of mixtures is high (e.g. 2048) and only the means of the gaussians are adapted. In our experiments, where only 256 gaussians are used, we got slightly better results when adaptating all parameters.

3. Neighborhood-adapted GMM

3.1. Principle

The idea consists in determining, during the training phase, a set of nearest neighbor models for the speaker (among a set of offline well-trained speaker models), and merging their models to get the model of the speaker.
Doing this, we hope to inherit some information for the modeling that it would not have been possible to estimate properly from few training data. This is based on the assumption that the neighborhood can be robustly determined with few training data.

In [1], we searched how to determine the neighborhood. During the training phase, the neighborhood is determined for a given speaker λ using the likelihood score of the training data X_λ of speaker λ computed for each well-trained speaker model γ_i:

$$d(\gamma_i) = -\log p(X_\lambda|\gamma_i)$$

where d is compatible with a distance.
To determine the neighborhood size, two approaches were experimented: one based on the K-nearest neighbors, the other based on the neighbors whose distance from the speaker is less than a fixed radius. Both approaches gave similar results, thus, for sake of simplicity, in this work the neighborhood is determined by selecting the K-nearest neighbors.

Once the neighborhood is found for a given speaker, the question is how to merge the K GMMs of the neighbors to get one GMM for the speaker. In the following, we recall the way the neighbor models are merged in [1], called index-based merging, and we present another way to merge the neighbor models, based on a distance criterion.

3.2. Index-based merging

When the GMM for each speaker is mean-only UBM-adapted, it can be assumed that the ith gaussian for speaker γ_m corresponds to the same acoustic event as the ith gaussian for speaker γ_n. This assumption is made for instance in [5], where decoding is done with UBM and rescoring is made with the corresponding gaussians in the speaker-adapted models. The good results obtained with this method confirm that this assumption is valid. Nevertheless, this assumption becomes more questionnable when all the parameters (and not only the means) of the mixtures are adapted

Hence, with this assumption, the new ith gaussians for the speaker is computed by averaging all the ith gaussians from the neighbors (all the gaussians with index i). The new parameters are then given by:

- Prior probabilities: $p^i = \frac{1}{K} \sum_{k=1}^{K} p_k^i$
- Means:
  $\mu^i = \frac{1}{K} \sum_{k=1}^{K} p_k^i \mu_k^i$
- Variances:

$$\sigma^{2^i} = \frac{1}{K} \sum_{k=1}^{K} p_k^i \left[{\sigma_k^i}^2 + {\mu_k^i}^2 \right] - (\mu^i)^2$$

3.3. Distance-based merging

In this approach, all the gaussians of the neighborhood GMMs are pooled together. Hence, if the neighborhood is made with K speakers, whose model size is N, a huge GMM of $K * N$ gaussians is obtained. The problem is then to reduced this model to obtain a GMM of N gaussians. The merging is now based on an iterative procedure which merges at each step the 2 closest gaussians to give a new one. The procedure is stopped when a given number of gaussians is reached. The distance between two gaussian densities is measured as the decrease in the likelihood of the corresponding training set when replacing the two gaussians by the merged one [6].

The merged gaussian k is estimated as if all the n_i and n_j samples that were assigned to gaussian i and j were now assigned to the new gaussian k. The parameters of the gaussian k are then given by:

- weights: $n_k = n_i + n_j$
- means : $\mu_k = n_i'.\mu_i + n_j'.\mu_j$
- variances :
  $\sigma_k^2 = n_i'.\sigma_i^2 + n_j'.\sigma_j^2 + n_i'.n_j'.(\mu_i - \mu_j)^2$

where: $n_i' = n_i/(n_i + n_j)$ and $n_j' = n_j/(n_i + n_j)$

Hence, the distance Δ between 2 gaussians i, j can be computed as:

$$\begin{aligned} \Delta &= -n_i \sum_{d=1,..,D} \log \sigma_{id} - n_j \sum_{d=1,..,D} \log \sigma_{jd} \\ &+ (n_i + n_j) \sum_{d=1,..,D} \log \sigma_{kd} \end{aligned}$$

where D is the dimension of the acoustical space and k the merged gaussian.

4. Weighting neighbors

In the previous section, each selected neighbor was assigned the same weight in the merging. In this section, a weighting function on the neighbor models to be merged is introduced.

In the index-based merging, instead of assigning a weight of $\frac{1}{K}$ on the gaussians in the reestimation formulae, a weight of w_k is assigned with the constraint $\sum_k w_k = 1$. Then, the reestimation formulae of the index-based merging becomes:

- Prior probabilities: $p^i = \sum_{k=1}^{K} w_k p_k^i$
- Means:
  $\mu^i = \sum_{k=1}^{K} w_k p_k^i \mu_k^i$
- Variances:

$$\sigma^{2^i} = \sum_{k=1}^{K} w_k p_k^i \left[{\sigma_k^i}^2 + {\mu_k^i}^2 \right] - (\mu^i)^2$$

In the distance-based merging, we apply the weighting by assigning new weights on the gaussians in the pool, before the merging procedure. We proceed in 2 steps. First, the weights $n_{i,k}$ assigned to the ith gaussian of speaker k in the pool are normalized so that the total weight for each speaker is the same.

$$n_{i,k} \rightarrow \frac{\sum_{l=1,..,K} \sum_{j=1,..,N} n_{j,l}}{K. \sum_{j=1,..,N} n_{j,l}} . n_{i,k}$$

Then, the weighting factor w_k is applied on the weight of each gaussian $n_{i,k}$:

$$n_{i,k} \rightarrow w_k n_{i,k}$$

We choose a weighting function based on the distance of each of the neighbors from the speaker. It is assumed that the smaller is the distance of the neighbor, the bigger is the weight assigned to this neighbor in the merging. The proposed weighting function is:

$$w_k = \frac{\delta_k}{\sum_{k=1,K} \delta_k}$$

where:

$$\delta_k = d(\gamma_K) - d(\gamma_k)$$

where K is the total number of the selected neighbors, and the neighbors $\{\gamma_k\}_{k=1,..,K}$ are sorted by increasing distance. By definition, this is a decreasing function of k and the weight assigned to the last selected neighbor is 0.

5. Experiments and results

5.1. Experimental set-up

This section presents the experimental evaluation of the text-independent speaker identification using the neighborhood-merged 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 42 coefficients (energy and the first 13 MFCC after cepstral mean substraction, plus their first and second derivatives).
The number of gaussian functions in the GMM is fixed to 256. 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*". The average length of a sentence is 4 s.
For closed-set identification, one test per sentence is made, that makes more than 6000 tests. We postpone evaluation on speaker verification because verification not only focuses on speaker modeling but also on score normalization, which is not treated here. As it is expected that the localization of the neighborhood is robust to sparse training data, the study is focused on the case where only one sentence (roughly 4 seconds of speech) is available to train the model for each speaker of $\mathcal{E}_1$.

5.2. Neighborhood-merged model

Figure 1 plots the correct identification rate for GMM models obtained by merging the neighbor models according to the index-based procedure and the distance-based procedure with or without the weighting function on the neighbors. The figure compares them with classical UBM-adapted GMM for one sentence of training speech data (i.e. one sentence of speech to determine the neighbors). It shows that merging the neighbors models capture significant information about the speaker, whatever the method of merging is. The weighting function of the neighbors provides significant improvement. Nevertheless, performances of the neigborhood merged model still remains worse than classical UBM-adapted GMM.

Notice that, when merging is done without weighting function, the performance of the merged model drops when the number of neighbors increases: when all the 200 neighbors are selected, their merged model is the same for every speaker; hence the performances obtained are those of a random classifier (2% of correct identification for 50 speakers, not shown in figure 1).

On the contrary, when merging is done with the weighting function, as the weighting function depends on the training data of the speaker to be modelled, even when all the 200 neighbors are merged, the resulting model is still characteristic of the speaker to be modelled.

5.3. Neighborhood-merged and adapted model

Once the model of a speaker 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 [1]. Figure 2 plots the correct identification rate for GMM models obtained after adapting the merged models according to the various proposed schedules. The figure shows that initializing model by neighborhood merged model instead of UBM provides significant improvement, confirming our prior assumption that the neighborhood could capture information we cannot estimate otherwise. Weighting neighbors in the merging process still provides improvements.

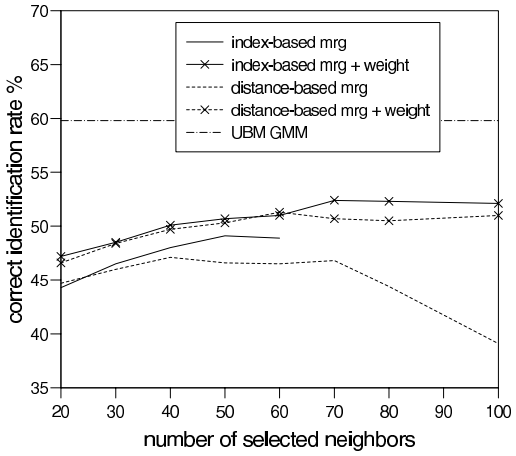


Figure 1: *Neighborhood-merged model: speaker identification performance versus number of selected neighbors.*

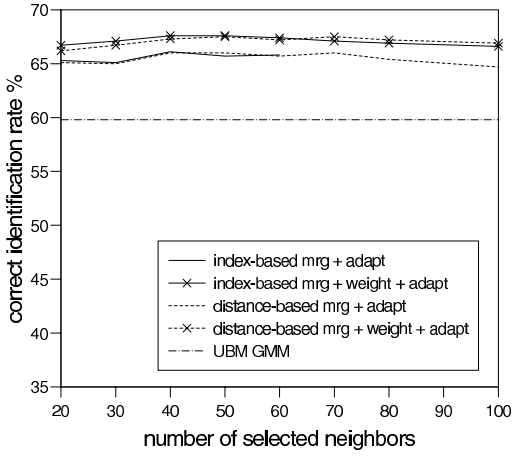


Figure 2: *Neighborhood-merged and adapted model: speaker identification performance versus number of selected neighbors.*

5.4. Influence of the GMM modeling

Some variables in the GMM modeling have been changed between [1] and this work. Indeed, the number of gaussians used (64 in [1], 256 here) and the size of acoustical vector (27 parameters in [1], 42 here) have changed, leading to a higher number of parameters for each speaker model. Results obtained with both modelings are plot in figure 3, where the results obtained in [1] with 64 gaussians and a frame vector of 27 parameters are refered with $x64, c27$ and the current results are refered $x256, c42$. These changes result in a slight improvement of the baseline GMM-UBM speaker modeling. But, above all, it results in increasing the interest of the relative estimation (neighborhood-merged model + adaptation) towards the classical approach: in [1], the relative reduction of identification error rate was less than 3%, whereas here, the reduction is 12%, for the same algorithm (index-based merging without weighting + adaptation), with few training data.

We suppose that 2 factors may explain this change. First, as the number of parameters to be estimated increases, the approaches that help estimate parameters with sparse training data are supposed to be more efficient. Second, as the size of the acoustical vector (hence, the dimension of the gaussian function) is higher, it is less probable that there is a "shift" of the gaussian during the adaptation of the GMM: the assumption that the ith gaussian for speaker γ_m corresponds to the same acoustic event as the ith gaussian for speaker γ_n seems to be valid with high-dimensionnal vector.

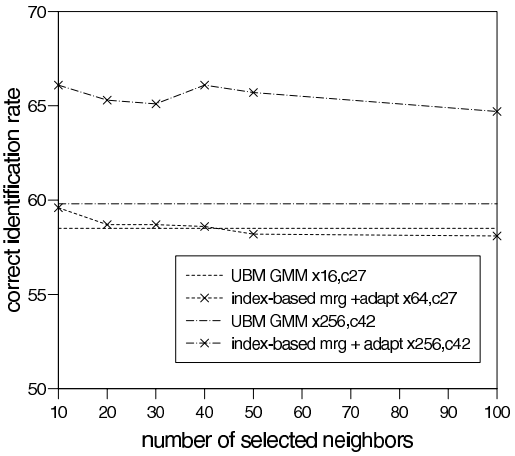


Figure 3: *Neighborhood-merged and adapted model: influence of GMM modeling*

5.5. Influence of training data

The figure 4 plots the correct identification rate for UBM-adapted GMM and neighborhood-adapted GMM (with 50 neighbors) for different amount of training data (from 1 sentence to 25 sentences, all uttered during the same call). Not suprisingly, the interest of the neighborhood-adapted GMM increases as the amount of training data decreases.

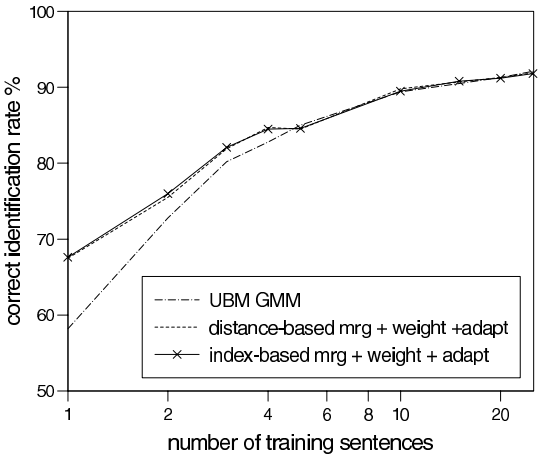


Figure 4: *Neighborhood-merged and adapted model: speaker identification performance versus number of training sentences.*

6. Conclusion

This paper completes a work begun in [1], in which it was questionned how the neighbors of a given speaker can help estimate his model in the GMM framework. Once the neighborhood is determined for a given speaker, we compare 2 methods to get the speaker model from the set of neighbor models. We then introduce a weighting function on the selected neighbors. Experiments on a telephone speech database show that using the neighborhood-merged model to initialize the training phase provides improvement compared to the UBM approach, when few training data is available. Further work should focus on integrating this relative estimation in a suitable background model in the context of speaker verification.

7. References

[1] Y. Mami and D. Charlet, "Speaker modeling from selected neighbors applied to speaker recognition," in *Eurospeech*, Geneva, Switzerland, 2003.

[2] E.J. Pusateri and T.J. Hazen, "Rapid speaker adaptation using speaker clustering," in *ICSLP*, Denver, USA, 2002, pp. 61–64.

[3] 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*, Aalborg, Denmark, 2001, pp. 1219–1222.

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

[5] D.A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification systems," in *Eurospeech*, Rhodes, Greece, 1997, pp. 963–966.

[6] J. Simonin, S. Bodin, D. Jouvet, and K. Bartkova, "Parameter tying for flexible speech recognition," in *ICSLP*, Philadelphia, USA, 1996, pp. 1089–1092.