Digital Signal Processing **10**, 198–212 (2000)
doi:10.1006/dspr.2000.0370, available online at http://www.idealibrary.com on IDEAL®

Segmental Approaches for Automatic Speaker Verification

Dijana Petrovska-Delacrétaz*,[1], Jan Černocký†, Jean Hennebert‡, and Gérard Chollet§

*Circuits and Systems Laboratory, Swiss Federal Institute of Technology, DE-CIRC, 1015 Lausanne, Switzerland; †Institute of Radio-electronics, Brno University of Technology, Czech Republic; ‡UpperSide Consulting, 52 Chaussée de Vleurgat, 1050 Brussels, Belgium; and §TSI Department, CNRS ENST, Paris, France

E-mail: dijana.petrovska@epfl.ch, cernocky@urel.fee.vutbr.cz, jean.hennebert@upperside.com, chollet@tsi.enst.fr

Petrovska-Delacrétaz, Dijana, Černocký, Jan, Hennebert, Jean, and Chollet, Gérard, Segmental Approaches for Automatic Speaker Verification, *Digital Signal Processing* **10** (2000), 198–212.

Speech is composed of different sounds (acoustic segments). Speakers differ in their pronunciation of these sounds. The segmental approaches described in this paper are meant to exploit these differences for speaker verification purposes. For such approaches, the speech is divided into different classes, and the speaker modeling is done for each class. The speech segmentation applied is based on automatic language independent speech processing tools that provide a segmentation of the speech requiring neither phonetic nor orthographic transcriptions of the speech data. Two different speaker modeling approaches, based on multilayer perceptrons (MLPs) and on Gaussian mixture models (GMMs), are studied. The MLP-based segmental systems have performance comparable to that of the global MLP-based systems, and in the mismatched train-test conditions slightly better results are obtained with the segmental MLP system. The segmental GMM systems gave poorer results than the equivalent global GMM systems.

1. INTRODUCTION

Various studies [10, 15, 21, 22] have shown that phonemes have different discriminant power for the speaker verification task. In [22] it was shown that nasals, fricatives, and vowels have better performance than plosives and

[1] Currently at AT&T Laboratories, Speech Research, 180 Park Avenue, Florham Park, NJ 07932. E-mail: dijana@research.att.com.

1051-2004/00 $35.00

liquids. In that study, segmentation in phonemes was achieved using Viterbi forced alignment with 42 context-independent phone hidden Markov models (HMMs). The segmentation step introduces an additional level of complexity in the speaker verification task. Speaker modeling with prior segment selection, based on large vocabulary continuous speech recognition (LVCSR), has already been studied [9, 19, 24, 29]. The conclusions are that such systems provided state-of-the-art performance, given "sufficient" train and test material (2 min for training and 30 s for testing). We believe that the different discriminant power of speech sounds could be exploited in text-prompted speaker verification systems and also when the speaker verification part is embedded in a system including automatic speech recognition.

As an example, let us imagine a banking environment. A customer makes a transaction, via a natural spoken dialogue interface, and the bank would like to also include a speaker verification module for more security. Then, when the customer iniated the dialogue with the machine to express his (or her) requests, a verification of the speaker's identity could be performed in parallel. Because the speech recognition is done for the dialogue manager, it also gives the segmentation in speech classes for the segmental speaker verification module. When the security level must be increased for certain transactions, text-prompted speaker verification could also be applied in a straightforward manner. Such systems have the following advantages: they can provide better security because no prerecorded speech of the client can be used to fool the system, and they offer the opportunity to segment the speech easily into classes via forced alignment. Another advantage could be that knowing which of the speech segments have better performance for the speaker verification task, one can use this information to design the prompt is such a way that it is mostly composed of speech segments that characterize the speaker better.

The example mentioned above requires an efficient speech recognition tool. Although current speech recognition methodology is making progress, the speech recognition systems are still language and task dependent and require collection and annotation of large speech corpora for a specific task. This database generation process is a major difficulty in adapting the systems to new tasks and languages. To avoid this problem, we propose to use automatic language independent speech processing (ALISP) tools [7] that provide a segmentation of the speech requiring neither phonetic nor orthographic transcriptions of the speech data.

In this paper the use of two different speaker modeling methods is described, keeping the speech segmentation preprocessing part the same. In Section 2, a general description of segmental speaker verification systems is given. In Section 2.1, the speech segmentation method is described in more detail. Section 2.2 introduces the general problem of speaker modeling. Section 3 concentrates on the use of MLPs for the segmental speaker modeling. In Section 3.1, a description of the MLP-based system is given, followed by Section 3.2, which depicts the experimental setup. The results are given in Section 3.3. Section 4, devoted to the use of Gaussian mixture models (GMMs)

for the segmental speaker modeling, has the same organization as that of Section 3. The conclusions concerning these two speaker modeling techniques are given in the last section.

2. SEGMENTAL SPEAKER VERIFICATION SYSTEMS

Current text-independent speaker verification systems are usually based on modeling globally the probability density function of the speaker feature vectors. In such systems, denoted here as *global systems*, the temporal information is not taken into account and all classes are represented using a unique speaker model. As noted in the Introduction, another possibility consists of dividing the speech signal into distinct categories (also called classes or segments) and of performing the speaker modeling independently for each class. Such systems are denoted here as *segmental systems*. In such cases, the speaker verification task is divided into two parts: speech segmentation followed by speaker modeling for each of the classes. The general flowcharts of global and segmental speaker verification systems are shown in Fig. 1.

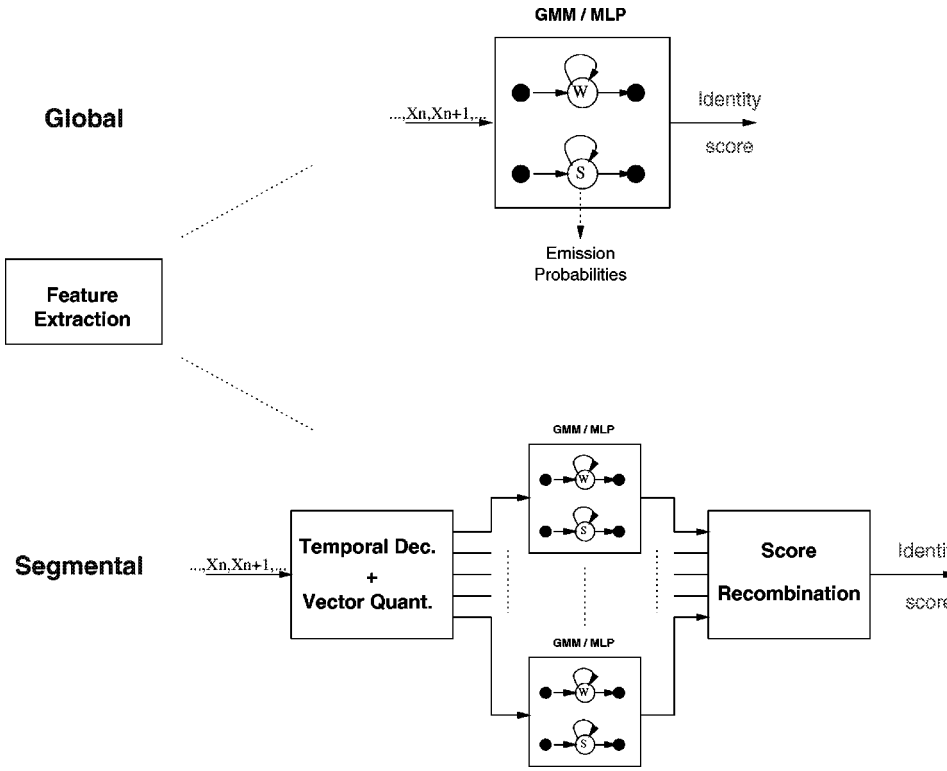

FIG. 1. Global and segmental speaker verification systems.

2.1. Speech Segmentation Based on Temporal Decomposition

The speech segments can be defined using two main approaches:

– The first possibility is to use large vocabulary continuous speech recognition (LVCSR) with previously trained phone models, and a language model, generally a bigram or a trigram stochastic grammar. This recognition segments the speech into phone classes;

– The second possibility is to use data-driven techniques based on automatic language independent speech processing tools [7] that provide a general framework for creating speech units with little or no supervision.

LVCSR systems, although very promising for segmental approaches, require huge phonetically annotated databases. This database generation process is a long and tedious task. It is also dependent on the speech signal characteristics (language and speech quality) and on the required application. For developments of multiple applications in different languages, this annotation phase requires an enormous amount of time and human effort. On the other hand, ALISP offers an alternative when no annotated training data are available, and we decided to use it for the speech segmentation module.

For the ALISP-based automatic speech segmentation, we used the temporal decomposition technique, introduced by Atal [1], followed by unsupervised clustering. The goal of the *temporal decomposition* is to detect quasi-stationary parts in the parametric representation of the speech. With this method, the trajectories of parameters $x_i(n)$ are approximated by a sum of m *targets* a_{ik} weighted by *interpolation functions*:

$$\hat{x}_i(n) = \sum_{k=1}^{m} a_{ik}\phi_k(n), \qquad \text{or} \qquad \underset{(P\times N)}{\widehat{\mathbf{X}}} = \underset{(P\times m)}{\mathbf{A}}\ \underset{(m\times N)}{\mathbf{\Phi}}. \tag{1}$$

In the matrix notation, the lower line indicates matrix dimensions. The initial interpolation functions are found using local singular value decomposition with adaptive windowing [4], followed by postprocessing (smoothing, decorrelation, and normalization). Target vectors are then computed by $\mathbf{A} = \mathbf{X}\mathbf{\Phi}^{\#}$, where $\mathbf{\Phi}^{\#}$ denotes the pseudo-inverse of the matrix. Interpolation functions and targets are iteratively locally refined by minimizing the distance between $\mathbf{X}$ and $\widehat{\mathbf{X}}$. Intersections of interpolation functions define speech segments.

Unsupervised clustering assigns segments to classes. Vector quantization (VQ) is used for automatic determination of classes. A K-means algorithm with binary splitting [12] is used to train the VQ codebook. Training is performed using vectors positioned at the centers of gravity of the interpolation functions, while the quantization takes entire segments into account using cumulated distances between all vectors of a segment and a code-vector. Temporal decomposition and vector quantization provide a symbolic transcription of the data in an unsupervised manner. The temporal decomposition was set to detect 15 events per second on average. This value corresponds to the average phonetic rate, and no experiments with other values were conducted. The vector

quantization is trained on the NIST 1997 data. This segmentation module is kept constant for all the experiments described in this paper.

As the main idea of introducing speech segmentation prior to speaker modeling is to exploit the different speaker discriminant power of different speech sounds, it is of major interest to study the way this segmentation is done. Since correct transcriptions of the evaluation data are not available, we cannot compare the correspondence of ALISP units and the usual phonetic units. This correspondence was studied for one speaker of the Boston University Radio Speech Corpus in [5, 8]. These experiments showed that there is some evidence of correlation of phonemes and ALISP units, in the case of a single speaker. In order to have an exact idea of what this correspondence is, in the case of multiple speakers, more experiments should be done. We should remind the reader here that the main point of our experiments is to investigate the applicability of segmental methods based on automatic segmentation tools. The first point of interest is if the results per class are different. The next step will be a more detailed study of the speech segmentation. Once we have more details about the specificity of the speech classes, we can study in more detail the possibility of different score recombination techniques.

The compromise we are faced with in segmental approaches is the amount of training data available per class. It is well known that the more training material we have, the better the models. If automatically determined speech units correspond to phonemes, the number of classes should approximately equal the number of phonemes. As we have no indications about the correspondence of the ALISP units with the usual phone units, we decided to choose, as a first experiment, not as many ALISP classes as phonetic units to ensure a proper training of all the ALISP classes. This is the reason the number of speech classes is set to 8 for all the experiments presented in this paper.

2.2. Speaker Modeling

The classical way to do pattern classification in text-independent systems is to assign a unique probability density function (pdf) to the whole vector sequence. One way to build the pdf is to use Gaussian mixture models [26] in which the multivariate distribution is modeled with a weighted sum of Gaussian distributions.

Another way to perform classification is to use artificial neural networks (ANNs) [16]. In previous studies [2, 11, 18, 20], ANNs have successfully been used for speaker verification. Among the different ANN architectures, multilayer perceptrons (MLPs) are often used. As explained in [3, 14, 28], the main advantages of MLPs include a discrimination-based learning procedure, a flexible architecture that permits easy use of contextual information, and weaker hypotheses about statistical distributions. The main disadvantages are that their optimal architecture has to be selected by trial-and-error procedures and that the temporal structure of speech signals remains difficult to handle.

Each of these methods can be used for segmental speaker modeling. In the first set of experiments, described in Section 3, we compare global and segmental

speaker verification results based on MLP speaker modeling. Results with state-of-the-art GMM on the same data set are given as a baseline comparison. In Section 4 we compare global and segmental speaker verification results based on GMM speaker modeling.

3. SEGMENTAL MLP SYSTEMS

Because the major problem with segmental systems is the reduced amount of training data (because the training speech is first divided into speech classes), our first choice was to use MLP for the segmental speaker modeling. The reason for this choice is that discrimination-based learning procedures could eventually compensate for the reduced amount of training data available for the segmental systems, because they involve maximizing the likelihood of the speaker data and minimizing the probability of the data belonging to the background (world) speakers.

3.1. MLP-Based System Description

The multilayer perceptrons used for speaker verification purposes are trained to distinguish between the client speaker and the background (world) speakers. Two outputs are used, one for the client class M_{c} and the other for the world class M_{w}. If each output unit k of the MLP is associated to a class category M_k, it is possible to train the MLP to generate a posteriori probabilities $P(M_k|x_n)$ of the state M_k given the acoustic input vector x_n. During the training, the parameters are iteratively updated, via a gradient descent procedure, in order to minimize the difference between the actual and desired outputs. Training is said to be discriminant because it minimizes the likelihood of the incorrect model and maximizes the likelihood of the correct model. Given the hypothesis that the number of parameters of the network and the data are well adapted, and given that the training does not get stuck in a local minimum, the MLP output approximates the a posteriori class probability, $P(M_k|x_n)$. The client score, S_{c}, and the world score, S_{w}, can be computed from the product of the a posteriori probabilities (output of the MLP) divided by the a priori class probabilities $P(M_k)$, computed on the training data. More details concerning pattern classification with neural networks are given in [3, 28].

In the case of global speaker modeling with MLPs, the sequence of N feature vectors x_n is fed into a unique classifier that outputs a score for the client model S_{c} and a score for the world model S_{w}, assuming independence of the observation vectors:

$$S_{\mathrm{c}} = \prod_{n=1}^{N} P(M_{\mathrm{c}}|x_n)/P(M_{\mathrm{c}}), \tag{2}$$

$$S_{\mathrm{w}} = \prod_{n=1}^{N} P(M_{\mathrm{w}}|x_n)/P(M_{\mathrm{w}}). \tag{3}$$

In the log domain, the ratio of the scores can be expressed as the difference of the terms:

$$\Lambda = \log(S_c) - \log(S_w). \tag{4}$$

When Z-normalization [27] is applied, the impostor speakers set used to estimate the impostor distribution should be independent of the background (world) speakers.

For segmental speaker modeling, the speech segmentation is achieved using temporal decomposition and the labeling step is performed with vector quantization, as introduced in Section 2.1. For the training phase, this procedure is applied to the training speech data of the client speaker, dividing it among $L^l : l = 1, \ldots, 8$ speech classes. The speech data of the background (world) speakers are also sorted into eight speech classes. In the training phase, the same technique as for the global MLP modeling is used, and L MLPs are trained for each client, using the client and the world speech data of the corresponding speech classes.

For the test phase, the test speech is also segmented into eight speech classes, and each test speech frame is tested against the corresponding MLP. The MLP trained with the speech data of that client, associated to the speech category L^l, provides the segmental client score, S_c^l, and the segmental world score, S_w^l according to

$$S_c^l = \prod_{x \in L^l} P(M_c^l|x)/P(M_c^l), \tag{5}$$

$$S_w^l = \prod_{x \in L^l} P(M_w^l|x)/P(M_w^l). \tag{6}$$

In the log domain, the ratio of the scores can be expressed as the difference of the terms:

$$\Lambda^l = \log(S_c^l) - \log(S_w^l). \tag{7}$$

When Z-normalization is applied to the segmental systems, it can be applied for each class separately, in the same way as for the global systems. The major inconvenience is that the number of tests that should be performed with the impostors is multiplied by the number of speech classes. The final score Λ, for the segmental systems, is defined by the sum of the segmental scores:

$$\Lambda = \sum_{l=1}^{L} \Lambda^l. \tag{8}$$

3.2. MLP-Based Experimental Setup

Segmental and global MLP-based systems are tested on the NIST1998 database, part of the Switchboard II, Phase II corpus, recorded over telephone lines. The speech is spontaneous and no transcriptions, orthographic or phonetic, are available. The database consists of 250 male and 250 female speakers representing the clients and the impostors of the system. The gender

mismatch is not studied, so that all experiences are strictly gender-dependent. Gender-dependent results are merged into a unique curve, for sake of simplicity. We used the 2F condition for training (2 min or more) and 30 s of speech for the test duration. To evaluate the robustness of the proposed segmental method, the tests are evaluated separately for matched and mismatched conditions (of the training and testing material). They are denoted respectively as same number (SN) and different microphone type (DT). An independent set of 100 female and 100 male speakers with mixed carbon and electret microphones was selected from the NIST1997 database for modeling the world speakers. For the pseudo-impostors (necessary for the Z-normalization), another independent set of 50 female and 50 male speakers with mixed carbon and electret microphones was selected from the NIST1997 database.

LPCC parameters are used for the feature extraction. A 30-ms Hamming window is applied every 10 ms in order to extract 12 LPCC coefficients. A liftering procedure (for more details see [25]) is applied to the cepstral vectors followed by cepstral mean subtraction in order to reduce the effects of the channel. The delta and delta-log-energy features (used only for the GMM modeling) are approximated as described in [25] by a first-order polynomial fit over a finite-length window of five frames centered on the current vector. The MLP parameters, number of hidden units, and input size, although not fully optimized, were experimentally tuned to reach acceptable performances for the different systems. MLPs with one input layer, one hidden layer, and one output layer of neurons are used. Hidden and output layers are computational layers with a sigmoid as activation function. During training, target vectors $t(x)$ are set to $[1, 0]$ and $[0, 1]$ when the input vector x is produced by, respectively, the client and the world speaker. The acoustic vectors are presented randomly from the available training set. The weight matrices of the MLP are iteratively updated, via a stochastic gradient descent procedure, in order to minimize the mean squared error criterion; i.e., weights are updated after every input presentation during the training process. The correction of the matrices values is weighted by a *learning rate* value which is updated after a presentation of the whole training set. A cross-validation set is used to modify the value of the learning rate in order to avoid overtraining. The MLPs used for the global systems have three layers with 120 neurons in the hidden layer and two output units. For the segmental MLPs, the number of neurons in the hidden layer is reduced to 20. In both cases, the input size of the MLP is defined by the number of contiguous frames set as input of the MLP. We use the notation Cxy to denote inputs with x frames to the left and y frames to the right of the central frame. No delta features are inserted at the input of the MLP.[2] In the case of the segmental systems the number of input frames is smaller because of the reduced quantity of training data available.

[2] The propagation function of the MLP is, after training, automatically extracting relevant dynamic information. Preprocessing this information through delta features is, in theory, not needed.

3.3. MLP-Based Results

We first compared the performance of the global MLP-systems with that of equivalent state-of-the-art global GMM systems. For the global systems a gender-dependent background model is used, followed by a Z-normalization of the likelihood ratio. The results in Fig. 2 show that global GMMs perform better than global MLPs. This difference might come from the fact that for the MLP, one part of the training data (roughly 10%) is kept apart for a cross-validation procedure to avoid overtraining. Although the global GMMs perform better, the discriminant training procedure used with the MLPs makes them better candidates for the segmental systems.

For the segmental systems based on MLPs, we first segmented the speech material into eight classes (as described in Section 2.1). A gender- and speech-class-dependent background model is used and the Z-normalization of the likelihood ratio is applied to each class separately, with prior segmentation of the pseudo-impostor speech material into eight speech classes.

Performances for the segmental system on a per-class basis are depicted in Fig. 3 for the same number condition. Only five of eight classes are illustrated for the sake of clarity. These five classes were chosen in such a way that they were between the minimum and maximum results, and were not too overlapping. The DET curves clearly show that classes perform differently and convey more or less information about the speakers.

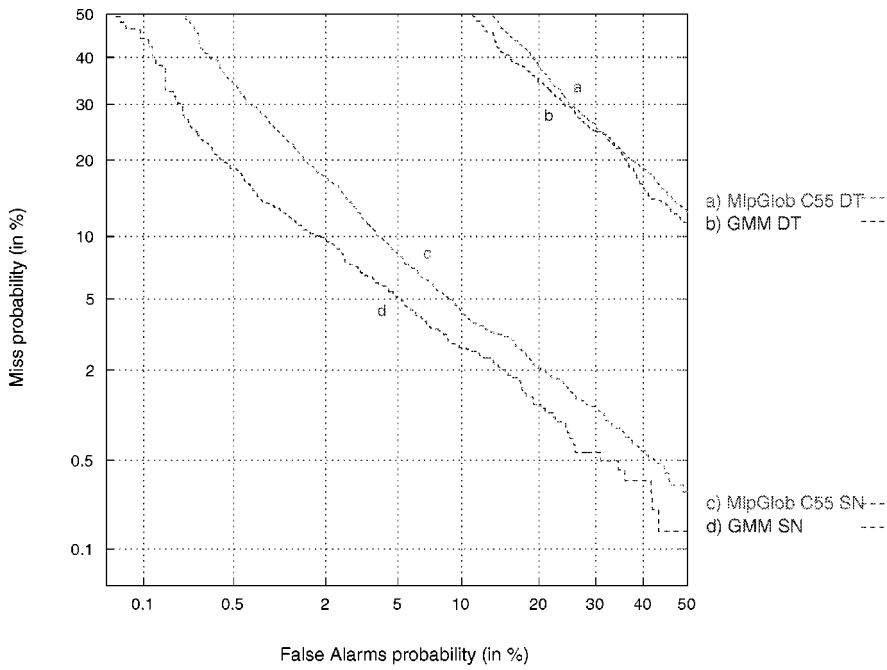

FIG. 2. DET curves for global GMM and global MLP systems, showing the performances for matched train/test conditions (SN) and mismatched train/test conditions (DT). Data from NIST1998, training conditions 2F (2 min or more), and 30 s for test segment duration.

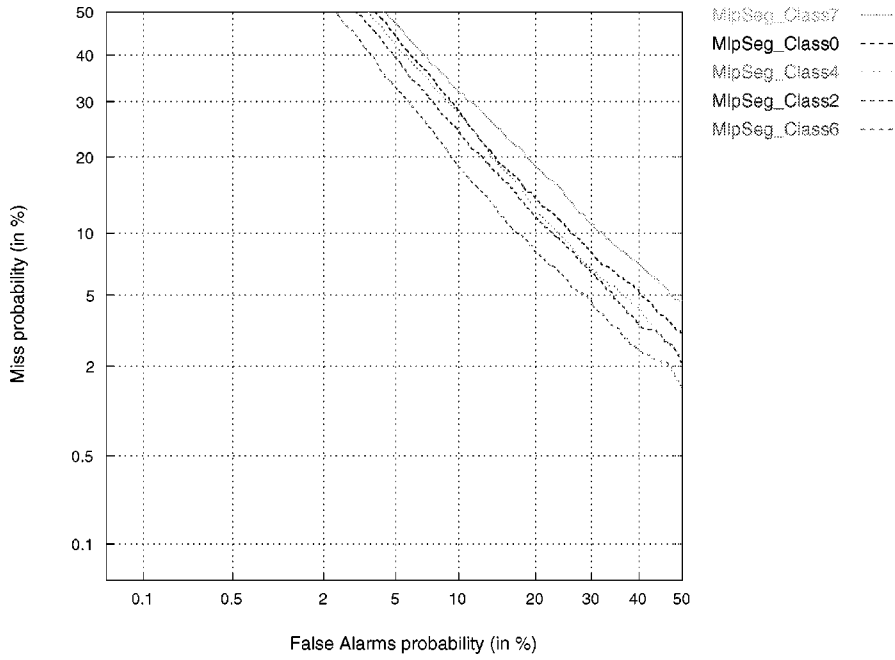

FIG. 3. DET curves for segmental MLP systems, showing the performances of five of eight classes, for matched train/test conditions (SN). Data from NIST1998, training conditions 2F (2 min or more), and 30 s for test segment duration.

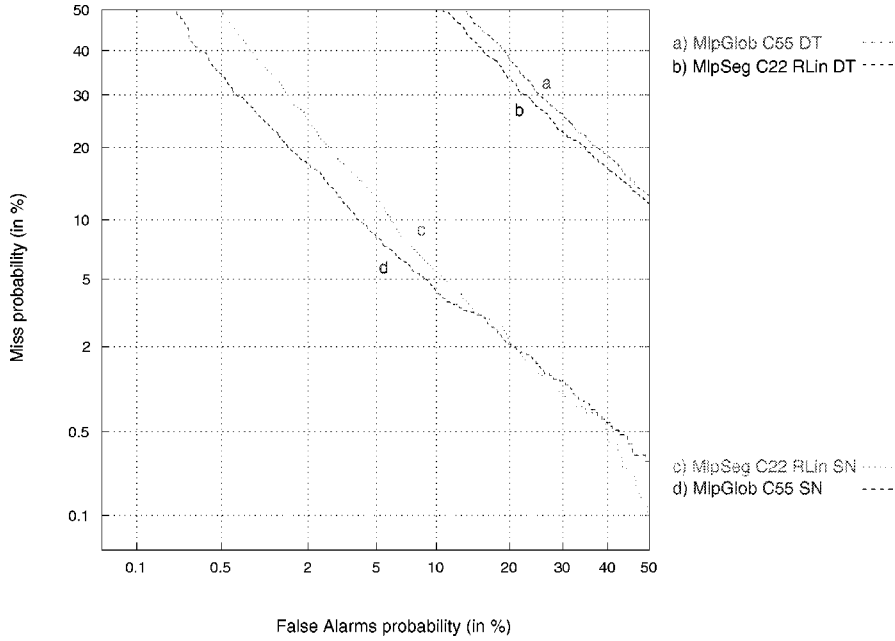

FIG. 4. DET curves for global and segmental MLP systems. MLPGlobC55 stands for the global system with 11 input frames and MLPSegC22RLin indicates the segmental system with linear score recombination using five input frames. Performances are reported for matched train/test conditions (SN) and mismatched train/test conditions (DT). Data from NIST1998, training conditions 2F (2 min or more), and 30 s for test segment duration.

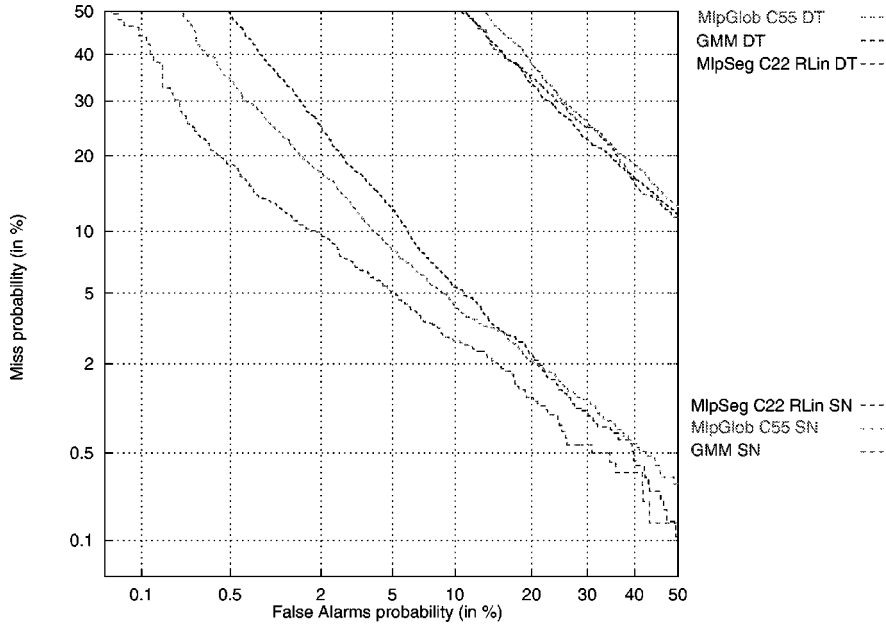

FIG. 5. DET curves for global GMM (as baseline comparison), with global and segmental MLP systems. MLPGlobC55 stands for the global system with 11 input frames and MLPSegC22RLin indicates the segmental systems (using five input frames for the MLP), with linear score recombination. Performances are reported for matched train/test conditions (SN) and mismatched train/test conditions (DT). Data from NIST1998, training conditions 2F (2 min or more), and 30 s for test segment duration.

For the segmental systems, a score recombination of the independent classifiers is necessary to obtain the final decision for the speaker verification, after the Z-normalization has been performed for each class. The differences in class performances, as reported in Fig. 3, suggest that the recombination of scores obtained for each category should be nonlinear, giving more weight to the most discriminant classes. For the sake of simplicity, a linear recombination of scores was used to obtain the final segmental scores. Performances of this approach are reported in Fig. 4, where we compare the global MLP system (denoted as MLPGlob C55) and the segmental MLP system (denoted as MLP SegC22 RLin).[3] Although a simple linear recombination of the individual scores has been used, the MLP segmental system has a performance almost equivalent to that of the global system in the case of matched conditions and performs better than the MLP global system in the case of mismatched conditions. Comparison with the baseline GMM system (Fig. 5) shows that the segmental system has comparable performance in the more difficult mismatched conditions and performs worse than the global GMM and MLP in the matched conditions.

[3] The number of input frames in the segmental systems is diminished to 5 because of the reduced amount of training data available for the segmental speaker models, due to the segmentation procedure.

4. SEGMENTAL GMM MODELING

GMM-based systems have proven their capabilities for speaker modeling. This led us to investigate their capabilities for segmental speaker modeling.

4.1. GMM-Based System Description

In this approach, the client as well as the world pdfs are modeled by the mixture of Gaussian distributions. The likelihood $\mathcal{L}$ of a model is expressed as a weighted sum of Gaussian mixtures, as a function of their means μ and variances Σ:

$$\mathcal{L}(x_n|M) = \sum w_j \mathcal{N}(x_n; \mu_j, \Sigma_j). \tag{9}$$

If L is the number of speech classes, L GMMs must be trained for each client. With the GMM modeling, a segmental world (L world GMMs) or a global world model can be chosen. The latter approach was tested in our experiments. In the testing phase, two possibilities exist for scoring:

– Each frame is first assigned to a speech class (using temporal decomposition and VQ) and only the corresponding GMM is used for the scoring. This approach can be denoted "hard" and was used for the MLP-based segmental speaker modeling.

– A "soft" weighting function of each vector to all speech classes is evaluated; the vector is scored by *all* GMMs and their outputs are weighted by the weighting function κ, which takes into account the distances of the test frames to all the speech classes.

Because we suspect that the segmentation is not perfect, we have used the latter approach with the heuristic weighting function $\kappa^l(x_n)$, which takes into account all the distances of the test vectors x_n to all the speech classes l,

$$\kappa^l(x_n) = \exp\left[\left(1 - \frac{d^l(x_n)}{\sum_{l=1}^{L} d^l(x_n)}\right)^4\right], \tag{10}$$

where $d^l(x_n)$ is the Euclidean distance of x_n from the centroid of the class l and $\sum_{l=1}^{L} d^l(x_n)$ is the sum of the distances over all classes. Weights $\kappa^l(x_n)$ are normalized to sum to 1. The global score S of the test file (composed of N frames) is then computed as

$$S = \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{l=1}^{L}\kappa^l(x_n)\left(\mathcal{L}(x_n|M_{\mathrm{c}}^l) - \mathcal{L}(x_n|M_{\mathrm{w}})\right)\right], \tag{11}$$

where M_{c}^l denotes the client model corresponding to the speech class l and M_{w}, denotes the background world model. They are normalized using a gender-dependent Z-normalization.

This "soft" weighting function of each test feature vector was not tested with the MLP-based segmental systems because of implementation difficulties.

4.2. GMM-Based Experimental Setup

The segmental GMM approach was tested on the ELISA-1 data [13], a subset of the NIST1998 data: 50 client speakers, each with 2 min of training data (condition 2S) and test segment duration 3 s. The parameterization was done using 16 LPCC coefficients with liftering. Each segmental client GMM had 64 mixture components for the LPCC parameters, 64 for delta-LPCC, and 16 for delta-log-energy. Delta features are approximated, as described in [25], by a first-order polynomial fit over a finite-length window of five frames centered on the current vector. The GMM speaker models were initialized by all the data from the given speaker and then each of the eight segmental models was retrained using only the data corresponding to each class. The world model was nonsegmental, gender-dependent, and of the same configuration as the segmental ones. It was trained using 50 electret and 50 carbon background speakers. For the Z-normalization, another set of 50 electret and 50 carbon pseudo-impostor speakers was used.

4.3. GMM-Based Results

First, a global GMM system was developed, with a gender-dependent background (world) model, and Z-normalization was applied to the likelihood ratio. Results for the matched (SN) and mismatched (DT) train/test conditions for the global and segmental systems can be seen in Fig. 6. Then the scoring was per-

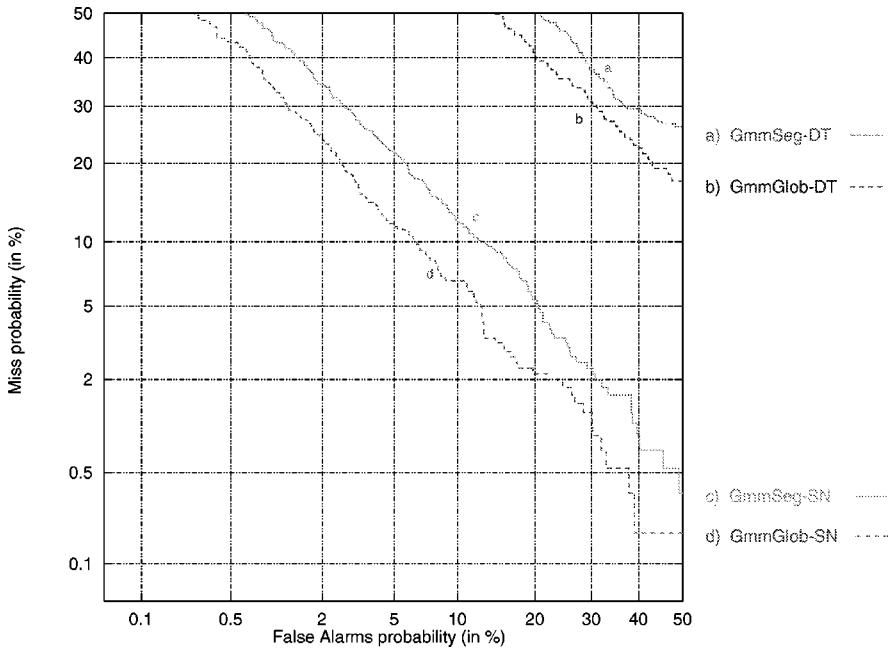

FIG. 6. Results for global and segmental GMM systems. Performances are reported for matched train/test conditions (SN) and mismatched train/test conditions (DT). Data from ELISA-1 (subset of NIST1998); training conditions 2s (2 min), and 3 s for test segment duration.

formed with the above-described procedure, applying the Z-normalization, after the score recombination.

It is obvious that the segmental GMM systems give lower performances than the global GMM systems for both conditions. The most probable reason is the limited amount of training data available for the reestimation of segmental models. An adaptation strategy [17] is a good candidate to give better results. Also, as for MLPs, the merging of class-dependent scores was not completely resolved: the weights depend on proximity of vectors to the centroids of classes but do not reflect the efficiency of classes in discriminating speakers.

5. CONCLUSIONS

In this paper, two methods for segmental text-independent speaker verification, with automatically derived classes of speech sounds, are presented. We have confirmed that the automatically derived ALISP speech segments are not equal in characterizing the speakers. Nevertheless an optimal grouping of acoustic ALISP segments for speaker verification has not been found so far.

Segmental MLP systems have performances similar to those of equivalent global MLP systems. This comparable performance was achieved only when the Z-normalization was done on a per-class basis. Segmental GMM systems have pourer performances than the equivalent global GMM systems.

The better performance of the segmental MLP systems is probably due to the discriminant learning procedures used with the MLPs. In such cases, although the training material is divided among eight classes, the discriminant MLP training procedure probably enables sufficient client modeling per class, which is not the case for the GMM segmental modeling, in the way we implemented it. Another important issue is the smaller amount of testing data for the segmental GMM systems. This test segment duration of 3 s seems to be a limiting factor for the segmental systems, suggesting possible use of the segmental systems only when enough training and testing material is available.

ACKNOWLEDGMENTS

We are grateful to Fréderic Bimbot (IRISA, France) for allowing us to use his temporal decomposition `td95` package.

REFERENCES

1. Atal, B., Efficient coding of LPC parameters by temporal decomposition. In *Proc. IEEE ICASSP 83,* 1983, pp. 81–84.
2. Bennani, Y. and Gallinari, P., Connectionist approaches for automatic speaker recognition. In *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland,* 1994, pp. 95–102.
3. Bourlard, H. and Wellekens, C. J., Links between Markov models and multi-layer perceptrons, *IEEE Trans. Pattem. Anal. Mach. Intell.* **12** (1990), 1167–1178.
4. Bimbot, F., An evaluation of temporal decomposition, Technical Report, Acoustic Research Department, AT&T Bell Labs, 1990.

5. Černocký, J., Baudoin, G., and Chollet, G., The use of ALISP for automatic acoustic–phonetic transcription. In *Proc. SPosSS—ESCA Workshop on Sound Patterns of Spontaneous Speech. Aix-en-Province, France,* September 1998.
6. Černocký, J., Petrovska-Delacrétaz, D., Pigeon, S., Verlinde, P., and Chollet, G., A segmental approach to text-independent speaker verification. In *Proc. Eurospeech, Budapest, Hungary,* September 1999.
7. Chollet, G., Černocký, J., Constantinescu, A., Deligne, S., and Bimbot, F., Towards ALISP: A proposal for automatic language independent speech processing. In *NATO ASI: Computational Models of Speech Pattern Processing* (Keith Ponting, Ed.). Springer-Verlag, Berlin/New York, 1999.
8. Chollet, G., Černocký, J., Gravier, G., Hennebert, J., Petrovska-Delactrétaz, D., and Yvon, F., *Toward Fully Automatic Speech Processing Techniques for Interactive Voice Servers, Speech Processing, Recognition and Artificial Neural Networks.* Springer-Verlag, Berlin/New York, 1999.
9. Corrada-Emmanuel, A., Progress in speaker recognition at dragon systems. In *Proc. ICSLP, Sydney, Australia,* 1998, pp. 1355–1358.
10. Eatock, J. P. and Mason, J. S., A quantitative assessment of the relative speaker discriminant properties of phonemes. In *ICASSP.* Vol. 1, 1994, pp. 133–136.
11. Farrell, K. A., Mammone, R., and Assaleh, K., Speaker recognition using neural networks and conventional classifiers, *IEEE Trans. Speech Audio Process.* **2**, No. 1 (1994), 194–205.
12. Gersho, A. and Gray, R., *Vector Quantization and Signal Compression.* Kluwer Academic, Dordrecht/Norwell, MA, 1992.
13. Gravier, G., The ELISA-1 system description, Technical Report, ENST Paris, February 1999.
14. Haykin, S., *Neural Networks. A Comprehensive Foundation.* Macmillan Co., New York, 1994.
15. Hennebert, J. and Petrovska-Delacréraz, D., Phoneme based text-prompted speaker verification with multi-layer perceptrons. In *RLA2C, Avignon, France,* 1998, pp. 55–58.
16. Hertz, J., Krogh, A., and Palmer, R. G., *Introduction to the Theory of Neural Computation.* Santa Fe Institute Studies in the Sciences of Complexity. Addison–Wesley, Reading, MA, 1991.
17. Leggetter, C. J. and Woodland, P. C., Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models, *Comput. Speech Language*, No. 9 (1995), 171–185.
18. Naik, J. M. and Lubenskt, D., A hybrid HMM-MLP speaker verification algorithm for telephone speech. In *Proc. IEEE ICASSP,* 1994, pp. 153–156.
19. Neuman, M., Gillick, L., Ito, Y., Mc Allaster, D., and Peskin, B., Speaker verification through large vocabulary continuous speech recognition. In *Proc. ICSLP, Philadelphia, PA,* 1996, pp. 2419–2422.
20. Oglesby, J. and Mason, J. S., Optimization of neural models for speaker identification. In *Proc. IEEE ICASSP,* 1990, pp. 261–264.
21. Olsen, J., A two-stage procedure for phone based speaker verification. In *First International Conference on Audio and Video Based Biometric Person Authentication (AVBPA).* Lecture Notes in Computer Science, Vol. 1206. Springer-Verlag, Berlin/New York, 1997, pp. 219–226.
22. Petrovska-Delacrétaz, D. and Hennebert, J., Text-prompted speaker verification experiments with phoneme specific MLPs. In *Proc. IEEE ICASSP, Seattle,* 1998, pp. 777–780.
23. Petrovska-Delacrétaz, D., Černocký, J., Hennebert, J., and Chollet, G., Text-independent speaker verification using automatically labelled acoustic segments. In *Proc. ICLSP, Sydney, Australia,* December 1998.
24. Peskin, B., *et al*., Topic and speaker identification via large vocabulary continuous speech recognition. In *ARPA Workshop on Human Language Technology, Princeton, NJ,* 1993, pp. 119–124.
25. Rabiner, L. and Juang, B. H., *Fundamentals of Speech Recognition.* Prentice–Hall, New York, 1993.
26. Reynolds, D. A., Automatic speaker recognition using gaussian mixture speaker models, *Lincoln Lab. J.* **8**, No. 2 (1995), 173–191.
27. Reynolds, D. A., Comparison of background normalization methods for text-independent speaker verification. In *Proc. Eurospeech,* 1997, pp. 963–966.
28. Richard, M. D. and Lippmann, R. P., Neural network classifiers estimate Bayesian a posteriori probabilities, *Neural Comput.* **3** (1991), 461–483.
29. Weber, F., Peskin, B., Newman, M., Corrada-Emmannuel, A., and Gillick, L., Speaker recognition on single- and multispeaker data, *Digital Signal Process.* **10** (2000), 75–92.