TEXT-INDEPENDENT SPEAKER VERIFICATION USING AUTOMATICALLY LABELLED ACOUSTIC SEGMENTS

Dijana Petrovska Delacrétaz [1] *Jan Černocký* [2,3] *Jean Hennebert* [1] *Gérard Chollet* [4] *

[1] Circuits and Systems Group, Swiss Federal Institute of Technology
[2] Technical University of Brno, Institute of Radioelectronics, Czech Republic
[3] ESIEE, Département Signal et Télécommunications, Paris, France
[4] CNRS URA-820, ENST, TSI Departement, Paris, France

ABSTRACT

Most of text-independent speaker verification techniques are based on modelling the global probability distribution function (pdf) of speakers in the acoustic vector space. Our paper presents an alternative to this approach with a class-dependent verification system using automatically determined segmental units. Segments are found with temporal decomposition and labelled through unsupervised clustering. The core of the system is based on a set of multi-layer perceptrons (MLP) trained to discriminate between client and an independent set of world speakers. Each MLP is dedicated to work with data segments that were previously selected as belonging to a particular class. The last step of the system is a recombination of MLP scores to take the verification decision. Issues and potential advantages of the segmental approach are presented. Performances of global and segmental approaches are reported on the NIST'98 data (250 female and 250 male speakers), showing promising results for the proposed new segmental approach. Comparison with state of the art system, based on Gaussian Mixture Modelling is also included.

1. INTRODUCTION

In recent years, speaker recognition technology has made quite a lot of progress, but open research problems still remain. The generic term of speaker recognition comprises all of the many different tasks of distinguishing people on the basis of their voices. There are speaker identification tasks - who among many candidate speakers pronounced the available test speech sequence; or speaker verification tasks - whether a specific candidate speaker said the available test speech sequence. In this paper we focus on speaker verification, which is actually a decision problem between two classes : the *true* speaker (also denominated as *client* or *target* speaker) and the *other* speakers (usually noted as *impostors* speakers).

As far as the speech mode is concerned, speaker recognition systems can be *text-dependent* or *text-independent*. In text-dependent experiments, the text transription of the speech sequence used to distinguish the speaker is known. In text-independent tasks, the foreknowledge of what the speaker said is not available.

*This work was supported by the Office Federal pour l'Education et la Science (OFES), Switzerland in the framework of the COST 250 European action, by the grant Marie Heimvögetlin of Swiss National Funds for Research and by the Ministry of Education, Youth and Sports of the Czech Republic – project No. VS97060.

As far as the accuracy is concerned, text-dependent systems perform generally better than text-independent systems. There are two reasons for this :

- the knowledge of what has been said can be exploited to align the speech signal into more discriminating classes (words or sub-word speech units);
- an optimized recombination of these class decisions can be done. Several studies on text-dependent systems [6] [13] [14] [9] have demonstrated that some phones show more speaker discriminative power than others [1], suggesting that a weighting of individual class decisions should be performed when computing the global decision.

We are interested here in building robust text-independent systems. Since the content of the speech signal is not accessible, text-independent speaker verification is usually based on modelling the **global** probability distribution function (pdf) of speakers in the acoustic vector space. We believe that such global approaches are reaching their limits because speaker modelling is too coarse in this case.

Therefore, we propose here to investigate a **segmental** approach in which the speech signal is pre-classified into more specific speech units. On the one hand, the segmental approach recovers text-dependent advantages since the speech signal is aligned into classes. On the other hand, the implementation is different since we have no clue about what has been said. As for text-dependent systems, we can underline two potential advantages. First, if the speech units are relevant, then speaker modelling is more precise and the system should present better performances than the global approach. Second, if speech units present different discriminative power, then better recombination of the decisions per class can be done.

The disadvantage of this method is that it requires an accurate recognition of speech segments. Two alternative procedures can be followed :

- Large Vocabulary Continuous Speech Recognition (LVCSR) provides the hypothesised contents of the speech signal on which classic text-dependent techniques can be applied. LVCSR uses previously trained

[1] This fact has been previously validated by phonetic studies [12] that have also shown that the information about the identity of a speaker is not uniformly distributed among speech segments (phones).

phone models and a language model, generally a bigram or trigram stochastic grammar.

- ALISP (Automatic Language Independent Speech Processing) tools [4] provide a general framework for creating sets of acoustically coherent units with little or no supervision.

LVCSR systems, although very promising for segmental approaches, require large annotated data sets for training which are either costly or not available. Furthermore, LVCSR are often dependent on the speech signal characteristics (language, speech quality, ...), making them difficult to port to new tasks. ALISP offers an alternative when no annotated training data is available.

This are the reasons that led us to investigate a text-independent segmental approach based on ALISP tools. Among the available ALISP techniques, we used the temporal decomposition (TD) followed by vector quantization (VQ) to obtain classes of sounds. The speaker verification part is based on multi-layer perceptrons (MLP) trained to discriminate between the client speaker and world speakers (independent from the impostor speakers).

We compare the performances of the segmental speaker verification versus a similar global system on the NIST (National Institute of Standardization in Technology) 1998 corpus [2] including 250 male and 250 female speakers.

2. SYSTEM DESCRIPTION

A classical text-independent speaker verification system, illustrated in Figure 1, can be represented in three blocks :

1. **Feature Analysis** : similarly to what is done for speech recognition, the speech signal is cut into analysis windows undergoing feature extraction. A feature vector is calculated for each window so that the speech signal is transformed into a sequence of feature vectors $X = \{x_1, x_2, \cdots, x_N\}$ of length N. Classical feature extraction algorithms are LPC-cepstral analysis and MFCC analysis. Futher datails can be found in [16] and [15].
2. **Pattern Classification** : The sequence of feature vectors is fed into a classifier that outputs a likelihood score for the client model and the world model, i.e. respectively S_c and S_w.
3. **Thresholding** : the verification (reject/accept) of the speaker is performed comparing the ratio of client and world score against a threshold value which is usually defined to minimize the cost C_{det} of false rejection and false acceptation :

$$\begin{aligned} C_{det} &= C_{fr}P(reject|client)P(client) \\ &\quad + C_{fa}P(accept|\overline{client})P(\overline{client}) \end{aligned} \quad (1)$$

where C_{fr} is the cost of false rejection, C_{fa} is the cost of false acceptance, $P(reject|client)$ is the probability to reject a client, $P(accept|\overline{client})$ is the probability to accept an impostor, $P(client)$ is the a priori probability of client and $P(\overline{client})$ is the a priori probability of impostor. $P(reject|client)$ and $P(accept|\overline{client})$ are both function of the threshold value T. The decision is performed with :

$$\log(S_c) - \log(S_w) > T \quad \rightarrow accept \quad (2)$$
$$\log(S_c) - \log(S_w) \leq T \quad \rightarrow reject \quad (3)$$

[2] NIST organizes every year an evaluation of speaker verification systems. A unique data set and evaluation protocol are provided to each participating laboratory, so that intra- and inter-laboratory algorithms comparisons are significantly easier.

2.1. Global systems

Figure 1 shows a classical way to do pattern classification in text-independent systems, referred here as the **global** method. Assuming independance of successive acoustic vectors, a unique pdf is assigned to the whole vector sequence.

GMM

One way to build the probability distribution functions is to do Gaussian Mixture Modelling (GMM) in which the multivariate distribution is modelized with a weighted sum of gaussians :

$$P(x_n|M) = \sum_{j=1}^{J} w_j \mathcal{N}(\S_\backslash, \mu_|, \Sigma_|) \quad (4)$$

with the constraint $\sum_{j=1}^{J} w_j = 1$. As illustrated in Figure 1, this can be viewed as a particular case of Hidden Markov Model (HMM) where there would be a unique state in the automaton. The HMM emission probability is then equal to the output of the unique probability density function tied to the state. Two models M are built in, one for the client, one for the world. Scores are equal to the product of frame likelihoods over the whole sequence :

$$S_c = p(X|M_{client}) = \prod_{n=1}^{N} p(x_n|M_{client}) \quad (5)$$

$$S_w = p(X|M_{world}) = \prod_{n=1}^{N} p(x_n|M_{world}) \quad (6)$$

MLP

Another possibility to modelize speakers is to build discriminant models with Artificial Neural Nets (for example multi-layer perceptrons MLP). The MLP (see figure 2) is an artificial neural network composed of hierarchical layers of neurons arranged so that information flows from the input layer to the output layer of the network, i.e. no feedback connections are allowed. A good introductory book on artificial neural networks and MLP architectures is [10]. The main advantages of MLP against other systems, include discriminant capabilities, weaker hypotheses on the acoustic vector distributions and possibility to include easily a larger acousic frame window as an input to the classifier. The main drawback using MLP's is that their optimal architecture must be selected by trials and errors.

MLPs, one per client speaker, are discriminatively trained to distinguish between the client speaker and a background world model. MLPs with two outputs are generally used, one for the client and the other for the world class. In [3] it has been proved that if each output unit k

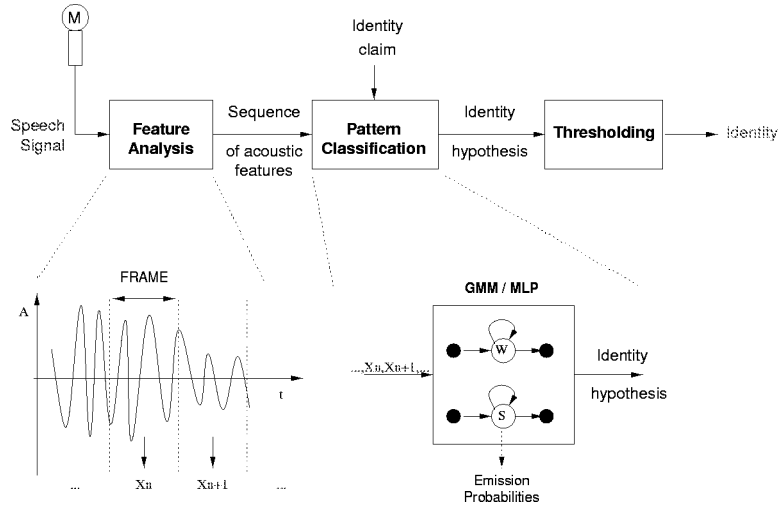

Figure 1: Speaker verification system.

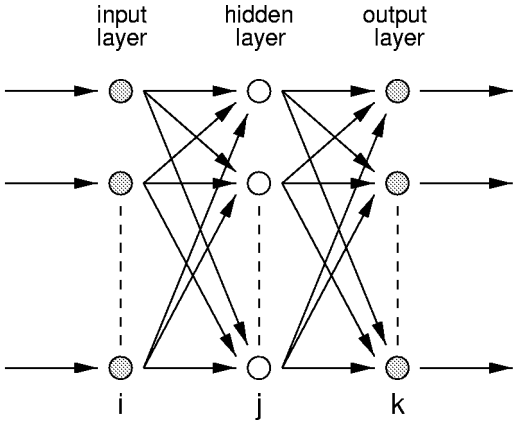

Figure 2: Muti-layer perceptron with 3 layers. The role of the input layer is to distribute the input features. No computation is actually performed at this stage. Hidden and output layers are the computational componenets with non-linear activation functions (generally sigmoids or tanh).

of the MLP is associated to class categories C_k, it is possible to train the MLP to generate a posteriori probabilities $p(C_k|x_n)$.

As explained in [17], the parameters of the MLP (weight matrices) are iteratively updated via a gradient descent procedure in order to minimise the difference between actual outputs and desired targets. In our case, during training, target vectors $d(x_n)$ are set to $[1, 0]$ and $[0, 1]$ when the input vector x_n is produced respectively by the client and by the world speaker. The training is said to be *discriminative* because it minimizes the likelihood of incorrect models (through the zeros of the target vector) and maximizes the likelihood of the correct model. The network attempts to model class boundaries, rather than accurate probability density functions for each class.

As shown with Bayes'rule :

$$P(C_k|x_n) = \frac{p(x_n|C_k)P(C_k)}{p(x_n)} \tag{7}$$

the MLP training incorporates priors $P(C_k)$ when modelizing posterior estimates. These priors are not equal to real-life class prior probabilities but are dependent on the number of samples in the training set :

$$P(C_k) = \frac{N_k^{ts}}{N^{ts}} \tag{8}$$

where N^{ts} is the total number of training patterns and N_k^{ts} is the number of training patterns of the class k. Posteriors have then to be scaled to remove the dependency to the number of patterns in the training set[3]. Posteriors $P(C_k|x_n)$ are then usually divided by priors $P(C_k)$ to obtain the so-called *scaled likelihoods* $p(x_n|C_k)/p(x)$. Scaled likelihoods can be used in place of likelihoods at recognition time since the factor $p(x)$ is not dependent on the class.

Client and world scores are then computed according to the Eqs. 5 and 6 where likelihoods $p(x_n|client)$ and $p(x_n|world)$ are replaced with scaled likelihoods.

2.2. Segmental system

Our aim is to develop a segmental text-independent speaker modelling system (see Figure 3) where the speech sequence is segmented and labelled into a category. The catagories are automatically determined by vector quantization. Segments are assigned to a model which is selected following its class label. Finally, a score recombination of the individual models is performed.

The temporal decomposition (TD) technique is used to compute speech segments and the vector quantization (VQ) to classify them into categories. The modelling part of the system can be similar to the one of the global systems. In our case, we used a set of multi-layer perceptrons (MLP) trained to discriminate between client and world speaker. Each MLP is dedicated to work with data segments that were previously selected as belonging to a particular class.

[3]This is even more necessary in the case of speaker verification where the number of patterns in the client class is extremely low in comparison to the number of patterns in the world class.

Segmentation
Segmentation is achieved using Temporal Decomposition. The purpose is to find quasi-stationary parts in parametric representations. This method, introduced by Atal [1] and refined by Bimbot [2], approximates the trajectory of i^{th} parameter x_n^i by a sum of m *targets* a_{ik} weighted by *interpolation functions* (IF):

$$\hat{x}_n^i = \sum_{k=1}^{m} a_{ik}\phi_k(n), \quad i = 1, \ldots, P, \tag{9}$$

where P is the dimension of the parameter vectors. Equation 9 can be written in matrix notation:

$$\underset{(P \times N)}{\hat{\boldsymbol{X}}} = \underset{(P \times m)}{\boldsymbol{A}} \; \underset{(m \times N)}{\boldsymbol{\Phi}}, \tag{10}$$

where the lower line indicates matrix dimensions. The initial interpolation functions are found using local Singular Value Decomposition with adaptive windowing, followed by post-processing (smoothing, decorrelation and normalization). Target vectors are then computed by: $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Phi}^{\#}$, where $\boldsymbol{\Phi}^{\#}$ denotes the pseudo-inverse of IFs matrix. IFs and targets are locally refined in iterations minimizing the distance of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$. More details on the computation of $\boldsymbol{A}$ and $\boldsymbol{\Phi}$ can be found in [2].

Intersections of interpolation functions permit to define speech segments $X_a^b = \{x_a, \ldots, x_b\}$ and the utterance is decomposed into I non-overlapping segments :

$$X = \{X_{s_1}^{s_2-1}, X_{s_2}^{s_3-1}, \ldots, X_{s_I}^{N}\} \tag{11}$$

with $s_1 = 1$ and $s_1 \leq s_2 \leq \ldots \leq s_I \leq N$.

Labelling
The next step is *unsupervised clustering*. Among several available algorithms (Ergodic HMM, self-organizing map, etc.), *Vector Quantization* (VQ) was chosen for its simplicity. The VQ codebook $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L\}$ is trained by K-means algorithm with binary splitting [8]. Training is performed using vectors positioned in gravity centers of the temporal decomposition interpolation functions, while the *quantization* takes into account entire segments using cumulated distances between all vectors of a segment and a code-vector as follows :

$$y_b^s = \min_i \left[\sum_{x \in X_s} d(x, y_i) \right] \tag{12}$$

where y_b denotes the winner centroid and s a particular segment. Temporal decomposition and vector quantization provide a symbolic transcription of the data in an unsupervised way. Each vector of the acoustic sequence is declared as a member of a category C_l determined through the segmentation and the labelling. The number of categories is fixed by the number of centroids in the VQ codebook.

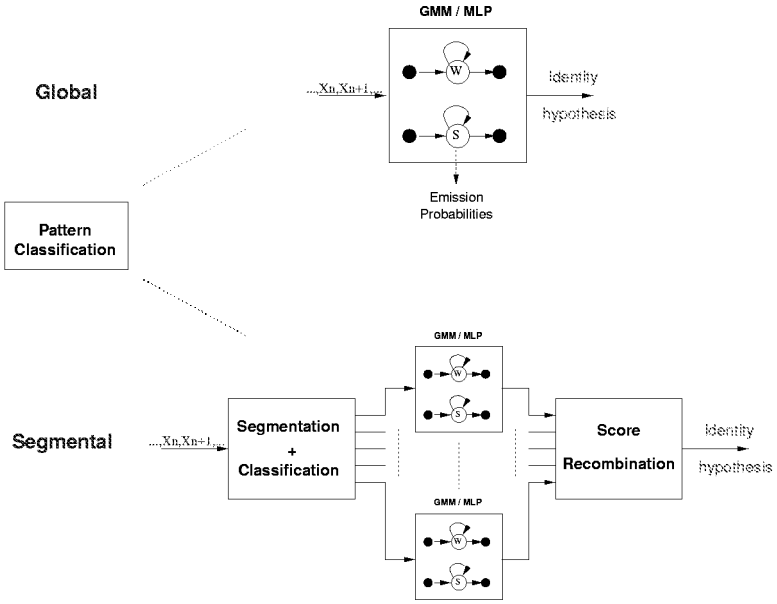

Figure 3: Global and segmental speaker verification system.

MLP Modelling
The same technique as for global modelling (see section **??**) is applied, but this time, L MLP's (same number as the number of centroids in the codebook) are used. They are respectively fed with feature vectors having corresponding labels. For example, the MLP associated with category C_l provides a segmental score as follows :

$$S_{cl} = \prod_{x \in C_l} p(M_{cl}|x)/P(M_{cl}) \tag{13}$$

$$S_{wl} = \prod_{x \in C_l} p(M_{wl}|x)/P(M_{wl}) \tag{14}$$

where products involve vectors beeing previously labelled as members of category C_l. Subscripts cl and wl denote respectively the client model for segmental category C_l and world model for segmental category C_l. Posterior probabilities in Eqs. 13 and 14 are also divided by prior probabilities, computed similarly as in Eq. 8.

Score recombination
Among several techniques for score recombination a **Linear recombination** is investigated in this study. The MLP scores are recombined linearly through a simple addition :

$$S_c = \sum_{l=1}^{L} S_{cl} \tag{15}$$

$$S_w = \sum_{l=1}^{L} S_{wl} \tag{16}$$

$$\tag{17}$$

3. EXPERIMENTS

3.1. Task description

The speaker verification system performances depend on many factors. Among the most influencing ones are the amount of training, the duration of the test segments, the microphone difference between training and testing data, the noise and the temporal drift, as pointed out by Doddington [5]. In the framework of evaluating the possible improvements brought by using a new method, we vary a choosen set among all of these factors.

Both segmental and global systems were compared in the framework of the NIST-'98 evaluation campaign. The data are selected from the SWITCHBOARD database, recorded over telephone lines. The speech is spontaneous and no transcriptions, neither orthographic nor phonetic, are available. The *training set* consists of 250 male and 250 female subjects representing *clients* of the system (the sex mismatch is not studied in these evaluations, so that all experiences are strictly sex-dependent). For each client, the system is trained under three training conditions (depending on the amount of data and denoted 1S for one session, 2S for more sessions and 2F for two sessions -full). The *test set* comprises 2500 test files per sex, and per test condition, each "pretending" to be 10 clients. Three test conditions are defined depending on the available amount of test data: 3, 10 or 30 seconds. The total number of trials to do within NIST evaluations is: 2 sexes × 9 train–test conditions × 2500 test files × 10 trials per test file = 450000. For this study only one training and testing configuration is considered: two minutes or more for the training and 30 s of speech for the tests. To evaluate the robustness of the new proposed segmental method, the tests are evaluated separately for matched and mismatched conditions, noted respectively as SN (same number) and DT (different microphone types). For modelling the world speakers, an independent set of 100 female and 100 male speakers with mixed carbon and electret microphones, was selected in the 1997 database of the previous evaluation.

3.2. Experimental setup

LPC-cepstral parameters are used for the feature extraction. A 30 ms Hamming window is applied to the speech every 10 ms in order to extract 12 LPC-cepstrum coefficients. The order of the LPC analysis is set to 10. A liftering procedure is applied to the cepstral vectors followed by cepstral mean substraction in order to operate a blind deconvolution.

The structure of the MLP's used for the global systems is a three layer MLP, with 120 neurons in the hidden layer. For the segmental MLP's, the number of neurons in the hidden layer is reduced to 20.

For the segmental system, the temporal decomposition was set up to detect 15 events per second in average. The vector quantization was trained on 1997 data with codebook size of $L = 8$. Coherence of acoustic labelling among speakers is verified in informal listening tests.

3.3. Length of the input speech sequence

The MLP offer the possibility to vary the length of the input speech. The feature extraction techniques we applied, use windows of 30 ms of speech for the calculations of the feature vectors. In this case when one frame is taken, it corresponds to 30 ms speech units. Taking more contiguous frames enables us to broaden this value. In the case of eleven conecutive frames, the input frame sequence is equivalent to 130 ms of speech.

ROC and DET curves
In Eq. 1, $P(reject|client)$ and $P(accept|\overline{client})$ are both function of the threshold value T. Given a test set, false acceptance (FA) and false rejection (FR) curves can then be plot giving different values to T. FA and FR curves are often represented as receiver operating curves (ROC)[7]. Generally, FA rate is plotted on the horizontal axis and and FR rate is plotted on the vertical. ROCs are sometimes not practical when similar systems need to be compared and an alternative representation of the ROC is used: Detection Error Tradeoff (DET)[11] in which the x and y scales are in the log domain. If the likelihood ratios are normally distributed (this is often observed in practice), the DET curve will be close to a straight line,enabling easy observation of system contrasts. Example of DET curves are given hereafter as, for example, in Fig. ??.

4. RESULTS

The experimental results are described as follows. First the global MLP performances are compared with the state of the art GMM based system. The influence of the mismatched training and testing conditions is pointed out. In the next section the influence of the length of the acoustic window is discussed. These experiments provide the necessary comparison points for the segmental system results described aftewards, where the performances per class are detailed. Finally, results with a simple recombination technique are given.

4.1. Global system results

Actually better results for speaker verification are achieved with GMMs. We use them here as the state of the art comparison point. The comparisons of the performances of the global MLP and GMM systems are shown in Figure 4. The importance of the mismatched training and testing conditions, as far as the microphone differences are considered, are also visible on this figure. When the test segments come from a different handset type then the training speech material (DT curves), the error rates are increased roughly by a factor of four. Global GMM and MLP have comparable results. Taking into account the further discriminant possibilities we can use with MLP, we adopted them for the segmental experiments.

4.2. Influence of the input window length

Our previous studies [14], [9] showed the importance of the acoustical window length used as the input of the MLP for speaker verification experiments. Figure 5 demonstrates the behaviour of the MLP with different input window length. The number of input frames spans from one (noted as C00, corresponding to 30 ms) to 11 input frames (noted as C55, equivalent to 130 ms). Using more contiguous input frames improves the performances of the global MLP systems, however a saturation is appearing when eleven frames are used as input.

4.3. Segmental system results

Results concerning the performances of five among eight classes are depicted in Figure 6. Only the ones that have dissimilar performances are illustrated. They demonstrate the fact that the classes perform differently and convey more or less informations about the speakers. In a similar manner as for the global system, one important issue is the number of input frames of the MLP's. For these preliminary tests we set the number of contiguous trames for the segmental trainings to five frames.

Another important factor is the amount of training material available per class. It is well known that more training material we have, better the models are. The number of classes is a compromise between the number of classes and the training material available. In the case we want the automatically determined speech units to corresond to phone-like units, the number of classes should be approximately equal to the number of phonemes. But with 2 minutes of speech training material we are not sure that enough training material per class is available. For this reason, the number of classes was set to eight, so that broad phonetic classes are detected.

When using fusion techniques to recombine the scores of all these classes, one should use the information that certain classes perform better than others.

4.4. Segmental recombination results

In our perspective to investigate the potentialities of the segmental approach, as a first step, we must ensure that doing the automatic segmentation does not degrade the performances. Figure 7 includes the results of the linear recombination of the scores of the eight classes (noted as MlpSeg C22 RLin) and the best global MLP system . With this simple recombination technique we observe comparable but a little worse results as with the global MLP system for the same number conditions. This ensures us that the segmentation is consistent. As far as the more difficult

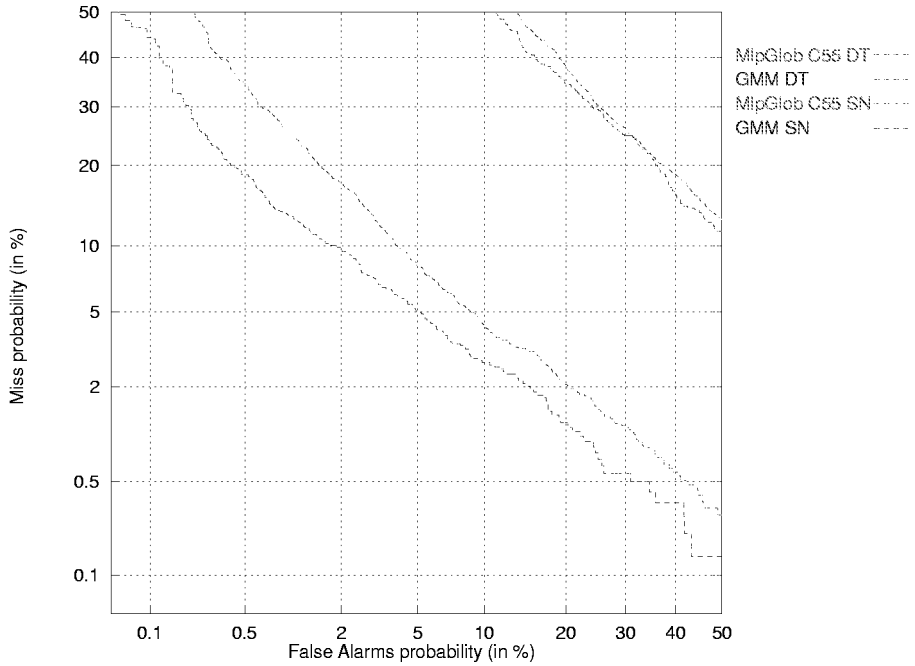

Figure 4: Global systems, GMM and MLP modelling, training condition 2 min or more, test duration 30 sec, same number (SN) and different type (DT), for train and test materials.

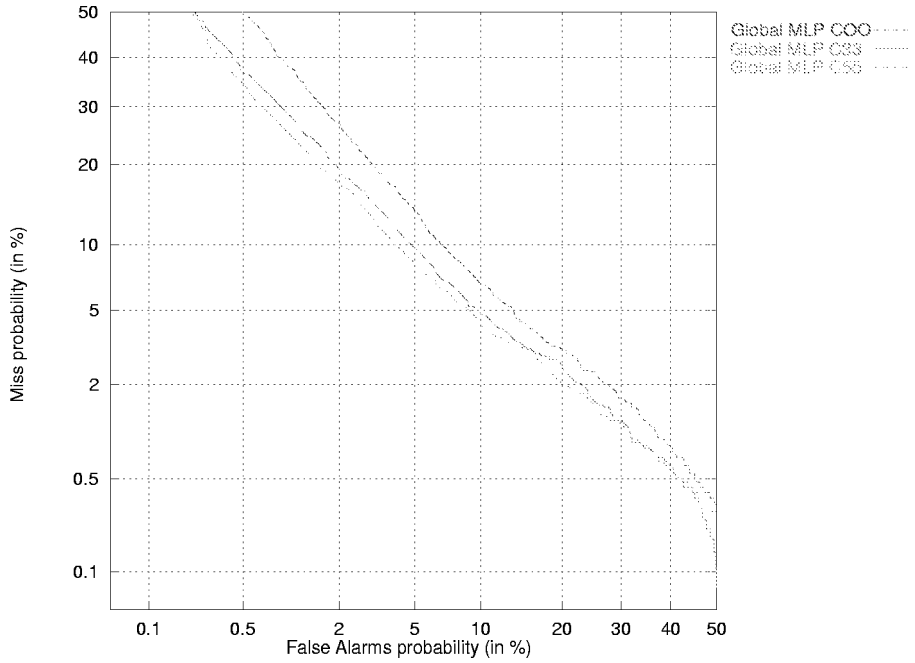

Figure 5: Global system, MLP modelling, training condition 2F (2 min or more), test duration 30 sec, same number, influence of the input training window length (varying fron C00=30ms to C55=130ms.

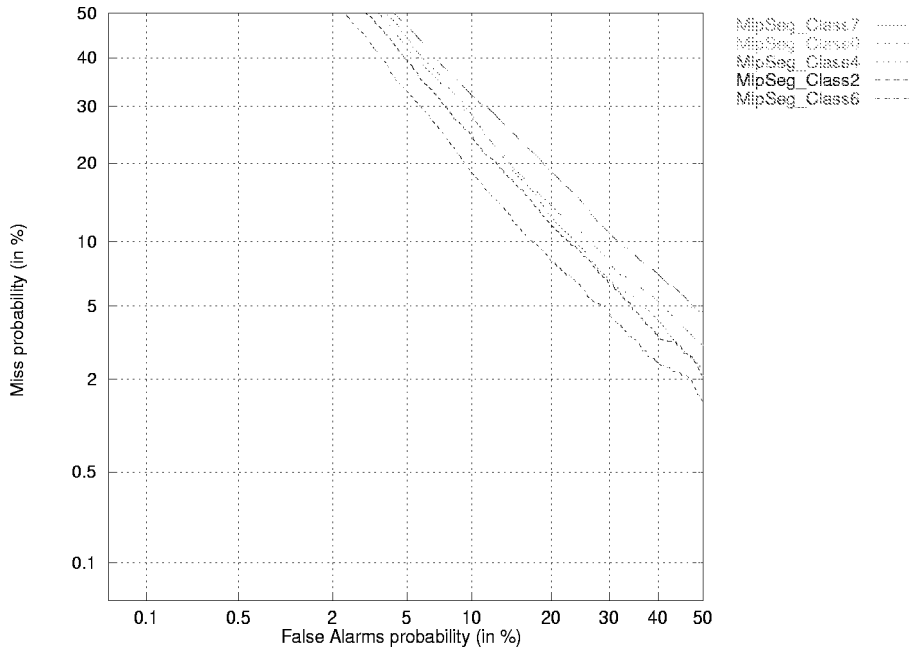

Figure 6: Segmental system, results by classes, training condition 2 min or more, test duration 30 sec, same train and test number.

experimental conditions (DT) are concerned the results of the segmental system are better than the global one. This opens the way to the fusion techniques, where individual tuning of parameters corresponding to each class can be done.

5. CONCLUSIONS AND FUTURE WORK

In this work, use of automatically derived speech units in text-independent speaker verification experiments was investigated. The automatic segmentation performed by temporal decomposition and vector quantization was coupled with artificial neural network scoring. The experimental results were obtained on NIST 98 test data. We verified, that the segmental system reaches similar performances as the global one, and even outperforms it in mismatched training/test conditions. In comparison with the baseline GMM system, GMMs showed still superior performance.

The main open issue of this problematic is the merging of class-dependent results to obtain the global score, taking into account the discriminant performances of classes. Some computations, done standardly at the end of the speaker verification chain (normalization, threshold setting) must be moved towards class-dependent score computations. More general issue of this work is the determination of classes, currently done independently to the following steps. The classes should be ideally determined as optimal for the given scoring system in an iterative refinement.

Besides speech recognition and very low bit-rate coding, we demonstrated that the ALISP techniques are potentially useful also in speaker verification as they limit the human interaction necessary (and hence the number of errors introduced by humans, and the cost) and they approach the system to the data rather than to units more or less related with the text (phonemes). However, a lot of work remains to be done when applying these methods efficiently in practice.

6. REFERENCES

1. B. S. Atal. Efficient coding of LPC parameters by temporal decomposition. In *Proc. IEEE ICASSP 83*, pages 81–84, 1983.
2. F. Bimbot. An evaluation of temporal decomposition. Technical report, Acoustic research departement AT&T Bell Labs, 1990.
3. Hervé Bourlard and C. J. Wellekens. Links between markov models and multi-layer perceptrons. *IEEE Trans. Patt. Anal. Machine Intell.*, 12(12):1167–1178, December 1990.
4. G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. *NATO ASI: Computational models of speech pattern processing*, chapter Towards ALISP: a proposal for Automatic Language Independent Speech Processing. Springer Verlag, in press.
5. George Doddington. Speaker recognition evaluation methodology - an overview and perspective -. In *Speaker Recognition and its Commercial and Forensic Apllications (RLA2C)*, pages 60–66, Avignon, France, 1998.
6. J. P. Eatock and J. S. Mason. A quantitative assessment of the relative speaker discriminant properties of phonemes. In *ICASSP*, volume 1, pages 133–136, 1994.
7. James Egan. *Signal detection theory and ROC analysis.* Academic Press, 1975.

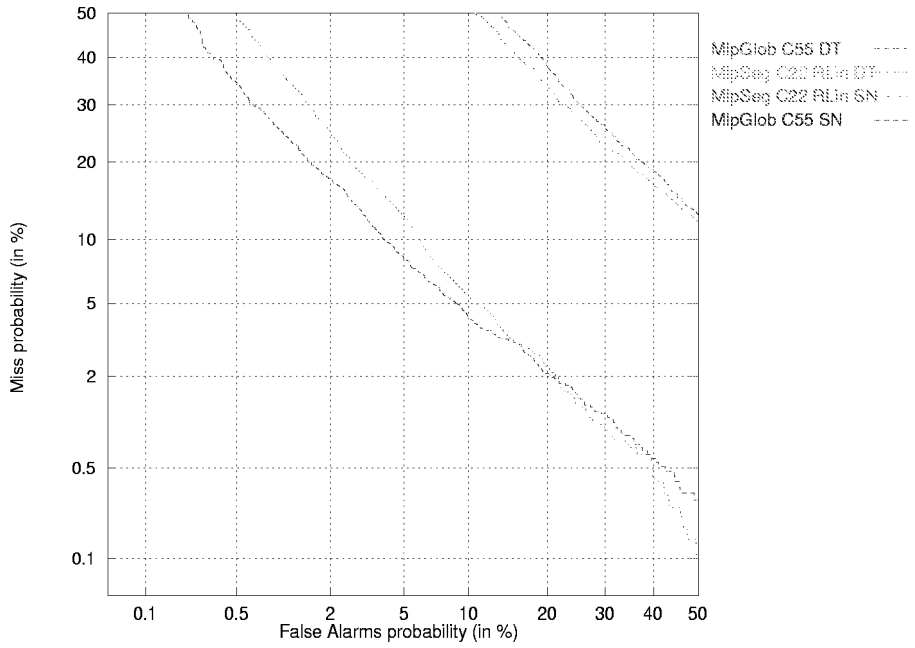

Figure 7: Global and segmental system, training condition 2F, test duration 30 sec, same number (SN) and different type (DT).

8. Allen Gersho and Robert Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.
9. J. Hennebert and D. Petrovska. Phoneme based text-prompted speaker verification with multi-layer perceptrons. In *RLA2C 98*, pages 55–58, Avignon, France, 1998.
10. John Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the theory of Neural Computation.* Santa Fe Institute Studies in the Sciences of Complexity. Addison Wesley, 1991.
11. Alvin Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assesment of detection task performance. In *Eurospeech 1997*, pages 1895–1898, Rhodes, Greece, 1997.
12. F. Nolan. *The Phonetic Bases of Speaker Recognition.* Cambridge University Press, 1983.
13. J. Olsen. A two-stage procedure for phone based speaker verification. In G. Borgefors J. Bigün, G. Chollet, editor, *First International Conference on Audio and Video Based Biometric Person Authentication (AVBPA)*, pages 219–226, Crans, Switzerland, 1997. Springer Verlag: Lecture Notes in computer Science 1206.
14. D. Petrovska and J. Hennebert. Text-prompted speaker verification experiments with phoneme specific mlp's. In *ICASSP*, pages 777–780, Seattle, 1998.
15. Joseph Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1214–1247, September 1993.
16. Lauwrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.
17. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing. Exploration in the Microstructure of Cognition*, volume 1. MIT Press, 1986.