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, Signal and Telecommunications Department, Paris, France
[4] CNRS URA-820, ENST, TSI Department, Paris, France

ABSTRACT

Most of the current text-independent speaker verification techniques are based on modelling the global probability distribution function of speakers in the acoustic vector space. We present an alternative approach based on class-dependent verification systems using automatically determined segmental units, obtained with temporal decomposition and labelled through unsupervised clustering. The core of the system is 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 are previously selected as belonging to a particular class. Issues and potential advantages of the segmental approach are presented. Performances of global and segmental approaches are tested on the NIST'98 database (250 female and 250 male speakers), showing promising results for the proposed new segmental approach. Comparison with a state of the art system, based on Gaussian Mixture Modelling is also included.

1. INTRODUCTION

We are concerned here with speaker verification systems [3]. In *text-dependent* experiments, the text transcription 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. Text-dependent systems perform generally better than text-independent systems, because the knowledge of what is said can be exploited to align the speech signal into more discriminant classes (words or sub-word speech units). Furthermore several studies on text-dependent systems [6] [11] [12] [8] have demonstrated that some phones show more speaker discriminant power than others, suggesting that a weighting of individual class decisions should be performed when computing the global decisions.

We are interested in building robust text-independent systems. They are usually based on modelling the **global** probability distribution function of speakers in the acoustic vector space. In our opinion, such global approaches are reaching their limits because the modelling is too coarse. We propose here to investigate a **segmental** approach in which the speech signal is pre-classified into more specific speech units. Performances of the segmental versus a similar global system are tested on the NIST'1998 corpus [1] including 250 male and 250 female speakers.

*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. We also like to thank Frédéric Bimbot for the temporal decomposition package and Guillaume Gravier for the znorm package.

2. SYSTEM DESCRIPTION

The segmental approach recovers some text-dependent advantages since the speech signal is aligned into classes but the implementation is different since we have no clue about what is said. Two potential advantages can be pointed out : firstly, if the speech units are relevant, then speaker modelling is more precise, thus allowing better performances than the global approach; secondly, 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 accurate recognition of speech segments is required. Two alternatives are possible. The first one is to use Large Vocabulary Continuous Speech Recognition (LVCSR) systems that provide the hypothesised contents of the speech signal on which classic text-dependent techniques can be applied. The second possibility is to use Automatic Language Independent Speech Processing (ALISP) tools [5], that 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 databases for training the phone models, which are either costly or not available and are often dependent on the speech signal characteristics (language, speech quality, etc.). These arguments make them difficult to adapt to new tasks. ALISP offers an alternative when no annotated training data is available. These are the reasons that led us to investigate a text-independent segmental approach based on ALISP tools. Temporal decomposition followed by vector quantisation are used 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.

2.1. Global systems

The classical way to do pattern classification in text-independent systems is to assign a unique probability distribution function (pdf) to the whole vector sequence. One way to build the pdfs is to use Gaussian Mixture Modelling (GMM) in which the multivariate distribution is modelled with a weighted sum of gaussians.

[1] The National Institute of Standards and Technology organises every year an evaluation of speaker verification systems, with a unique data set and evaluation protocol provided to each participant.

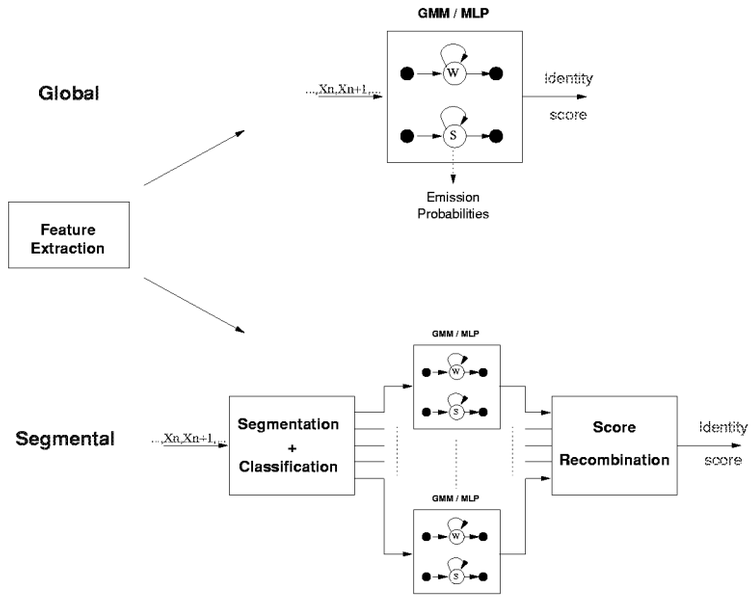

Figure 1: Global and segmental speaker verification systems.

Another way to perform classification is to use Artificial Neural Nets [9]. Multi-layer perceptrons (MLPs) are often used. They include discriminant capabilities and weaker hypotheses on the acoustic vector distributions. The main drawback using MLPs 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. If each output unit k of the MLP is associated to class categories C_k, it is possible to train the MLP to generate a posteriori probabilities $p(C_k|x_n)$ [4]. During the training, the parameters of the MLP are iteratively updated via a gradient descent procedure in order to minimise the difference between actual outputs and desired targets. The training is said to be discriminant because it minimises the likelihood of incorrect models and maximises the likelihood of the correct model. The network attempts to model the class boundaries, rather than the accurate probability density functions for each class.

When using either GMM or MLPs, the sequence of feature vectors is fed into a unique classifier that outputs a score for the client model and the world model, i.e. respectively S_c and S_w (see Figure 1, top part). The verification (reject/accept) of the speaker is performed comparing the ratio of client and world score against a threshold value as follows :

$$\log(S_c) - \log(S_w) > T \qquad \rightarrow accept \qquad (1)$$

$$\log(S_c) - \log(S_w) \leq T \qquad \rightarrow reject \qquad (2)$$

2.2. Segmental system

Our aim is to develop a segmental text-independent speaker modelling system (see Figure 1, bottom part) where the speech sequence is segmented and labelled into categories. Each MLP is dedicated to work with data segments that were previously selected as belonging to a particular class. Segmentation is achieved using temporal decomposition (TD). The purpose of the TD 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). The initial interpolation functions are found using local singular value decomposition with adaptive windowing, followed by post-processing (smoothing, decorrelation and normalisation) [2]. Intersections of interpolation functions permit to define speech segments and the utterance is decomposed into non-overlapping segments. The next step is *unsupervised clustering*. Among several available algorithms (Ergodic HMM, self-organising map, etc.), *Vector Quantisation* (VQ) was chosen for its simplicity. The VQ codebook is trained by K-means algorithm with binary splitting [7]. Training is performed using vectors positioned in gravity centers of the temporal decomposition interpolation functions, while the *quantisation* takes into account entire segments using cumulated distances between all vectors of a segment and a code-vector. TD and VQ 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. In the modelling step, the same technique as for global modelling is used. L MLPs (same number as the number of centroids in the codebook) are trained for each client. 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 :

$$S_{cl} = \prod_{x \in C_l} P(M_{cl}|x)/P(M_{cl}) \qquad (3)$$

$$S_{wl} = \prod_{x \in C_l} P(M_{wl}|x)/P(M_{wl}) \qquad (4)$$

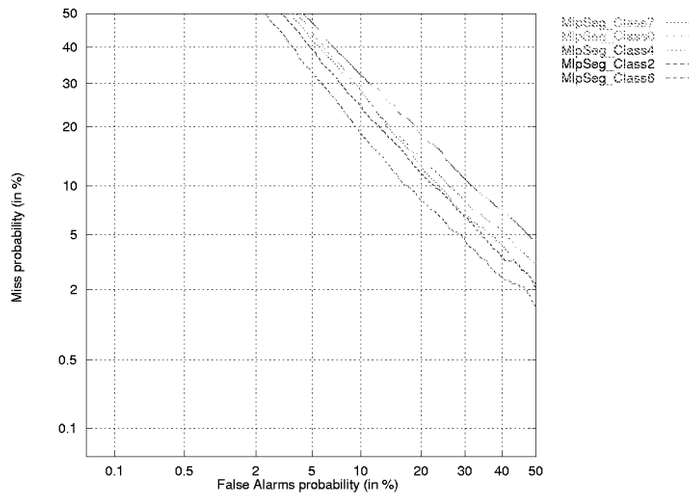

Figure 2: Segmental system, results by classes, training condition 2 min or more, test duration 30 s, same train and test number.

where products involve vectors being 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.

3. EXPERIMENTS

Task description: Segmental and global systems are tested on the NIST'98 database, part of the SWITCHBOARD II database, recorded over telephone lines. The speech is spontaneous and no transcriptions, neither orthographic nor phonetic, are available. The database consists of 250 male and 250 female subjects representing the clients and the impostors of the system. The sex mismatch is not studied, so that all experiences are strictly sex-dependent. Sex-dependent results are merged in a unique curve, for sake of simplicity. Only one training and testing configuration is considered: 2 min or more for the training and 30 s of speech for the test duration. To evaluate the robustness of the new proposed segmental method, some of the tests are evaluated separately for matched and mismatched conditions (of the training and testing material). They are noted respectively as SN (same number) and DT (different microphone type). An independent set of 100 female and 100 male speakers with mixed carbon and electret microphones was selected from the NIST'1997 database for modelling the world speakers.

Experimental setup: LPC-cepstral parameters are used for the feature extraction. A 30 ms Hamming window is applied 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 reduce the effects of the channel. The structure of the MLPs used for the global systems is a three layer MLP, with 11 contiguous frames as input, and with 120 neurons in the hidden layer. For the segmental MLPs, the number of neurons in the hidden layer is reduced to 20 and 5 contiguous frames are used as the input for the MLPs. The temporal decomposition is set to detect 15 events per second in average. The vector quantisation is trained on the 1997 data with codebook size of $L = 8$. Coherence of the acoustic labelling among speakers is verified through informal listening tests. Znorm is applied for each system.

ROC and DET curves: Performances of speaker verification systems are usually given in terms of False Alarms and Miss Probability, often represented as Receiver Operating Curves (ROC). When similar systems need to be compared, it is more practical to use a Detection Error Tradeoff (DET)[10] in which the x and y scales are in the log domain.

4. RESULTS

Performances on a per-class basis for the segmental system (SN conditions) are depicted in Figure 2. Only five classes having dissimilar performances are chosen for illustration. Classes perform differently and convey more or less informations about the speakers. One important factor is the amount of training material available per class. It is well known that the more training material we have, better the models are. In the case when the automatically determined speech units are supposed to correspond to phonemes, the number of classes should approximately equal the number of phonemes. However two minutes of training material might not be sufficient to ensure a proper training of all the classes. This is the reason why the number of classes is set to eight, so that broad phonetic classes are detected.

In Figure 3 we compare global GMM, global MLP and segmental MLP modelling. GMMs are our baseline system that stands as the state of the art comparison point. The importance of the mismatched training and testing conditions, as far as the microphones are concerned, is also shown. When the test segments are issued from different handset type than the training speech material (DT curves), the error rates are increased roughly by a factor of five. This figure includes the results of the linear score recombination of the eight classes (noted as MLP SegC22 RLin) and the best global MLP system. With this simple recombination technique we observe a slight degradation for the SN conditions. For the more difficult DT condition, the segmental system outperforms the

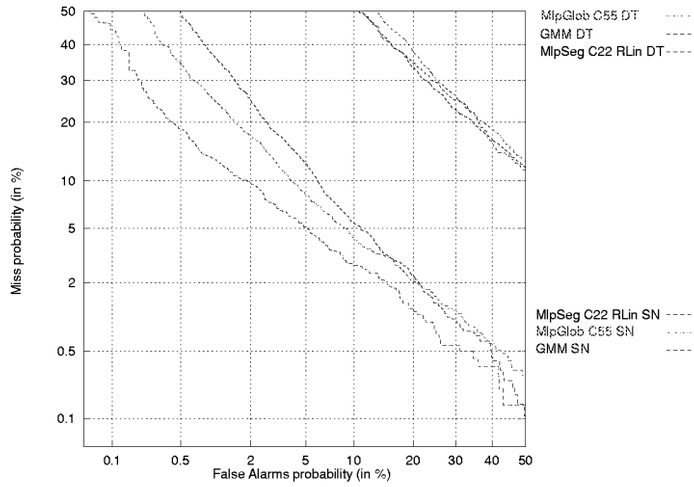

Figure 3: Global GMM and global and segmental MLP systems, training condition 2 min or more, test duration 30 sec, same number (SN) and different type (DT), for train and test materials. RLin denotes a linear recombination of the class scores.

global MLP system, and even the GMM results. This fact opens the way to fusion techniques with potential new improvements of the results. When using fusion techniques to recombine the scores of all the classes, one should furthermore take into account that certain classes convey more speaker informations than others.

5. CONCLUSIONS

In this work, use of automatically derived speech units in text-independent speaker verification experiments is investigated. The automatic segmentation performed by temporal decomposition and vector quantisation is coupled with artificial neural network scoring. The proposed system is tested on NIST'98 database. The proposed segmental system reaches similar performances as the global MLP system, and even outperforms it in mismatched training/test conditions. We show that ALISP techniques are potentially useful also in speaker verification because they are automatic and unsupervised, limiting the human interaction necessary, and hence the number of errors introduced by human operators. Two issues are still open regarding the segmental approach. Firstly, per-class individual tuning of the parameters should be investigated (thresholding, normalisation, etc.). Secondly, better merging of the class-dependent results to obtain the global scores, taking into account the discriminant performances of the classes should be analysed.

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. F. Bimbot and G. Chollet. *EAGLES Handbook on Spoken Language Systems*, chapter Assessment of speaker verification systems. Mouton de Gruyter, 1997.
4. 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.
5. 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.
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. Allen Gersho and Robert Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.
8. J. Hennebert and D. Petrovska. Phoneme based text-prompted speaker verification with multi-layer perceptrons. In *RLA2C 98*, pages 55–58, Avignon, France, 1998.
9. 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.
10. Alvin Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. In *Eurospeech 1997*, pages 1895–1898, Rhodes, Greece, 1997.
11. 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.
12. D. Petrovska and J. Hennebert. Text-prompted speaker verification experiments with phoneme specific mlp's. In *ICASSP*, pages 777–780, Seattle, 1998.