A SEGMENTAL APPROACH TO TEXT-INDEPENDENT SPEAKER VERIFICATION

J. Černocký [1], *D. Petrovska-Delacrétaz* [2], *S. Pigeon* [3], *P. Verlinde* [3,4] *and G. Chollet* [4]

[1]Brno University of Technology, Inst. of Radioelectronics, Czech Republic, cernocky@urel.fee.vutbr.cz
[2]EPFL Lausanne, DE-CIRC, Switzerland, Dijana.Petrovska@epfl.ch
[3]Royal Military Academy Brussels, SIC, Belgium, {verlinde,spigeon}@elec.rma.ac.be
[4]ENST Paris, Département Signal, France, {verlinde,chollet}@tsi.enst.fr

10.21437/Eurospeech.1999-556

ABSTRACT

Current text-independent speaker verification systems are usually based on modeling globally the probability density function (PDF) of the speaker feature vectors. In this paper, segmental approaches to text-independent speaker verification are discussed. Unlike the schemes based on Large Vocabulary Continuous Speech Recognition (LVCSR) with previously trained phone models, our systems are based on units derived in unsupervised manner using the ALISP (Automatic Language Independent Processing) tools. Speaker modeling is then done independently for each class of speech sounds. Among the techniques to merge the class-dependent scores, linear combination was tested and logistic regression and a method based on the Mixture of Experts technique are under investigation. The experimental results were obtained on the data from the NIST-NSA'98 campaign.

Keywords: text-independent speaker verification, segmental approach, data fusion.

1. INTRODUCTION

Current text-independent speaker verification systems are usually based on modeling globally the probability density function (PDF) of the speaker feature vectors. In such systems, the temporal information of the speech sequence is not taken into account and all the phonetic classes are represented using a unique model. It is however possible to divide the speech sounds into categories and perform the modeling independently for each of them. Ideally, this should lead to more precise speaker modeling and to lower error-rate of the system. The categories can be defined using two approaches:

1. The first possibility is to use a Large Vocabulary Continuous Speech Recognition (LVCSR) with previously trained phone models. This recognition provides the segmentation and classification of segments. The drawback of this approach is a necessity of large annotated database for training of the phone models.
2. The second possibility is to use Data-driven techniques based on ALISP (Automatic Language Independent Speech Processing) tools [3]. The segmentation can be obtained automatically on raw data without any transcriptions. The drawback of this approach is a need to ensure the coherence of classes across speakers and to determine the verification power of classes, not known a-priori.

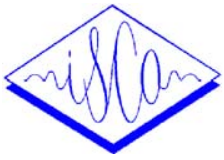
This paper concentrates on the latter approach and discusses its advantages and drawbacks. It is organized as follows: section 2 deals with the unsupervised segmentation and labelling of speech signals. A comparison with phonetic alignment is presented. The two following sections deal with two different methods of modeling the speaker PDF in classes: Multi-Layer Perceptrons (MLP) and Gaussian Mixture Models (GMM). Both sections are completed by the description of the experimental setup and results. Section 5 deals in more detail with the combining of class-dependent scores using mixture of experts techniques.

2. SEGMENTATION

For performing the segmentation, ALISP tools [3, 9] were used. Unlike LVCSR, this approach requires neither phonetic nor orthographic transcription of the corpus, as it is based on the speech data. The speech segmentation is achieved using *temporal decomposition* (TD) [1, 2]. 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 is trained by K-means algorithm with binary splitting [4]. 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 class C_l determined through the segmentation and the labelling. The number of classes is fixed by the number of centroids in the VQ codebook. In the experimental work, 8 classes have been used.

2.1. Correspondence of phonetic and ALISP segmentations

To study the correspondence of the automatic segmentation and labelling, a comparison with phonetically aligned data was done. 4.8 minutes of hand labelled SWITCHBOARD data (used in the CLSP'96 and '97 workshops) were used for the comparison. The phonemes were grouped into 7 phonetic categories: vowels, stops, fricatives, affricatives,



6th European Conference on Speech Communication and Technology (EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

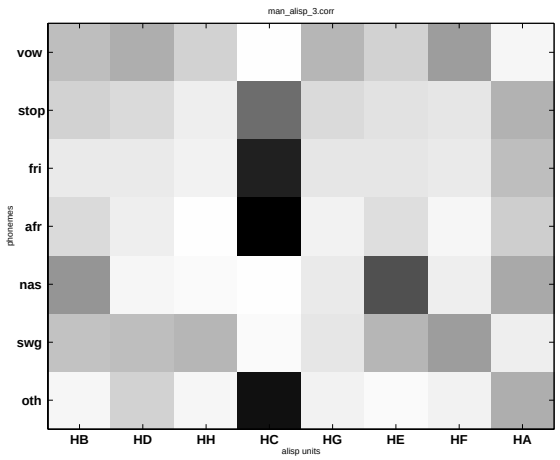


Figure 1: Confusion matrix of broad phonetic classes and ALISP units. White color corresponds to zero correspondence, black to maximal correspondence

nasals, semivowels+glides, and others (silence). The ALISP units were derived using the procedure described in previous paragraph. The correspondence was evaluated using relative overlaps of ALISP units (denoted HA...HH) with the phonemes over the entire dataset.

The resulting confusion matrix, with phonetic classes as references, is shown in Fig. 1. It is obvious, that the correspondence is far from one-to-one; the ALISP unit `HC` represents for example well the noise and pause parts of signal, but especially vowels and semi-vowels are spread over almost all ALISP classes. Therefore, the VQ with such a low number of classes does seem to be comparable with phonetic categories; in [10] we have shown, that a larger number of classes maps more precisely to phonemes. However, in the speaker verification experiments, only 8 classes were used.

3. MLP MODELING

One of the main reasons for using MLPs for modeling purposes, is their discriminant capability [9]. In this approach, each segmental MLP (one per class) is trained in a discriminative manner, to distinguish between the client speaker and a background world model, using only feature vectors having the corresponding class label. For example, the MLP associated with class C_i provides the following segmental LLR_i score:

$$LLR_i = \log(S_{ci}) - \log(S_{wi}), \text{where} \quad (1)$$

$$S_{ci} = \prod_{x \in C_i} P(M_{ci}|x)/P(M_{ci}), \quad (2)$$

$$S_{wi} = \prod_{x \in C_i} P(M_{wi}|x)/P(M_{wi}) \quad (3)$$

The products involve vectors being previously labelled as members of class C_l. Subscripts ci and wi denote respectively the client and the world MLP-outputs for the segmental class C_i.

3.1. MLP: Experiments and results

The segmental MLPs were tested on NIST-NSA'98 data and the experimental setup and results were thoroughly described in [9]. Each segmental MLP had 20 neurons in the hidden layer and worked with the context of 5 acoustic frames. The class-dependent scores were recombined with equal weights. The results can be summarized as follows: the segmental MLP system reached almost the same performances as a global system with one MLP for the easy training-test condition SN (data in training and in test files come from the same telephone number) and outperformed the global MLP in difficult training-test condition DT (different type – the test file was recorded using different handset type).

4. GMM MODELING

In this approach, the client as well as the world PDF are modeled by the mixture of Gaussian distributions: $\mathcal{L}(x_n|M) = \sum w_j \mathcal{N}(x_n; \mu_j, \Sigma_j)$. If N is the number of classes, N GMMs must be trained for each client. In case of the world, we can choose either a segmental approach (N world GMMs) or a global world model. The latter approach was tested in our experiments. In the testing phase, two possibilities exist for scoring using the class-GMMs:

- each frame is first assigned to a class (using temporal decomposition and VQ) and only the corresponding GMM is used for the scoring. This approach can be denoted "hard".
- a "soft" appertaining function of each vector to all classes is evaluated, the vector is scored by *all* GMMs and their outputs are weighted by the appertaining function.

We have used the latter approach with the following function quantizing the appertaining of vector x to class i:

$$w_i = \exp\left[\left(1 - \frac{d_i}{\sum d_i}\right)^4\right], \quad (4)$$

where d_i is the Euclidean distance of x from the centroid of class i and $\sum d_i$ is the sum of distances over all classes. Weights w_i are normalized to sum up to 1. The global LLR of test file (L frames) is then computed as:

$$LLR = \sum_{n=1}^{L} \left[\sum_{i=1}^{N} w_i \left(\mathcal{L}(x_n|M_{ci}) - \mathcal{L}(x_n|M_w)\right)\right], \quad (5)$$

and finally normalized by the number of frames L. It is obvious, that in this case, no weighting taking into account discriminative powers of classes is performed.

4.1. GMM: Experiments and results

Segmental GMM approach was tested on the ELISA-1 [5] subset of NIST-NSA'98 data: 50 client speakers, each with 2 minutes of training data (condition 2S) and about 240 test files per sex (duration 3 seconds). The parameterization was done using 16 LPCC coefficients with liftering. Each segmental client GMM had 64 mixture components for

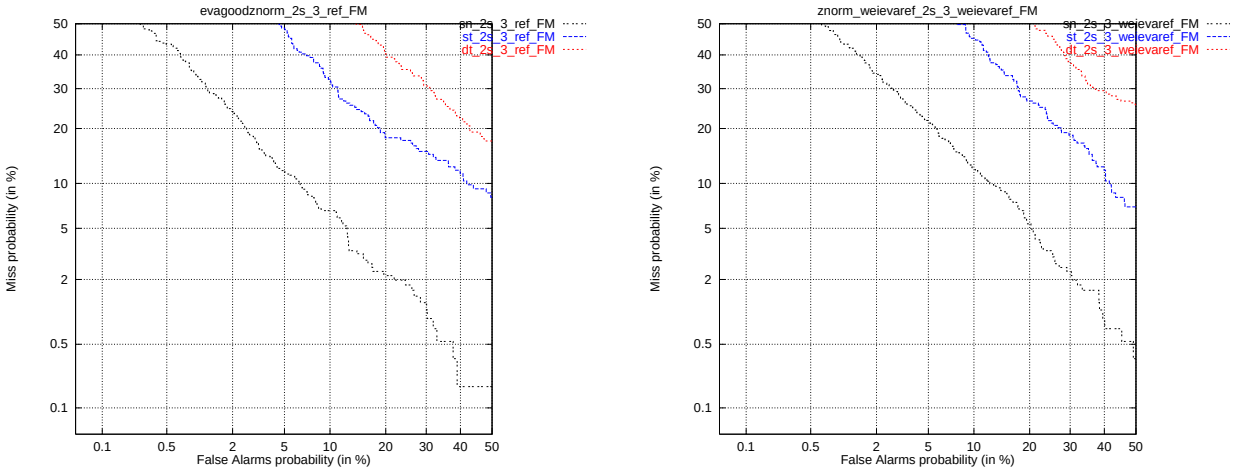


Figure 2: Results of GMM systems for three training-test conditions: SN (same number), ST (same type) and DT (different type). Left panel: one GMM per client. Right panel: 8 segmental GMMs per client.

basic parameters, 64 for ΔLPCC and 16 for Δlog-energy. The models were initialized by all the data from the given speaker and then each of 8 segmental models was retrained using only the data corresponding to i-th class. The world model was non-segmental, gender dependent and of the same configuration as segmental ones. It was trained using 50 electret and 50 carbon background speakers.

First, a global GMM system was developed, all the test files were scored and normalized using impostor accesses (50 electret and 50 carbon impostors) and *znorm* (handset-independent normalization). Its results for the 3 conditions can be seen in the left panel of Fig 2. Then, the scoring was performed with above described segmental system, with similar normalization (see the right panel of Fig 2 for results).

It is obvious that the segmental GMM system reaches lower performances than the global one for all three training-test conditions. The most probable reason is a limited amount of training data available for the re-estimation of segmental models. An adaptation strategy [7] is a good candidate to bring better results. Also, similarly 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. COMBINING THE OUTPUTS OF THE SEGMENTAL EXPERTS

5.1. Logistic regression

Instead of using equal weights for combining the outputs of all segmental models (experts), these outputs can also be combined using a data fusion method based on the logistic regression model presented in [11, 12]. This method performs a statistical analysis of the observed data (training data) and the discrimination function it implements is the *logistic distribution function*, which is formalized hereafter:

$$E(Y/s) = \Pi(s) = \frac{e^{g(s)}}{1 + e^{g(s)}} \tag{6}$$

In this expression $E(Y/s)$ is the conditional probability for the (binary) output variable Y given the N-dimensional input vector s, with $g(s) = \beta_0 + \beta_1 s_1 + \cdots + \beta_N s_N$ and $s = (s_1, s_2, \cdots, s_N)$. This equation gives as a result for the input pattern s, the probability $\Pi(s)$ of belonging to the class of clients ($Y = 1$) and, in an indirect manner, the probability $[1 - \Pi(s)]$ of belonging to the class of impostors ($Y = 0$). Since each β_i with $i \neq 0$ multiplies one of the N experts, and if *all* experts output scores are in the same range, the value of β_i is a measure of the importance of the i-th expert in the fusion process. A large β_i indicates an important expert, a small β_i indicates an expert that does not contribute very much. The idea is then that this principle should lead to the weighting of the different classes according to their discriminative powers.

5.2. Mixture of Experts

Instead of using the weights obtained by the appertaining function defined in Eq. 4 for combining the outputs of all segmental models (experts), these outputs can also be combined using a data fusion method based on the Mixture of Experts paradigm presented in [6]. To weight the likelihood ratio outputs LLR_i of each of the segmental experts, we add a MLP, which will serve as *gating network*. This gating network receives the same acoustic vectors as input as the segmental experts and has eight output neurons with *softmax* activation functions. This softmax function assures that the outputs of gating network sum to unity and are non-negative, thus implementing the (soft) competition between the different segmental experts [8]. These N different output values are noted W_i, and they will be used to weight the N output LLR_i of the N segmental experts in the following manner:

$$\text{Total LLR} = \sum_{i=1}^{N} W_i LLR_i \tag{7}$$

The gating network is trained using speech segments from the claimed speaker. For these speech segments, the target

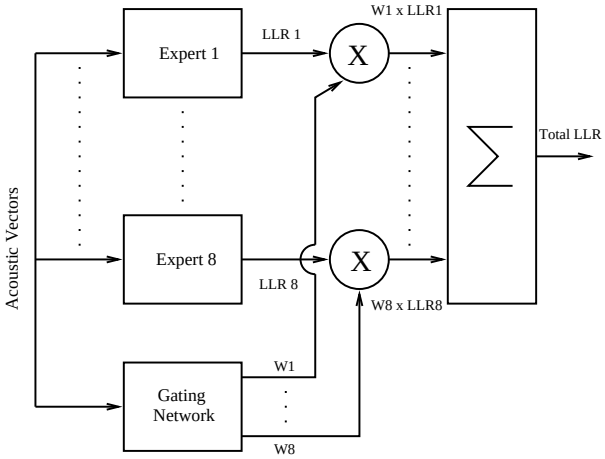


Figure 3: Combining the outputs of the different segmental experts

vector is 1 for the output neuron corresponding with the largest LLR_i, and 0 for the $N-1$ other outputs. During the test phase, the N output neurons of the gating network are going to vary with the presented input segment. This means that if an input segment is lying close to k class segmentation prototypes, this will be translated by the fact that k different output neurons will tend to have significant outputs. In this manner, k segmental experts will significantly and proportionally contribute to the total LLR. The structure of this data fusion paradigm is represented in Fig. 3.

Experiments with the logistic regression and gating network are being conducted on the ELISA-1 subset of NIST'98 data.

6. CONCLUSIONS

Several methods of segmental, text-independent speaker verification with automatically derived classes of speech sounds were presented. We have confirmed, that speech segments are not equal in characterizing speakers. Nevertheless an optimal grouping of acoustic segments for speaker verification has not been found so far. In comparison with linear recombination of class-dependent scores, the "mixture of experts" approach is elegant and needs to be further investigated. It is likely that a system based on speaker independent segmental HMMs (LVCSR) adapted to each client is the next thing to try. In this framework, the need for all class-dependent models scoring in parallel is unclear.

7. ACKNOWLEDGEMENTS

We are grateful to Joe Picone (ISIP/Mississippi State Univ.) for having provided us with the SWITCHBOARD data used in the CLSP '96 and '97 workshops. Thanks also to Fréderic Bimbot (IRISA, France) for having allowed us to use his temporal decomposition `td95` package, and to Jean Hennebert (UBILAB, Switzerland) for the permission to use his POST speech processing software. This work is supported by the Ministry of Education, Youth and Sports of Czech Republic under the project No. VS97060.

8. REFERENCES

[1] B. S. Atal. Efficient coding of LPC parameters by temporal decompositio n. 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] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. *Computational models of speech pattern processing*, chapter Towards ALISP: a proposal for Automatic Language Independent Speech Processing, pages 375–388. NATO ASI Series. Springer Verlag, 1999.

[4] Allen Gersho and Robert Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[5] G. Gravier. The ELISA-1 system description. Technical report, ENST Paris, February 1999.

[6] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[7] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, (9):171–185, 1995.

[8] P. Moerland. Mixtures of experts estimate a posteriori probabilities. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Proc. Intl. Conf. on Artificial Neural Networks (ICANN'97)*, number 1327 in Lecture Notes in Computer Science, pages 499–504, Berlin, 1997. Springer-Verlag.

[9] D. Petrovska-Delacrétaz, J. Černocký, J. Hennebert, and G. Chollet. Text-independent speaker verification using automatically labelled acoustic segments. In *International Conference on Spoken Language Processing (ICLSP)*, Sydney, Australia, December 1998.

[10] J. Černocký, G. Baudoin, and G. Chollet. The use of ALISP for automatic acoustic-phonetic transcription. In *Proc. SPoSS - ESCA Workshop on Sound Patterns of Spontaneous Speech*, pages 149–152, Aix-en-Provence, France, September 1998.

[11] P. Verlinde and G. Chollet. Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In *Proc. 2-nd Intl. Conf. on Audio- and Video-based Biometric Person Authentication*, pages 188–193, Washington D. C., USA, March 1999.

[12] P. Verlinde, P. Druyts, G. Chollet, and M. Acheroy. Applying Bayes based classifiers for decision fusion in a multi-modal identity verification system. In *International Symposium on Pattern Recognition "In Memoriam Pierre Devijver"*, Brussels, Belgium, February 1999.