PHONEME BASED TEXT-PROMPTED SPEAKER VERIFICATION WITH MULTI-LAYER PERCEPTRONS

Jean Hennebert and Dijana Petrovska Delacrétaz

Chair of Circuits and Systems
Swiss Federal Institute of Technology

RÉSUMÉ

Ce papier présente une étude d'un système de vérification du locuteur utilisant des Chaînes de Markov Cachées (HMMs) et des Perceptrons Multi Couches (MLPs). Les objectifs du travail sont (1) d'évaluer les propriétés discriminantes des phonèmes avec différentes tailles de contexte à l'entrée des MLPs et (2) d'étudier deux techniques d'échantillonnage des vecteurs acoustiques pendant la phase d'entraînement des MLPs.

ABSTRACT

Results presented in this paper are obtained in the framework of a text-prompted speaker verification system using Hidden Markov Models (HMMs) and Multi Layer Perceptrons (MLPs). The aims of the study described here are (1) to assess the relative speaker discriminant properties of phonemes with different temporal frame-to-frame context at the input of the MLP's and (2) to study the influence of two sampling techniques of the acoustic vectors while training the MLP's.

1. INTRODUCTION AND MOTIVATIONS

The work presented here is the continuation of a previous study [14] in which the relative speaker discriminant properties of phonemes were investigated. In short, it was shown that, with similar experimental conditions, nasals, fricatives and vowels convey more speaker specific informations than plosives and liquids. Other studies [5] [13] have also reported that some phonemes have more discriminant power than other as far as the speaker verification is concerned.

Text-independent and text-dependent speaker verification systems (passwords, pin codes, ...) are too weak from a security point of view because they can easily be broken with pre-recorded speech of the client. Text-prompted systems, in which the text to utter is prompted with different word sequences from session to session, have been introduced in order to close the door to system breakers using pre-recorded speech [10]. Such a procedure works efficiently if the vocabulary of the system is large enough. Indeed, modern digital recorders can play back an arbitrary sequence of keywords so that text-prompted systems with fixed small vocabulary, like digits, can also be broken.

The characteristics of the text-prompted system which is proposed here, are as follows. The speech recognition part is performed by a set of context independent phoneme (CIP) HMMs which enable the recognition of large vocabulary. The speaker verification part is performed by MLPs trained to classify the acoustic vectors into the claimed speaker or a world speaker class. Results presented in this paper focus on the speaker verification part of the system and it is assumed for the rest of the discussion that the speech recognition is performed error-free.

The set of CIP HMMs are used to provide a segmentation of the speech signal into phonemes with a simple Viterbi forced alignment. The feature vectors, labelled with the corresponding phonemes, are then used to train MLPs, one per phoneme and per client. In previous studies , MLPs have successfully been used for text-independent [12] [6] and for fixed-text [11] speaker recognition tasks. The main advantages of MLPs against other systems like Gaussian Mixture Modelling include, among others, discriminant capabilities, weaker hypotheses on the acoustic vector distributions and possibility to include a larger acoustic frame window as input of the classifier.

Similarly to what is done in speech recognition with hybrid HMMs/MLP systems [3], this approach combines the ability of HMMs to handle efficiently the sequential character of speech and the discriminant properties of ANNs. The main drawback using MLPs is that its optimal architecture (essentially the number of hidden nodes) must be selected by trials and errors. Another drawback lies in problems that can occur when the amounts of training patterns are too dissimilar between classes.

The main aim of the study described in this paper is to assess the relative speaker discriminant properties of phonemes while investigating the importance of the temporal frame-to-frame context at the input of the MLPs. A Swiss German telephone speech database is used for the experiments. Results are also reported regarding two different sampling techniques of the acoustic vectors while training the MLP's.

2. SYSTEM DESCRIPTION

2.1. Feature Extraction

The extraction of salient features for speaker verification is not addressed in this paper. Lpc-cepstrum are known to present good performances while being very inexpensive to compute and are used for both the speech recognition and speaker verification modules. The speech data is initially processed by the application of a pre-emphasis filter $H(z) = 1 - z^{-1}$. A 30 ms Hamming window is applied to the speech signal 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 subtraction in order to operate a blind deconvolution. Energy and dynamic information (delta coefficients)

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 and by the grant Marie Heimvögetlin of Swiss National Funds for Research.

were used for the speech recognition part but not for the speaker verification part.

2.2. Speech Recognition Part

As previously said, results presented in this paper focus on the speaker verification part of the system, assuming no errors in the speech verification step. 42 Swiss-German CIP HMMs are trained using the whole set of speakers available in the database and are then used to generate segmentation into phonemes using a simple Viterbi forced alignment.

2.3. Speaker Verification Part

MLPs, one for each phoneme/speaker, are discriminatively trained to distinguish between the client speaker and a background world model. MLPs with two outputs are used, one for the client class C_1 and the other for the world class C_2. In [4] it has been proved that if each output unit k of a MLP used in a classification problem, is associated to a class C_k of our problem, it is possible to train the MLP to generate a posteriori probabilities $p(C_k|\mathbf{x}_n)$ when $\mathbf{x}_n$, a particular acoustic vector, is provided to its input.

2.3.1. Architecture and Training Procedures

Concerning the architecture, 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. It has been previously shown [12] that using more than one hidden layer did not improve the performance for a speaker identification task and thus this architecture has not been investigated here.

During training, target vectors $d(\mathbf{x}_n)$ are set to $[1, 0]$ and $[0, 1]$ when the input vector $\mathbf{x}_n$ is produced by, respectively, the client and by the world speaker. Two different sampling procedures are used to present the acoustic vectors as input to the MLP's. For the first procedure, referred as baseline (BL), the acoustic vectors are picked randomly in the whole training set, build up with the client and world vectors. For the second procedure, referred as Equal File Sampling (EFS), the acoustic vectors are also picked randomly in the training set, but this time taking successively one vector in the client training set and one vector in the world training set. The EFS sampling mode is expected to avoid problems due to large difference of size between client and world training sets. Performing EFS does not change the behaviour of the MLP, in the sense that it will still estimate posterior probabilities, but with equal class priors values $P(C_k) = 0.5$, as explained in the Bayes formula:

$$P(C_x|\mathbf{x}_n) = \frac{p(\mathbf{x}_n|C_k)P(C_k)}{p(\mathbf{x}_n)} \tag{1}$$

The error criterion used for training is defined as

$$E = \sum_{n=1}^{N} ||g(\mathbf{x}_n) - d(\mathbf{x}_n)||^2 \tag{2}$$

where g is the non-linear vector function operated by the MLP on the input vector $\mathbf{x}_n$. As explained in [15], the parameters of the MLP (weight matrices) are iteratively updated via a gradient descent procedure in order to minimise the error criterion in (2). The weights are updated after every input presentation during the training process. The correction of the matrices values is weighted by a *learning rate* value η which is updated after a presentation of the whole training set (epoch) with the following rule:

- set $\eta_{i+1} = \frac{\eta_i}{2}$ if the error measure E on a independent cross-validation data set is increasing from epoch $i-1$ to epoch i.
- set $\eta_{i+1} = \eta_i$ if the error measure E on a independent cross-validation data set is decreasing from epoch $i-1$ to epoch i.

Observing an increasing error measure on a independent cross-validation data set from one epoch to another is a sign of over-fitting on the training data set. In order to avoid over-fitting, the update of the weight matrices is discarded before setting the new learning rate value and pursuing with the next epoch. Training is stopped when η falls below a pre-determined value.

When using the EFS sampling, a training epoch is over when all the feature vectors of the largest training set are visited once. This implies that vectors in the smaller set may be visited more than one time during the same epoch.

2.3.2. Decision Making

The output of the MLP provides estimations of the client and world a posteriori probabilities at the frame level. The client and world scores for a sequence of N vectors belonging to a phoneme k can be obtained as follows, assuming independence of the observation vectors.

$$S_{1k} = \sum_n log(p(C_1|\mathbf{x}_n)) \tag{3}$$

$$S_{2k} = \sum_n log(p(C_2|\mathbf{x}_n)) \tag{4}$$

In this paper, the recombination of scores S_{1k} and S_{2k} in order to take a decision at the word level is not investigated. Instead, speaker verification EER are computed directly on the S_k measures in order to study the discriminative power of the different phonemes.

A thresholding procedure is applied in order to find the EER, point of intersection between the false acceptance and false rejection curves. It could be argued that if MLPs are actually estimating the posterior probabilities of the classes, it would not be necessary to use a thresholding procedure. The same discussion can take place also for non-discriminant likelihood approaches in which in theory, a majority vote on the class likelihood should be enough to determine the EER (if class priors are equal). The problem lies, for likelihood estimators and for a posteriori probability estimators, in the fact that they are biased estimators due to the lack of training datas.

3. DATABASE DESCRIPTION

The HER Swiss German telephone speech database has been used for the experiments. HER has been recorded in the framework of the *Himarnnet P6488 Esprit Project* dedicated to speech recognition using HMMs and Artificial Neural Network (ANNs). This Swiss German spoken telephone speech database contains 108 phonetically balanced isolated words uttered by 536 speakers. The 108 words were recorded in one session by each talker and are identical from speaker to speaker.

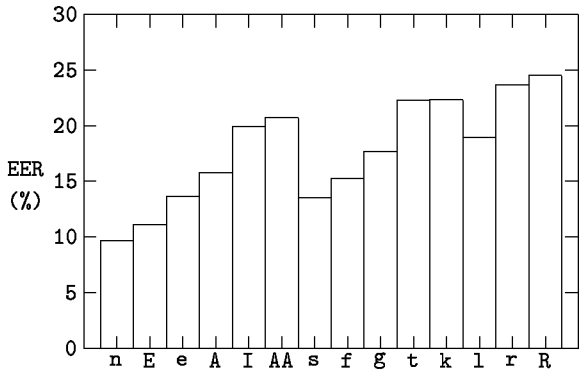

Figure 1: EER averaged per phoneme with 20 hidden nodes and 3 input frames for the MLPs.

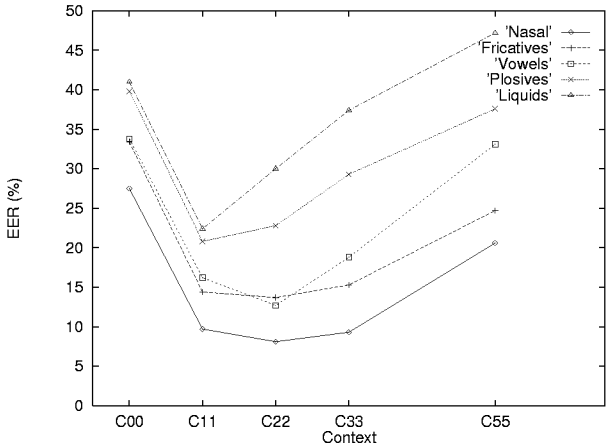

Figure 2: EER averaged per phonemic group with different acoustic window length at the input of the MLP. The number of hidden nodes equals 20 and is kept constant for all the experiments.

25 male speakers were selected as the clients of the system. 25 other male speakers were selected to constitute the background model and 25 male speakers were used as impostor speakers. Cross-sex tests, and female against female tests, as described in the Eagles recommendations [2] are under investigation.

In order to minimise the influence of lack of training data when building the models, a reduced set of 14 phonemes having more than 22 occurrences in the database was selected. The training data set for each phoneme model is obtained from a concatenation of 8 client segments and 200 world segments. Independent cross-validation sets were defined in the same way, concatenating 5 segments of each phonemes for the client and 125 segments for the world. 2 true-identity tests were defined for each speaker phoneme model, concatenating 4 and 5 segments. 125 impostor tests were defined for each speaker phoneme model, concatenating 4 segments.

In order to have more testing material, 8 distinct train, cross and test data sets were defined from the 22 phoneme occurrences available by concatenating segments in different order. The total true-identity tests and impostor tests are then respectively 400 and 25000.

4. RESULTS AND DISCUSSIONS

4.1. Results by phoneme with different MLP input frame context

Figure 1 shows the averaged EER for the 14 selected phonemes. Results were obtained with a 20 hidden nodes MLP trained on 3 consecutive acoustic frames as input with the baseline sampling procedure. The best performance is obtained with phoneme *n* which is a nasal. Vowels (*E, e, A, AA, I*) and fricatives (*s, f*) give good and similar performances while plosives (*g, t, k*) and liquids (*l, r, R*) convey less speaker specific informations. Per speaker detailed results show that some phonemes perform better with some speakers while the same phonemes perform badly with other speakers.

It should be pointed out that results are obtained training MLPs with the same occurrence of each individual phonemes and no length normalisation of the segments has been performed. Very similar results are reported in [5] in which a phonetically hand-labelled database is used to train a VQ based speaker verification system.

The influence of the acoustical window length at the input of the MLP has been investigated adding symmetrically left and right frames to the central frame. Experiments with 1, 3, 5, 7 and 11 successive frames at the input of the MLP are reported on figure 2 and are noted as Cxy (meaning x left frames and y right frames of context taken into account). EER are averaged in phonetic classes for clarity's sake. Significant improvements are brought when increasing the acoustic window size from $C00$ to $C11$. This result suggests that the frame-to-frame temporal informations convey important speaker specific informations. Increasing furthermore the size of the input to $C22$ improved somehow performances for nasal, fricatives and vowels while plosives and liquids got worse performances. Performing the training of the speaker phoneme specific MLP's with 7 ($C33$) and 11 ($C55$) successive frames make worse results, except for phonemes *e* and *f* in the case of $C33$. Results in figure 2 show also that each phoneme class has its optimal MLP input size which gives the best EER. For nasal *n*, vowels and plosives, the best results are obtained with $C22$, while for plosives and liquids better performances are obtained with $C11$.

Results reported in figure 1 and 2 are all obtained with MLPs having 20 nodes on the hidden layer and the baseline sampling procedure was used. The influence of the number of hidden nodes is an important issue but is not addressed here. Some results and discussions regarding this issue can be found in our previous work [14]. The inclusion of delta coefficients at the input of the MLP has not been investigated since a delta computation is a simple linear operation on successive acoustic vectors and since such an operation can be performed by the MLP at its first layer if used with context.

The improvements obtained when larger acoustic frame windows are used are quite important for the configuration investigated here. In the literature, different text-independent predictive systems have been proposed with mitigated results to specifically include the temporal information: predictive MLP's [9] [7] [1] giving encouraging results and Auto-Regressive (AR) vector models [8] giving contradictory results. Comparing with these results obtained with text independent systems, the large amelioration re-

	Nasal	Fricatives	Vowels	Plosives	Liquids
C00-BL	27.5	33.4	33.8	39.8	41.0
C00-EFS	23.5	30.9	32.9	38.6	39.2
C11-BL	9.7	14.4	16.2	20.8	22.4
C11-EFS	10.6	13.4	15.4	21.6	22.6

Table 1: EER averaged per phonemic group with two different training sampling modes of the MLP.

ported here is probably due to the fact that it is easier to exploit the temporal information at the phoneme level than at the word level.

4.2. Influence of the sampling mode for the MLP training

The world training set is, on average, 25 times larger than the client training set and training problems due to this large difference were suspected. As explained in section 2.3.1, an alternative sampling procedure referred as EFS is compared to the standard sampling used for training MLPs. EER using the baseline (BL) and EFS sampling procedure are given in table 1 for the $C00$ and $C11$ configuration. For $C00$, the EFS procedure shows some improvements while for $C11$, there are no significant differences. A clear explanation for this behaviour is not straightforward. One could argue that the $C11$ configuration brings a better class separability which smooth out problems of unbalanced class populations.

5. FUTURE WORK

Future work will be dedicated to the following points :

- Validate results on a multisession database.
- Investigate phone score recombination strategies, trying to take advantage of the different discriminative power of phonemes.
- Compare performances of the phoneme based text-pompted system presented here with a baseline text-independent system in order to put in evidence the pros and cons of the approach.
- Study the performances of the global system including the speech verification part and the speaker verification part.

6. CONCLUSIONS

The text-prompted system which is proposed here combines the ability of HMMs to handle efficiently the sequential character of speech and the discriminant properties of MLPs.

The discriminative power of the most frequently appearing phonemes is reported and the influence of the acoustic window length at the input of the MLP is studied in the framework of a telephone database. According to the experiments, nasals, fricatives and vowels are found to provide the best performances, followed by plosives and liquids. Significant improvements are reported from the inclusion of several acoustic frames at the input of the MLPs and each phoneme class seems to have its optimal MLP input size which gives the best EER.

An alternative input vector sampling procedure used during training and referred as EFS seems to improve the quality of the MLP when no context is used in input. When MLPs are trained with context, no difference between the EFS and the baseline sampling is observed.

All the results presented in this paper are of course specific to the database, to the language and to the configuration that has been used throughout the experiments.

7. REFERENCES

[1] T. Artières. *Méthodes prédictives neuronales: application à l' identification du locuteur.* PhD thesis, Université de Paris XI Orsay, 1995.

[2] F. Bimbot and G. Chollet. *EAGLES Handbook on Spoken Language Systems*, chapter Assessment of speaker verification systems. Mouton de Gruyter, 1997.

[3] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition.* Kluwer Academic Publishers, 1994.

[4] Hervé Bourlard and C. J. Wellekens. Links between markov models and multi-layer perceptrons. *IEEE Trans. Patt. Anal. Machine Intell.*, 12(Inconnu):1167–1178, Inconnu 1990.

[5] 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.

[6] K. A. Farrell, R. Mammone, and K. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on Speech and Audio Processing*, 2(1):194–205, 1994.

[7] H. Hattori. Text-independent speaker verification using neural networks. In *Workshop on Automatic Speaker Recognition and Verification*, pages 103–106, Martigny, Switzerland, April 1994.

[8] I. Magrin-Chagnolleau, J. Wilke, and F. Bimbot. A further investigation on ar-vector models for text-independent speaker identification. In *ICASSP*, pages 401–404, Atlanta, GA, May 1996.

[9] C. Montacie, P. Deleglise, F. Bimbot, and M. J. Caraty. Cinematic techniques for speech processing : temporal decomposition and multivariate linear prediction. In *ICASSP*, volume 1, pages 153–156, San-Francisco, 1992.

[10] J. M. Naik. Speaker verification: A tutorial. *IEEE Communications Magazine*, 28(1):42–48, January 1990.

[11] Jayant M. Naik and David M. Lubenskt. A hybrid hmm-mlp speaker verification algorithm for telephone speech. In *ICASSP*, pages 153–156, 1994.

[12] J. Oglesby and J. S. Mason. Optimization of neural models for speaker identification. In *ICASSP*, pages 261–264, 1990.

[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 *Submitted to ICASSP*, Seattle, 1998.

[15] 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.