TEXT-PROMPTED SPEAKER VERIFICATION EXPERIMENTS WITH PHONEME SPECIFIC MLPS

Dijana Petrovska Delacrétaz and Jean Hennebert

Circuits and Systems Group
Swiss Federal Institute of Technology
email : petrovska,hennebert@circ.de.epfl.ch

ABSTRACT

The aims of the study described in this paper are (1) to assess the relative speaker discriminant properties of phonemes and (2) to investigate the importance of the temporal frame-to-frame information for speaker modelling in the framework of a text-prompted speaker verification system using Hidden Markov Models (HMMs) and Multi Layer Perceptrons (MLPs). It is shown that, with similar experimental conditions, nasals, fricatives and vowels convey more speaker specific informations than plosives and liquids. Regarding the influence of the frame-to-frame temporal information, significant improvements are reported from the inclusion of several acoustic frames at the input of the MLPs. Results tend also to show that each phoneme has its optimal MLP context size giving the best Equal Error Rate (EER).

1. INTRODUCTION AND MOTIVATIONS

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 [11] [15]. 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.

In a large vocabulary text-prompted system, the speaker verification is done in two steps. First, the content of the speech signal is verified in order to check if the speaker said what he was asked to say. This implies using a speech recognition system which is generally speaker independent and which is usually based on sub-word models in order to recognize a very large set of words. In practice, a full recognition will not be necessary because the lexical content of the prompted text is known. Once the pre-verification is done, the real speaker verification part can take place. Any usual technique can be applied but using the lexical information which is known from the prompt can bring advantages in two ways.

1. During testing and training time, the a priori knowledge of the lexical content of the utterance can be used to automatically and efficiently segment the speech input into sub-word units like phonemes. Speaker and impostor models can then be trained at sub-word levels, which has been shown to give good performances [9].
2. Studies [5] [14] have already reported that some phonemes have more discriminant power than other as far as the speaker verification is concerned. This fact could be advantagely used to build a more robust scoring procedure taking into account the different discriminant power of phonemes. Furthermore, the prompting could also be driven by this information in order to get as much as possible occurrences of those phonemes in the speech input.

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 and 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 verification 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 succesfully been used for text-independent [13] [6] and for fixed-text [12] 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 classifyer.

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.

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 unexpensive to compute and are

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

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 substraction in order to operate a blind deconvolution. Energy and dynamic information (delta coefficients) 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 Procedure

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 [13] 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 $d(\mathbf{x}_n)$ were 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. During the training phase of the MLP's, the acoustic vectors were presented randomly from the available training set.

The error criterion used for training is defined as

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

were g is the non-linear vector function operated by the MLP on the input vector $\mathbf{x}_n$. As explained in [16], the parameters of the MLP (weight matrices) are iteratively updated via a gradient descent procedure in order to minimize the error criterion in (1). 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.

The case of an increasing error measure on a independent cross-validation data set from one epoch to another is a sign of overfitting on the training data set. In order to avoid overfitting, 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.

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.

$$S_{1k} = \sum_n log(p(C_1|\mathbf{x}_n)) \quad (2)$$

$$S_{2k} = \sum_n log(p(C_2|\mathbf{x}_n)) \quad (3)$$

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 alarm 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. 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

A Swiss German telephone speech database called the HER 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.

25 male speakers were selected as the clients of the system. 25 other male speakers were selected to constitute the backgroud 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 minimize 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 background segments. Independent cross-validation sets were defined in the same way, concatenating 5 segments of each phonemes for the client and 125

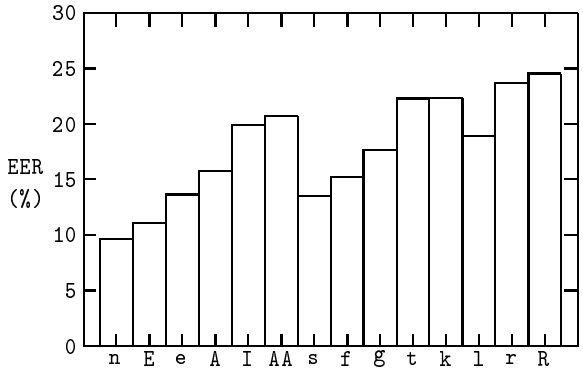

Figure 1: EER averaged per phoneme with 20 hidden nodes and 3 input frames for the MLPs

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 set 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

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

4.2. Influence of the MLP input frame context

	Nasal	Fricative	Vowels	Plosives	Liquids
C00-20	27.5	33.4	33.8	39.8	41.0
C11-20	9.7	14.4	16.2	20.8	22.4
C22-20	8.1	13.7	12.7	22.8	30.0

Table 1: EER averaged per phonemic group with different acoustic window length at the input of the MLP. The number of hidden nodes is kept constant

The influence of the acoustical window length at the input of the MLP has been investigated adding symetrically left and right frames to the central frame. Experiments with 1, 3, and 5 successive frames at the input of the MLP are reported on table 4.2 and noted as, respectivelly $C00$, $C11$ and $C22$. Cxy meaning simply x left frames and y right frames of context taken into account. EER are averaged in phonetic classes for a sake a clarity. 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, fricative and vowels while plosives and liquids got worse performances. Results presented in table 1 show that each phoneme has its optimal MLP input size which gives the best EER.

The improvements obtained when temporal frame-to-frame information are quite important for the configuration investigated here. In the litterature, different text-independent predictive systems have been proposed with mitigated results to specifically include the temporal information: predictive MLP's [10] [7] [1] giving encouraging results and Auto-Regressive (AR) vector models [8] giving contradictory results. Comparing with these studies, the large amelioration reported here is probably due to the fact that the problem of exploiting the temporal information is done at the phoneme level which is not the case in the above mentionned works that are text independent.

4.3. Influence of the number of MLP parameters

Performances with different number of parameters in the MLPs (number of hidden units) is shown on table 2. The MLP input is fixed to one acoustic frame with no context. A controlled experiment between a $C00$ and a $C11$ configuration in which the number of parameters is kept constant has been performed. The idea was to evaluate if the improvement when using a larger MLP input is due to the larger number of parameters or from informations in the frame context. A $C11$ net with 20 hidden nodes (760 weights) was compared to a $C00$ net with 54 nodes (756 weights). The $C00-54$ configuration performed worse than the $C11-20$ but unfortunately, it seems that there is a lack of training data or a problem of local minimum in order to train the $C00-54$ nets since the performance is even worse that the $C00-20$ configuration. Increasing the number of hidden nodes can result in a more powerful and complex classification function but which is more subject to overfitting and "getting stuck" in a local minima.

	Nasal	Fricative	Vowels	Plosives	Liquids
C00-10	28.1	32.6	36.4	41.4	41.7
C00-20	27.5	33.4	33.8	39.8	41.0
C00-54	31.9	36.2	38.4	43.3	43.8

Table 2: EER averaged per phonemic group with different number of hidden nodes in the MLP.

5. CONCLUSION

The speaker verification part of a large vocabulary text-prompted system has been investigated. MLPs were used for speaker specific phoneme modeling, using the automatic segmentation provided with speaker independent HMMs. The discriminative power

of the most frequently appearing phonemes was investigated. According to the experiments, nasals, fricatives and vowels are found to provide the best performances, followed by plosives and liquids. The influence of the acoustic frame window length at the input of the MLP is studied and significant improvements are reported from the inclusion of several acoustic frames. Results tend to show that each phoneme has its optimal MLP input size giving the best EER.

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 experiements. Future work will be carried on in order to validate the forementionned conclusions on other databases.

6. 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] Romoko Matsui and Sadaoki Furui. Concatenated phoneme models for text-variable speaker recognition. In *ICASSP*, volume 2, pages 391–394, Minneapolis, April 1993.

[10] 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.

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

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

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

[14] 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.

[15] A. E. Rosenberg and F. K. Soong. Recent research in automatic speaker recognition. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 701–738. Marcel Dekker, New York, 1992.

[16] 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 of the Microstructure of Cognition*, volume 1. MIT Press, 1986.