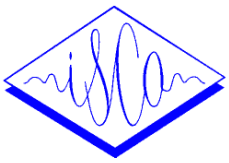

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

MULTI-STATE PREDICTIVE NEURAL NETWORKS FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

T. Artières, P. Gallinari

LAFORIA UA CNRS 1095
Tour 46-00, Boite 169
Université Paris 6, 4 place Jussieu
75252 Paris cedex 05 France
artieres@laforia.ibp.fr, gallinari@laforia.ibp.fr
fax : (33)-1-44-27-70-00

ABSTRACT

Both Hidden Markov Models and Neural Networks have already been used as production systems for speaker identification or verification.

Recently [9] has shown that ergodic multi-state hidden Markov Models do not outperform one-state "hidden" Markov Models, i.e. Gaussian Mixture Models, for speaker recognition. She put in evidence that the important characteristic of these models is the total number of mixtures and not the number of states.

These HMMs are thus unable to make use of temporal information for performing speaker recognition. On the other hand, recent experiments have shown that, for neural predictive systems, modelization of non stationarity allowed to significantly improve the performances [6].

We are interested here in the development of such models which will be refereed to as multi-state predictive neural networks (MSPNNs). We study the ability of these systems for speaker identification and discuss the superiority of multi-state upon one-state models. We provide results on 15 talkers from the TIMIT database.

1. INTRODUCTION

We analyse in this paper neural prediction systems for automatic speaker identification (A.S.I.). Speaker identification has been tackled by different non discriminant methods which have been developed and used successfully for speech recognition tasks. This is the case for example with Hidden Markov Models (HMM) [9,11,12,13], Vectorial Auto-regressive Models (VAM) [4,10], and more recently Predictive Neural Networks (NN) [1,2,5,6].

These methods are based on a modelization of the conditional class probability for each talker, and parameters estimation is usually performed independently for each model. The main benefit of these production systems is that they allow incremental modifications of the speaker database, which is an essential requirement for most ASI systems. However, speech utterance modelization being an indirect way to perform classification, this approach has intrinsic limitations [1].

The use of non stationary models allows to overcome the limitations of stationary (one-state) models and to increase the performances of neural predictive models ([6]) although this is not the case for HMMs as discussed in [9,15].

We present our neural predictive systems in §2, and put into evidence some of its inherent limitations in §3. We study in §4 the potential of non stationary neural predictive models, and propose a temporal alignment procedure for these multi-state models.

2. EXPERIMENTS AND DATABASE

The experiments reported in this paper have been performed on the international TIMIT database. We used the 15 female speakers from the first dialect. The 5 SX sentences were used for training and the 3 SI and two SA for testing. One sentence lasts approximately 3 seconds. This base being very clean and recorded in a single session, speaker identification is easy when the test utterance is long enough. In order to test our methods, we have used short segments from different lengths. Input data for the models are vectors resulting from a 16-order LPCC analysis, using 25,6 ms Hamming windows, with an overlap of 15,6 ms. The duration of an n-frame-length utterance is thus $\frac{n}{100}$ seconds. The neural nets used in the experiments reported here are Multi-Layer Perceptrons (MLP), with one hidden layer, trained to produce frame X_t, given a prediction context of the frame. In all the experiments we used the two preceding frames X_{t-1} and X_{t-2} as the prediction context for the frame X_t. We will note C_t the prediction context for frame X_t. Different temporal context could be used and may lead to increased performances. A discussion about this phenomenon can be found in [3].

3. STATIONARY MODELS

3.1. Adequacy problem

In an earlier paper [1], we described preliminary experiments showing that stationary predictive neural models (i.e. one state PNNs) were intrinsically limited for speaker recognition. By varying the size of the hidden layer, which can be viewed as a measure of the predictors complexity, we showed that models with a larger number of hidden cells improve prediction performances up to a certain extent but do not necessarily exhibit increased classification performances.

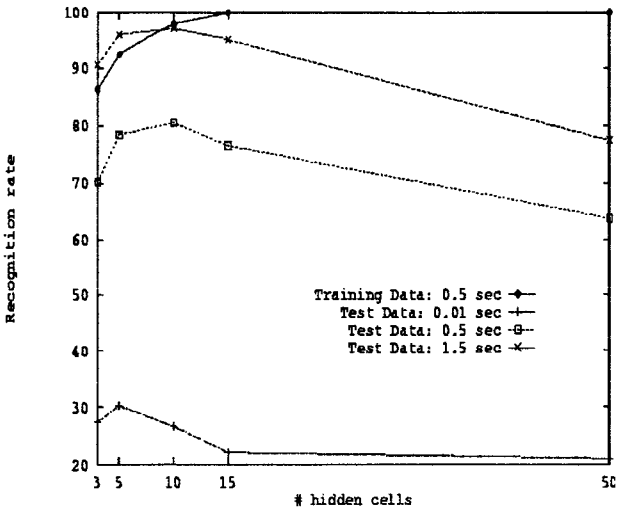

Figure 1 : Classification performances of stationary neural predictive models as a function of their complexity (i.e. # hidden cells). Performances are given on training data (0.5 sec. utterances) and on test data for 0.01, 0.5 and 1.5 sec. test durations.

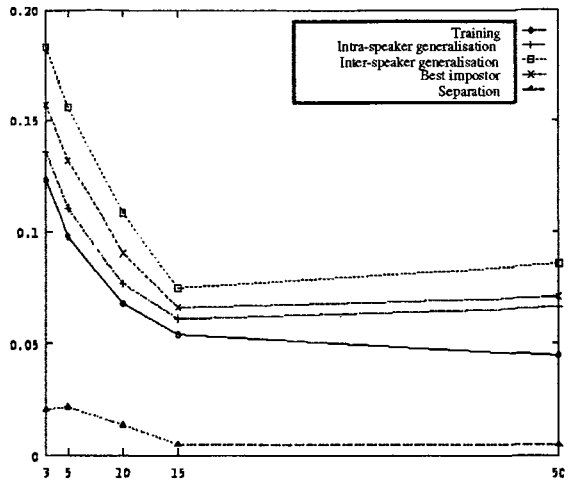

Figure 2 : The five curves represent the mean prediction error for a speaker predictor on his training data (Training), his test data (Intra-speaker generalisation), the test data of other speakers (Inter-speaker generalisation), the mean error of the best impostor's predictor (Best impostor), and the difference between "Best impostor" and "Intra-speaker generalisation" (Separation) corresponding to the "distance" between a speaker and the closest impostor.

Figures 1 and 2 clearly illustrate this phenomenon. The curves in Fig. 1 represent the correct identification rate of stationary PNNs on speech segments of various durations, on training and test data. Figure 2 shows the corresponding mean prediction errors. It can be seen on the two figures that the best models for classification (10 hidden cells) are different from the best models for prediction (15 hidden cells). The comparison of the two figures clearly shows that this is not an overtraining phenomenon (which arises here for 50 hidden cells predictors) but rather a consequence of the mismatch between the training and test criterion which are respectively the quality of prediction and the minimum classification error rate.

The above problem is expected to appear with any non discriminant algorithm, whatever the models are, neural networks or HMMs. This limitation is inherent to the modelization based approach since the training and test criteria would be equivalent if the speaker models were perfect, which is not the case in practice. However, it is expected that the mismatch will be reduced by using more "adequate" models.

3.2. Effect of test duration

It should be noted (from Fig. 1 and 2) that the best model depends on the test duration : 5 hidden cells predictors are the best for 0.01 sec. test utterances whereas 10 h.c. predictors are superior for longer test sequences. The Separation criterion in Fig. 2 measures the difference between the MSE of the talker and the closest competitor for each frame and is thus relevant for short speech classification performances.

This can be explained by the correlation between successive prediction errors of the predictors. When making simple hypothesis on these correlations, one can estimate analytically the probability of classification error in the simple case of a two speakers identification problem. Figure 3 shows the empirical and theoretical error rates for such an experiment.

This theoretical modelization complies with the experimental results. It is however difficult to work out similar modelizations for more than two speakers.

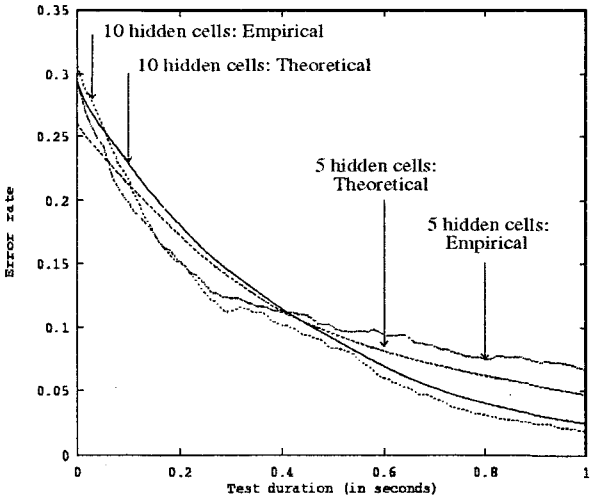

Figure 3 : Theoretical and empirical error rates for a two speakers identification problem, for PNNs with 5 and 10 hidden cells.

4. MULTI-STATE MODELS

MSPNNs allow to overcome some of the limitations of simple models described in §3.1 and offer improved

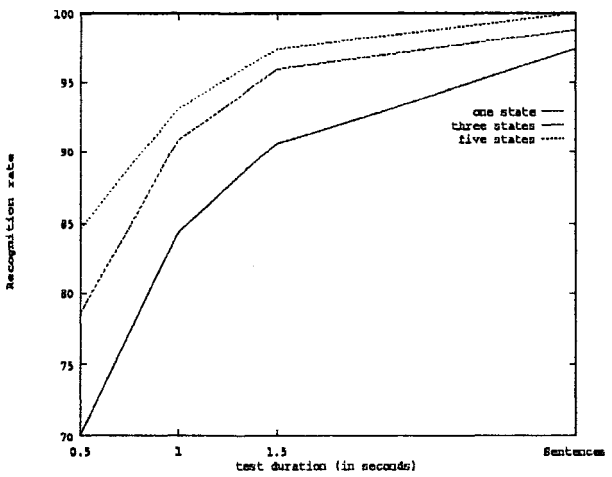

Figure 4 : Classification performances of N-states MSPNNs for different test speech lengths with N=1, 3, 5. Each state is implemented by a 3 hidden cells neural predictor.

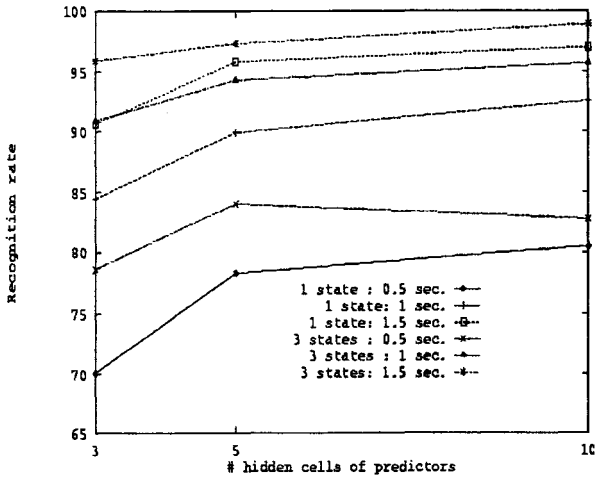

Figure 5 : Classification performances of 1 and 3-state MSPNNs for various predictor complexities.

performances over the best one-state models. These models allow to limit the averaging phenomenon observed with stationary models which are unable to modelize efficiently the speakers. For speaker recognition multi-state models are usually ergodic with only a few states (3 to 8) and we will used such models here. We discuss the training strategy in §4.1, and study the performances of MSPNNs as a function of the number of states (§4.2) and the predictors complexity (§4.3). We then introduce a temporal alignment procedure in §4.4.

4.1. Training strategy

There are two main approaches for training multi-state models. In the categorisation approach, the states are trained on short speech segments identified via an initial clustering or segmentation. Each predictor is then optimised separately on its own data and dedicated to the modeling of a broad phonetic class defined by the initial clustering. The justification for this is that similar speech segments (e.g. vowels) are close in the data space and then should be modeled with the same predictor. The free model approach, on the opposite, makes use of hidden states and alternates segmentation and optimisation. These two kinds of training have been used for MSPNNs in speech or speaker recognition tasks [2,6,7,8].

We have compared these strategies for 3-state models. In either case, we use a Viterbi-like procedure for training and recognition. The two strategies (constrained or free segmentation) give very similar classification performances although the behaviour of the models may be different. Categorisation models have significantly larger training prediction errors but generalise as well as free models. In addition, it is found that the states of a free model tend to automatically focus on broad phonetic classes (as reported in [11]) although they are less specialised than categorisation models. In the following, for simplicity, we use MSPNNs trained by categorisation.

4.2. Number of states.

We have performed experiments for analysing the influence of the number of states in MSPNNs. Figure 4 shows the behaviour of models with one, three and five states for a fixed predictor architecture while figure 5 compares 3-state models for various predictor complexities.

These two experiments clearly show that non stationarity is indeed useful in neural predictive systems. In Fig. 4, the 5-states models are superior to the 3-states models which are superior to the one-state models (i.e. stationary models). In Fig. 5, 3-state MSPNNs always outperform the corresponding (i.e. same number of hidden cells) stationary ones for any test speech duration. Due to the limited training data, we didn't explore more than five states for the models and 10 hidden units for the predictors. These results confirmed those obtained in [6].

4.3. Predictors complexity

We study here the influence of the complexity of the individual neural predictors. As said in §4.2, in the experiments reported in Fig. 4 and Fig. 5, MSPNNs always outperform the corresponding one-state models. However, the improvement obtained with the multiplication of states in a model is dependent both of the predictors complexity and of the prediction task. It can be observed for example that the improvement of 3-state MSPNNs on PNNs decreases with the complexity of the predictors used.

Furthermore, it is likely that the inadequacy phenomenon described previously in §3.1 with stationary models will appear with more complex MSPNNs. This means that there exists (independently of the overtraining phenomenon) for a fixed predictor architecture, an optimal number of states for classification. This optimal number of states of MSPNNs is larger for simple predictors (e.g. 3 hidden cells) than for complex predictors (10 hidden cells).

There is an optimal trade-off between these two model parameters.

4.4. Temporal alignment

It is now well known that transition probabilities of HMM-like non stationary models are ineffective for speaker identification, at least with ergodic models [9,13,15]. This ineffectiveness is a consequence of the difference of nature between emission probabilities (continuous density) and transition probabilities (a priori, discrete probabilities) in HMMs.

However, alignment methods could be useful for improving neural predictive multi-state models. Indeed, non discriminant training is one of the main cause of error in MSPNNs: a NN-state may abusively generalise even on data he has never seen, thus obtaining better performances than the correct model. In order to reduce this phenomenon, we suggest to implement the following decomposition of the likelihood:

$$P(X / C, \text{Speaker}) = \sum_{\text{State}} P(X / C, \text{State}, \text{Speaker})\, P(\text{State} / C, \text{Speaker})$$

where X is a frame and C the prediction context (i.e. preceding frames). The likelihoods P(X/C,State,Speaker) are implemented by the neural predictors and the posterior probabilities P(State/C,Speaker) are computed via a neural classifier, one for each speaker, trained to learn the segmentation of the MSPNN on his training data. This decomposition is particularly interesting with categorisation models because posterior probabilities are easier to estimate. We used a simplified Viterbi-like algorithm to compute the probability of a sequence $X_1^T = (X_1, ..., X_T)$ with a speaker model. This is performed in two steps:

1. Compute the state sequence $(q_1, ..., q_T)$ using the alignment module where q_t denotes the state at t.

2. Compute the likelihood of the sentence by:

$$P(X_1^T) = \prod_t P(X_t/C_t, q_t)$$

This temporal alignment procedure is close to multi-expert architectures recently developed for prediction tasks [14].

We used here 3-state MSPNNs with 5 hidden units for neural predictors. The alignment modules are MLPs with 10 hidden units (with 32 input units corresponding to the prediction context). The performances of alignment modules are measured by their ability to predict, given a prediction context, the predictor of the MSPNN leading to the minimal prediction error. These rates are approximately of 85%, 76%, and 66% on training data, intra-speaker generalisation data and inter-speaker generalisation data. From these results, it is clear that the proposed algorithm will penalise the models in inter-speaker generalisation more than in intra-speaker generalisation.

We report preliminary results using this additional technique. The tables 1 and 2 show prediction and classification performances of MSPNNs with or without time alignment procedure. The improvement in recognition given by the alignment procedure is small for short segments but increases with the test duration (Tab. 2). Furthermore, the separation criteria is also increased, meaning a greater discrimination between the speaker models (Tab. 1). This separation criteria, which was also shown in Fig. 2 and discussed in §3.2 is a indicator of the effectiveness of the models to classify. This method could be further improved by optimising simultaneously the neural predictors and the state-classifier of a speaker model.

Temporal alignment	Intra-speaker	Inter-speaker	Separation
No	**0,093**	**0,136**	0,022
Yes	0,102	0,152	**0,026**

Table 1: Mean Prediction Errors of 3-state MSPNNs with or without temporal alignment. The predictors have 5 hidden units. Best results are in bold.

Temporal alignment	0.5	1	1,5	Phrases
No	84,1	94,3	97,3	98,7
Yes	**84,2**	**94,8**	**98,3**	**100,0**

Table 2: Identification rate for various test durations (in seconds) for 3-state MSPNNs with or without temporal alignment. The predictors have 5 hidden units. Best results are in bold.

5. CONCLUSIONS

We have presented multi-state neural predictive models as a way to overcome limitations of simple stationary models. We have shown that this modeling technique is a way to build effective models for classification. Due to the inadequacy problem, it is not easy nor immediate to build effective models for classification. Other techniques have been proposed in [3]. We have also proposed an additional alignment procedure which seems to be a promising way to handle the temporal nature of the signal when the standard HMMs transition probabilities have been shown to be ineffective [9,13].

REFERENCES

[1] Artières T., Gallinari P., 93 : neural models for extracting speaker characteristics in speech modelization systems, Eurospeech.
[2] Artières T., Gallinari P., 94 : adequacy of neural predictors for speaker identification, WCNN.
[3] Artières T., 95 : predictive systems for speaker identification: heuristics for model selection, submitted to ICANN.
[4] Bimbot F., Mathan L., Lima A., Chollet G., 92 : standard and target driven AR-vector models for speech analysis and speaker recognition, ICASSP.
[5] Hassanein K., Deng L., Elmasry M., 94 : a neural predictive hidden Markov model for speaker verification, ESCA Workshop, Martigny.
[6] Hattori H., 92 : text independent speaker recognition using neural networks, ICASSP, II 153-156.
[7] Iso K., Watanabe T., 90 : speaker-independent word recognition using a neural prediction model, ICASSP.
[8] Levin E. 93 : hidden control neural architecture modeling of non linear time varying systems and its applications, IEEE Trans. on NN, Vol. 4.
[9] Matsui T., Furui S., 92 : comparison of text-independent speaker recognition methods using VQ-distorsion and discrete/continuous HMMs, ICASSP 157-160.
[10] Montacié C., Le Floch J.L., 92 : AR-vector models for free-text speaker recognition, ICSLP.
[11] Poritz A.B., 82 : linear predictive HMMs and the speech signal, ICASSP, Vol. 2, 1291-1294.
[12] Savic M., Gupta S.K., 90 : variable parameter speaker verification system based on hidden Markov modeling, ICASSP, 281-284.
[13] Tishby N., 91 : on the application of mixture AR HMMs to text-independent speaker recognition, IEEE Trans. on Signal Processing, Vol. 39, N° 3, March 91.
[14] Tresp V., Tanigushi M., 95 : combining estimators using non constant weighting functions, NIPS 7.
[15] Zhu 94 X., Gao S.R., Chen F., 94 : text-independent speaker recognition using VQ, mixture gaussian VQ and ergodic HMMs, ESCA Workshop, Martigny, Switzerland.