PHONEME RECOGNITION WITH STATISTICAL MODELING OF THE PREDICTION ERROR OF NEURAL NETWORKS

F. Freitag, E. Monte

Universitat Politècnica de Catalunya
Department of Signal Theory and Communications
Barcelona, Spain

felix@gps.tsc.upc.es

ABSTRACT

This paper presents a speech recognition system which incorporates predictive neural networks. The neural networks are used to predict observation vectors of speech. The prediction error vectors are modeled on the state level by Gaussian densities, which provide the local similarity measure for the Viterbi algorithm during recognition. The system is evaluated on a continuous speech phoneme recognition task. Compared with a HMM reference system, the proposed system obtained better results in the speech recognition experiments.

1. INTRODUCTION

Most of today's speech recognition systems are based on Hidden Markov Models (HMMs). Given a speech pattern in terms of an observation vector sequence, the HMMs can represent local stationary segments of speech by means of states and can model the variability in duration of the speech patterns through self-transitions and transitions between states. The standard HMM, however, does not model the dependence between the observation vectors in a stationary segment, but rather considers that these observation vectors are generated randomly according to the state output probability distribution.

Recently, hybrid speech recognition systems have been proposed where a certain sub-task of the speech recognition system is carried out by neural networks (for instance [1]). Among the hybrid systems for speech recognition, there are the speech recognition systems which are based on predictive neural networks, such as proposed in [2][3][4][5].

In our approach based on predictive neural networks, the observation vector sequence of speech is assumed to be generated by a dynamic system, where a causal relation between neighboring observation vectors exists. We model the function describing this dependence by predictive neural networks. Further, we assume that the function provides class-specific information which can be useful for speech recognition. In our system we obtain this information by the prediction error measure and incorporate it in the Viterbi algorithm for speech recognition.

2. SYSTEM IMPLEMENTATION

In the speech recognition system which we propose we use HMM based phoneme models where each phoneme of the database is represented by a model consisting of one or more modeling states. The phoneme model has a non-emitting entry and exit state and modeling states in between. Different to the standard HMM, however, each modeling state incorporates one multilayer perceptron neural network of five hidden neurons, which is used to predict the current observation vector given one past observation vector, so that we can write

$$\hat{c}_i(t) = f(\mathbf{o}_{t-1}), \qquad (1)$$

where f is the nonlinear function described by the neural network, and $\mathbf{o}_{t-1}$ is the past observation vector.

We denote $\hat{c}_i(t)$the *ith* predicted observation vector coefficient and $c_i(t)$ the *ith* coefficient of the observation vector of size N at time t. Then, we can introduce the prediction error vector $\mathbf{d} = [d_1, d_2, \ldots, d_N]$, which is composed of the coefficients

$$d_i = c_i - \hat{c}_i \,. \qquad (2)$$

On the hypothesis of a Gaussian distributed parameter $\boldsymbol{d}$, we can model its distribution by a multivariate Gaussian density $N(\mathbf{d}; \mu, \Sigma)$, so that our speech recognition system consists of the parameter set $\lambda = (\mathbf{W}, \mathbf{B}, \mathbf{A})$, where $\boldsymbol{W}$ indicates the weights of the neural network, $\boldsymbol{B}$ the means and variances of the Gaussian distribution of the parameter $\boldsymbol{d}$, and $\boldsymbol{A}$ the state transition probabilities.

In most recognition systems based on predictve neural networks the prediction error of the neural networks is incorporated in the Viterbi search in form of the Euclidean distance [2][4][5], where each prediction error coefficient obtains the same weight in the distance measure. The Euclidean distance measure, however, does not take into account the different variance of the coefficients. For vectors with cepstrum and delta cepstrum parameters, for instance, the use of Gaussian modeling of the prediction error coefficients can be advantageous since the normalization to the variance of the coefficient increases the weight of the delta cepstrum parameters, which have a small dynamic range compared to the cepstrums.

This work has been supported by the TIC95-1022-C05-03

For the training of our system the predictive neural networks have to be trained (parameters ***W***) and the parameter values of the HMM have to be estimated, which are the values of the Gaussian densities and the state transition probabilities (parameters ***B***, ***A***). The training steps are the following: 1) Training of the predictive neural networks with the observation vector sequence assigned to each state 2) Estimation of the means and variances of the Gaussian densities for the new parameter ***d*** for each state and estimation of the values of the state transition probabilities.

The training of our system was started with a phonetically segmented bootstrap database. After this initial training, a larger training database was used for which a rough phonetic segmentation existed. The system was trained with a small number of training epochs on this database, then the recognition result on the validation database was taken. The validation database was used to control the performance of the system on unseen data. After obtaining the result on the validation database, the system re-segmented the training database and the described process started again. When the best result on the validation database was obtained, then the recognition rate of the system on the test database was taken.

Different to the reference HMM system where training and testing was carried out in a sequential way, in the proposed system with predictive neural networks the training and testing was done in the above described parallel way in order to avoid overtraining of the system on the training database.

Speech recognition was performed with the Viterbi algorithm, which determined the most likely state sequence *S* with

$$\hat{P}(O|M) = \max_{S}\{a_{S(0)S(1)} \prod_{t=1}^{T} b_{S(t)}(\mathbf{d}_t) a_{S(t)S(t+1)}\} \qquad (3)$$

where $b(\mathbf{d}_t)$ is the state output probability for the new speech parameter vector ***d***. In the proposed system a single continuous Gaussian density was used for each state to model the distribution of the parameter ***d***.

In comparison to the HMM reference system, where the HMMs modeled an observation sequence consisting of mel-cepstrum based observation vectors $\mathbf{o}_t$, the vector sequence for the proposed system consisted of a sequence of prediction error vectors obtained by means of the predictive neural networks. Therefore, in the proposed system the state output probability is computed by $b(\mathbf{d}_t)$, where the reference system computed $b(\mathbf{o}_t)$.

The performance of the proposed system was compared with that of the reference HMM system. This HMM system also used a single continuous Gaussian density per state and was trained with the Baum-Welch algorithm on the training database. Although the same test database was used both for the reference HMM system and the proposed system, the reference system had a slightly increased training database due to the fact that for the reference system no validation database was needed.

Both the proposed system and the reference system used a very simple grammar network for recognition which only required the beginning and end of the recognized phoneme sequence to be silence while any other combination of succeeding recognized phonemes within the sequence was allowed.

3. DATABASE AND RECOGNITION TASK

The bootstrap database was a small phonetically segmented database of Spanish speech with a total of 2259 phoneme samples, with which the initial training of the models was carried out. Then, for training and testing of the recognition system the Spanish EUROM1 database was used. The training database contained 29738 training phoneme samples, the validation database had 7236 phonemes and the test database consisted of 12928 phonemes. The databases were labeled into 26 different phonemes.

The speech data of the databases was clean continuous speech sampled at 16 kHz. We used speech frames of 25 ms and a frame shift of 10 ms. The speech data was Hamming windowed and pre-emphasized. Speech was parametrized into 12 liftered mel-frequency cepstrums with delta parameters, providing an observation vector of 24 coefficients for each speech frame.

The task of the speech recognition system was continuous speech phoneme recognition. Further, the test database was speaker- and text-independent of the training database.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section we report the speech recognition results obtained with the proposed system which used the local distance measure based on the new parameter vector ***d***. First, we present some preliminary experiments which illustrate some statistic characteristics of the predicted observation vectors and the prediction error measure. In a second part we evaluate our system in a continuous speech recognition task and compare the results with the reference system.

In the preliminary experiment our objective is to compare the variances of the coefficients c_i, $\hat{c}_i$, and d_i, where c_i is the *ith* coefficient of the mel-cepstrum based observation vector, c_i is the *ith* coefficient of the predicted observation vector, and d_i is the *ith* coefficient of the prediction error vector ***d*** as given in equation (2). For the preliminary experiment a speech recognition system with five states phoneme models was trained as described in section 2. After training the variances of the coefficients c_i, $\hat{c}_i$, and d_i, were computed. In Figure 1 the variances of the coefficients for the third state of the model for the phoneme D are shown.

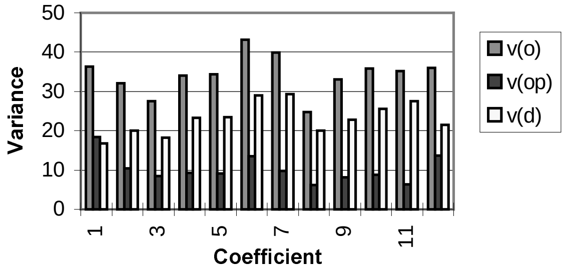

Figure 1: Variances of 12 coefficients of the observation vector of the third state of the models D. The denominations are: v(o) = variance of coefficients of the observation vector; v(op) = variance of coefficients of predicted observation vector; v(d) = variance of coefficients d_i.

In Figure 1 it can be observed that generally $\sigma_{\hat{c}_i} < \sigma_{d_i} < \sigma_{c_i}$. The fact that $\sigma_{d_i} < \sigma_{c_i}$ indicates that the predicted coefficient $\hat{c}_i$ is correlated with the coefficient c_i of the observation vector, so that $d_i = c_i - \hat{c}_i$ has a smaller variance than c_i. Also, it can be seen that the dynamic range of the predicted coefficient $\hat{c}_i$ is smaller than that of the coefficient c_i of the observation vector, since $\sigma_{\hat{c}_i} < \sigma_{c_i}$. From the preliminary experiment it can be seen that $\hat{c}_i(t)$ generally approximates the coefficient $c_i(t)$ since $\sigma_{d_i} < \sigma_{c_i}$.

In order to evaluate the proposed system, we performed the speech experiments with phoneme models of three, four, and five states, where in an experiment the same size of phoneme model was used for all the phonemes. The modeling of the prediction error vectors on the state level was made by a single continuous Gaussian density per state. The phoneme models allowed self-transitions and transitions to the next state.

The system was trained first on the bootstrap database and then training was carried out on the EUROM1 training database. The training process of the system was controlled by the results on the validation database. When the best performance on the validation database was obtained, then the recognition results on the test database were taken.

In order to compare the results of the proposed recognition system, we carried out speech recognition experiments with the HMM reference system. The reference system modeled the sequence of observation vectors consisting of the mel-cepstrum vectors and delta parameters.

In the Figures 2 and 3 the recognition results are presented for recognition systems using three, four, and five state phoneme models.

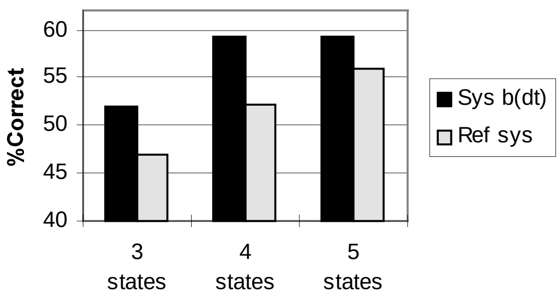

Figure 2: Recognition results on the test database in %Correct of the proposed system based on predictive neural networks (Sys b(dt)) and the reference system.

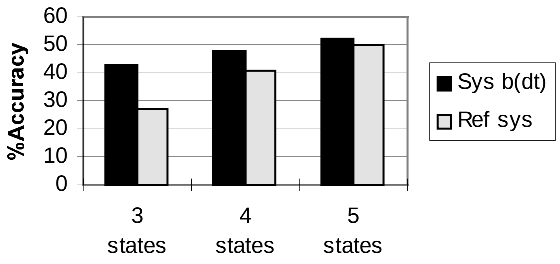

Figure 3: Recognition result on the test database in %Accuracy of the proposed system based on predictive neural networks (Sys b(dt)) and the reference system.

In Figure 2 and Figure 3 the speech recognition results are presented in %Correct and %Accuracy, respectively. Considering the %Correct and %Accuracy results, it can be seen that the proposed recognition system increased the recognition rates with the increase of the number of states. This observation is also made for the reference system, which also increased the performance with the increase of the number of states. Secondly, it can be observed that the proposed system outperformed the reference system in the three, four and five states phoneme models, both in the %Correct and %Accuracy measure.

From an architectural point of view the proposed system is different from the reference system in the sense that its phoneme models based on HMMs incorporate additionally predictive neural networks in each of the modeling states. These neural networks perform a state-specific nonlinear feature transformation of the mel-cepstrum vector sequence into a sequence of prediction error vectors. Therefore, different to the reference system where the HMMs model the original mel-cepstrum sequence, the HMM based phoneme models of the proposed system model a state specific vector sequences of prediction error vectors.

The recognition results of the experiments indicate that the incorporation of the predictive neural networks in the system improved the recognition rates. In what concerns the state transitions and the number of continuous Gaussian densities per state, the classifier of both the proposed system and the reference system is based on the same HMM structure. Therefore, the experimental results indicate that using the new features obtained by the nonlinear feature transformation with the neural networks provides a better performance.

It is known that the performance of a speech recognition system is influenced by the interaction of many of its parameters. Therefore, it is difficult to identify in which sense the new observation vectors consisting of the prediction error vectors influenced on the performance of the recognition system. Nevertheless, we shall point out to some aspects of the system, where the new parameters could have influenced:

1. Since both systems used the same HMM architecture, it could have occurred that a more accurate modeling of the density of the prediction error vectors was possible with the single continuous Gaussian density than it was for the density of the mel-cepstrum vectors. Let us assume that an observation vector is obtained in function of a past observation vector and some Gaussian noise component such that $o_t = f(\mathbf{o}_{t-1}) + \varepsilon$. If the predictive neural network after training encodes well the deterministic function, then the prediction error vector as given in equation (2) may consist mainly of the random component which the neural net cannot predict. If we further assume that this prediction error has a single-modal Gaussian density, then one Gaussian density per state which we used in the proposed system could have been an accurate model for the probability density of such vectors. Differently, it is known that if for the mel-cepstrum vectors a mixture of Gaussian densities per state is used then the recognition results improve.

2. The HMM was a more suitable model for the observation vector sequence consisting of the prediction error vectors than it was for the mel-cepstrum vectors. The standard HMM assumes that no correlation exists between observation vectors of a stationary segment and that these observation vectors are emitted randomly by the HMM state according to the state output probability density. However, one can assume that there is some correlation between neighboring mel-cepstrum observation vectors, since it allowed the prediction by the neural network. Due to the fact that the HMM does not model the correlation of observation vectors, it can be argued that this model is not completely accurate to represent well the process which generates such an observation vector sequence. As outlined above, compared to the mel-cepstrum observation vector sequence, the prediction error vectors could consist rather of random components, for which the correlation between observation vectors in the vector sequence is less. Then, for such an observation vector sequence the HMM was a more appropriate model. Due to the better fit of the HMM to the vector sequence of prediction error vectors, the recognition rate of proposed system has improved compared to the reference system.

5. CONCLUSIONS

The proposed system based on predictive neural networks outperformed for all tested phoneme models of 3, 4, and 5 states both in %Correct and %Accuracy the reference continuous density HMM system (Figure 2 and Figure 3). Our results showed 1) the incorporation of predictive neural networks in a speech recognition system; 2) that with predictive neural networks we improved the recognition rates compared to a reference continuous density HMM system, which points out that the features derived from the predictive neural networks can be useful for speech recognition.

REFERENCES

1. Bourlard, H., Morgan, N. “Continuous Speech Recognition by Connectionist Statistical Methods”, *IEEE Trans. on Neural Networks, vol. 4, no. 6, pp. 893-909, November 1993.*

2. Tebelskis, J., Waibel, A., Petek, B., Schmidbauer O., “Continuous speech recognition using Linked Predictive Neural Networks”. *Proc. ICASSP 91, pp. 61-64, 1991.*

3. Iso, K., Watanabe, T., “Large vocabulary speech recognition using neural predictive model”, *Proc. ICASSP 91, pp. 57-60, 1991.*

4. Freitag, F., Monte, E., Salavedra, J. “Predictive neural networks applied to phoneme recognition”, *Proc. EUROSPEECH, pp. 2831-2834, Rhodes, September 1997.*

5. Levin, E. “Hidden Control Neural Network Architecture Modeling of Nonlinear Time Varying Systems and Its Applications”, IEEE Trans. On Neural Networks, vol. 4, no. 1, January 1993.