Internal Use Only

TR-I-0233

Text-Independent Speaker Recognition Using Neural Networks

Hiroaki HATTORI

1991.12

Abstract

This paper describes a text-independent speaker recognition method using predictive neural networks. The speech production process is regarded as a non-linear process so the speaker individuality in the speech signal also includes non-linearity. Therefore, the predictive neural network, which is a non-linear prediction model based on multi-layer perceptrons, is expected to be a more suitable model for representing speaker individuality. For text-independent speaker recognition, an ergodic model which allows transitions to any other state, including self-transitions, is adopted as the speaker model and one predictive neural network is assigned to each state. The proposed method was compared to distortion based methods, HMM based methods, and a discriminative neural network based method through a text-independent speaker recognition experiments on 24 female speakers. The proposed method gave the highest recognition accuracy of 100.0%, and the effectiveness of predictive neural networks for representing speaker individuality was clarified.

©ATR Interpreting Telephony Research Laboratories
©ATR 自動翻訳研究所

1 Introduction

The speaker recognition/identification technique is one of the important techniques in speech processing. It can be directly applied to security checks, and also to speaker selection or speaker clustering so as to improve the performance of speech recognition systems. There are two types of speaker recognition methods: text-dependent method and text-independent method. Text-dependent speaker recognition uses phoneme context information so that high recognition accuracy is easily achieved. On the other hand, text-independent speaker recognition has an advantage that it does not require specially designed utterances. Therefore, it is easier to build a user friendly system. This paper discusses text-independent speaker recognition.

The VQ based speaker recognition method is one of the well-known text-independent speaker recognition methods and has reported its high performance[1]. It uses the static information of speaker individuality in a short term spectrum. However, speaker individuality contains not only static, but also dynamic features. Therefore, a model which represents both static and dynamic features of spectra is required for higher performance. Some approaches for acquiring dynamics features of speaker individuality can be found in papers[2]-[4].

The predictive neural network is a non-linear prediction model based on multi-layer perceptrons[5]. It non-linearly predicts the next frame from the preceding several frames. The speech production process is regarded as a non-linear process so the speaker individuality in the speech signal also includes non-linearity. Therefore, a predictive neural network is expected to be a more suitable model for representing speaker individuality. In this paper, we propose a new speaker model based on predictive neural networks. This model is an ergodic model which allows transitions to any other state, and one predictive neural network is assigned to each state.

In the following section, the proposed method and other speaker recognition methods such as distortion based methods, HMM based methods, and a discriminative neural network based method are described and the performance of these methods on text-independent speaker recognition is reported.

2 Speaker Recognition Algorithms

2.1 Distortion Based Method

In VQ based speaker recognition method, a codebook is designed for each speaker from training data using LBG algorithm. Input speech is quantized using each speaker's codebook and the input speaker is recognized as the speaker whose codebook gives the minimum distortion. Matrix quantization (MQ) is a straightforward extension of the VQ method to handle the dynamic features in a vector sequence.

2.2 HMM Based Method

A statistical approach such as the hidden Markov model (HMM) has shown its high performance for modeling of speech[6]. There have been some reports about the use of this model for both speaker recognition[7] and speaker adaptation[2],[3].

In text-dependent speaker recognition, a left-to-right model can be used because the phoneme sequence of an input sentence is predetermined. However, in text-independent speaker recognition, it is difficult to know the phoneme sequence beforehand, so the ergodic model, which allows transitions to any other state, is adopted. We evaluate two types of HMMs, a discrete model and a continuous model. For the discrete model, the output probability of a speaker-independent codebook is given to each transition between states. For the continuous model, the diagonal Gaussian mixture density is given to the transition.

One model is trained for each speaker using the forward-backward algorithm. During recognition, the forward probability of input speech is calculated for each speaker model. An input speaker is recognized as the speaker whose model gives the maximum probability.

2.3 Discriminative Neural Network

Discriminative neural networks have been used for various classification problems, including speaker recognition[8]. The discriminative neural network for speaker recognition is a commonly used 3-layer network whose output units correspond to speakers. The structure of a discriminative neural network is shown in Fig. 1.

This network is trained using all reference speakers's training samples using the back-propagation algorithm. During recognition, input speech is given to the network frame by frame and the values of the output units are accumulated. An input speaker is recognized as the speaker who corresponds to the unit with the maximum output value.

2.4 Predictive Neural Network

Predictive neural networks have been originally proposed for speech recognition[5]. The structure of a predictive neural network is shown in Fig. 2 . This network non-linearly predicts the next frame from the preceding two frames. This network is trained to minimize the prediction error using the back propagation algorithm. We use two types of speaker models which are shown in Fig. 3. One predictive neural network is assigned to each state. In the case of the 1-state model, only one predictive neural network is used for the prediction of entire samples, hence the prediction error is large. To decrease the prediction error, vector variations which are predicted by one neural network should be small. Therefore, in the case of the 4-state model, training samples are divided into four groups using clustering technique, which correspond to states of the model, and samples in each group are used to train the corresponding predictive neural network. During recognition, input speech is time-aligned with the model using the Viterbi algorithm to get the minimum prediction error. An input speaker is recognized as the speaker whose model gives the minimum prediction error.

3 Database

The TIMIT database is used for the evaluation experiments. The TIMIT database contains more than 600 speakers and ten sentences for each speaker. The sentences consist of two dialect calibration sentences(SA), three random contextual variant sentences(SI), and five phonetically compact sentences(SX). Only SA sentences are the same for all the speakers. We use SX sentences as training data and SI sentences as test data. Therefore, no sentences are the same in training and test data.

3.1 Analysis

The data was analyzed every 5ms using the auditory model proposed by Seneff[9]. The model has two outputs: the synchrony output and the mean rate response. It has been reported that these outputs are complementary and the combination of these outputs gives a better result than they do separately[10]. However, we have only used the mean rate response as the feature vector because of the computational cost. An analyzed utterance is normalized so as to have a mean of zero for each frame and a maximum value of 1.0 in one sentence according to the following equations,

$$s_t^k = o_t^k - \frac{1}{K}\sum_k o_t^k,\ x_t^k = \frac{s_t^k}{max|s_t^k|},$$

where, o_t^k is the original mean rate output , s_t^k is a mean-zeroed vector, and x_t^k is the normalized vector. The superscript k denotes the kth channel's output and the subscript t denotes the tth frame. This normalization is carried out for all experiments.

3.2 Speaker Selection

The use of all the speakers in the TIMIT database requires an enormous amount of CPU time. Therefore, to carry out evaluation experiments efficiently, a speaker set which has similar features is selected, using a speaker clustering technique based on VQ distortion. The algorithm is as follows:

1. Regard each speaker as one cluster.
2. Calculate the distance between two clusters for all combinations and then merge the closest two clusters. The distance between clusters is calculated as follows:
   $D_{cc}(A,B) = \frac{1}{N_A N_B}\sum_{a\in A}\sum_{b\in B} distortion(a,b)$,
   $distortion(a,b) = \frac{1}{2}\left(dist(a|b) + dist(b|a)\right)$,
   where, N_A and N_B are the numbers of speakers in the cluster A and B, and $dist(a|b)$ is the VQ distortion when speaker a's utterance is quantized by speaker b's codebook.
3. Calculate the distance between the speaker and the cluster for all combinations and move the speaker to the closest cluster. The distance between speaker a and cluster B is calculated as follows:
   $D_{pc}(a,B) = \frac{1}{N_B}\sum_{b\in B} distortion(a,b)$.
4. Repeat 3 until there is no movement.
5. If the number of clusters is bigger than two, go to 2.

This algorithm was applied to 177 female speakers in the TIMIT database, and finally 24 speakers in Table 1 were selected. The histogram of the distortion between two speakers, $distortion(a,b)$, in the selected speakers and all speakers are shown in Fig. 4. The average and variance of the distortion in the selected speakers were 0.270 and 0.021, while in all speakers they are 0.345 and 0.058, respectively.

4 Evaluation Experiments

4.1 Distortion Based Method

The speaker recognition accuracies obtained using the VQ based method are shown in Fig. 5. The accuracies are shown as functions of input sentence length, and the right-most points show the accuracies obtained using all the test sentences. It can be seen that there were no significant difference in the performance of difference size codebooks and 64 codewords seems sufficient for a codebook. The highest recognition accuracy of 91.7% was obtained using 64-codeword codebook.

For the MQ based method, three successive frames were treated as a matrix, i.e. the length of matrix was 15ms. The results of the MQ based method are shown in Fig. 6. These results shows that the number of codewords does not affect the performance. In comparison to Figure 5, it can be seen that the use of a matrix does not improve performance. The highest recognition accuracy of 87.5% was obtained using 256 codeword codebook.

Next, matrices with different lengths were evaluated. To prevent an increase in the number of parameters, the first ,middle , and last frames of a segment were used to make a matrix. Therefore, all matrices had the same number of parameters, i.e. 120, and the size of codebooks was fixed at 256 codewords. The results are shown in Fig. 7. This figure shows that performance decreases as the length of the matrix increases. This means that the amount of training data was not enough to create a matrix codebook. While, a longer matrix can represent dynamic features in a longer time period, it requires a lot more training data to cover the phoneme context variations. In this experiment, the training data of five sentences was not enough to design a matrix codebook and performance decreased as a result.

4.2 HMM Based Method

The 4-state model is adopted as a speaker model. The output probabilities of transitions to the same state are tied. Therefore, there are 4 independent output probabilities instead of 16. For the discrete model, the output probability of a speaker-independent codebook was given to each transition. For the continuous model, the diagonal Gaussian mixture density was used as the probability density function.

The speaker recognition accuracies obtained using discrete HMM and continuous HMM are shown in Fig. 8. The dashed and solid lines show the accuracies of discrete HMM and

continuous HMM, respectively. The accuracies of discrete HMM increased as the size of codebook increased. However, the performance of the 4,096-codeword model was still less than that of the VQ based method. This was due to the lack of distortion information. In the case of discrete HMM, an input utterance is transformed into a label sequence, and the distortion information is discarded. The highest accuracy of 79.2% is achieved by the 4,096-codeword model.

In the case of continuous HMM, a model with 16 mixture densities performed better than a model with 32 mixture densities. This may suggest an insufficiency of training data. The increase in the number of mixtures decreases the reliability of parameter estimation when the training data is limited. The highest speaker recognition accuracy of 91.7% is achieved using the 16-mixture model.

Comparing these results, the continuous HMM performed better than the discrete HMM. This is because the continuous HMM is essentially free of the VQ distortion.

4.3 Discriminative Neural Network

The number of input units and output units were 120 and 24 respectively, and the number of hidden units ranged from 48 to 120. Sigmoid functions were used for the hidden units and output units. 2000 samples were randomly selected for each person every iteration. Center initialization[11] was used for the initialization of the network, and gain and momentum were determined according to the results of preliminary experiments. The number of iterations and the initial values were selected to give the best recognition rate for the test data.

The results of the discriminative neural network are shown in Fig. 9. The dashed lines show recognition accuracies for the training data and the solid lines show accuracies for the test data. The best accuracy of 95.7% is achieved using the network with 72 hidden units.

4.4 Predictive Neural Network

The number of input units and output units were 80 and 40 respectively, and the number of hidden units ranged from 5 to 40. Sigmoid functions were used for the hidden units and output units. 2000 samples were randomly selected every iteration. Center initialization[11] was used for the 1-state model and the trained 1-state model was used as the initial model for the 4-state model. Gain and momentum were determined according to the results of preliminary experiments. The results of the 1-state model are shown in

Fig. 10. The dashed lines show recognition accuracies for the training data and the solid lines show accuracies for the test data. It can be seen that recognition accuracies for training data increased as the number of hidden units increased. However, accuracies for the test data increases until 10 hidden units had been used and then decreased as the number of hidden units increased. This probably means that a network with many hidden units tends to over-learn the features of the training data. The best result for the test data was achieved when 10 hidden units were used, with an accuracy of 95.7%.

Next, the effectiveness of the non-linearity was evaluated. The performance of a model with sigmoid functions and a model with linear functions are shown in Fig. 11. The number of hidden units was fixed at 10. The performance of a model with sigmoid functions was better than that of a model with linear functions for both training data and test data. This clearly shows that the non-linearity of the predictive neural network is essential to capture speaker individuality. The recognition accuracies of the 4-state models are shown in Fig. 12. The accuracy of any 4-state model was greater compared to that of a 1-state model which had the same number of hidden units. Recognition accuracies of 100.0% are obtained for the models with 5 hidden units and 10 hidden units.

Next, the connections between the bias unit and the output units were investigated. The connections weights and the average vector of samples used to train the predictive neural network are shown in Fig. 13. It can be seen that the connection weights were quite similar to the average vectors; it is assumed that the static features of speech were mainly captured by the connections between the bias unit and the output units, and that the dynamic features were captured by the other connections.

5 Discussion

The best result for each method is shown in Fig. 14. The discrete HMM shows the lowest performance due to the lack of distortion information. There is no significant difference between the VQ based method, and the MQ based method, and the continuous HMM based method. This is because the amount of training data used in this experiments was not enough for the MQ based method and the continuous HMM method. However, in spite of the limited training data, the proposed method performed quite well. The best accuracy of 100.0% is achieved by the proposed model. This result shows that non-linear modeling is effective in representing speaker individuality in speech.

Though the proposed method performed well, its performance varies depending on the number of training iterations and the initial values, and it is difficult to determine suitable

values beforehand. This is still a problem for both the predictive neural network and the discriminative neural network.

6 Summary

This paper proposed a text-independent speaker recognition method using predictive neural networks. The proposed method was compared to distortion based methods, HMM based methods, and a discriminative neural network on text-independent speaker recognition and the proposed method gave the best recognition accuracy of 100.0%. The results clarified the effectiveness of using predictive neural networks for representing speaker individuality.

Acknowledgments

This work was carried out while the author stayed at the Laboratory for Computer Science at MIT. The author would like to thank Dr. Akira Kurematsu and Dr. Victor Zue for their support of this work. He would also thank members of the Spoken Language Systems group, particularly Hong Lueng, James Glass, and Michel Phillips, for their discussion and software support.

References

[1] F.K. Soong et al., "A Vector Quantization Approach to Speaker Recognition", Proc. ICASSP85, pp.387-390, 1985.

[2] T. Imamura, "Speaker-Adaptive HMM Based Speech Recognition With A Stochastic Speaker Classifier", Proc. ICASSP91, pp.841-844, 1991.

[3] H. Hattori, "Speaker Adaptation Based On Markov Modeling Of Speakers In Speaker-Independent Speech Recognition", Proc. ICASSP91, pp.845-848, 1991.

[4] S. Nakamura et al., "A Neural Speaker Model for Speaker Clustering", Proc. ICASSP91, pp.853-858, 1991.

[5] K. Iso et al., "Speaker-Independent Word Recognition Using A Neural Prediction Model", Proc. ICASSP90, pp.441-440, 1990

[6] K-F. Lee et al., "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", Proc. ICASSP85, pp.387-390, 1985.

[7] A. E. Rosenberg, et al., "Connected Word Talker Verification Using Whole Word Hidden Markov Models", Proc. ICASSP91, pp.381-384, 1990

[8] L. Rudasi et al., "Text-Independent Talker Identification with Neural Networks", Proc. ICASSP91, pp.389-392, 1991

[9] S. Seneff et al., "A Joint Synchrony/Mean-Rate Model Of Auditory Speech Processing", Proc. J. of Phonetics, vol. 16, pp55-76, 1988

[10] H.M. Meng et al., "Study of Classification of Vowels Using Multi-Layer Perceptrons", Proc. ICSLP90, pp.343-347, 1990

[11] H.C. Leung et al., "Phonetic Classification Using Multi-Layer Perceptrons", Proc. ICASSP90, pp.525-528, 1990

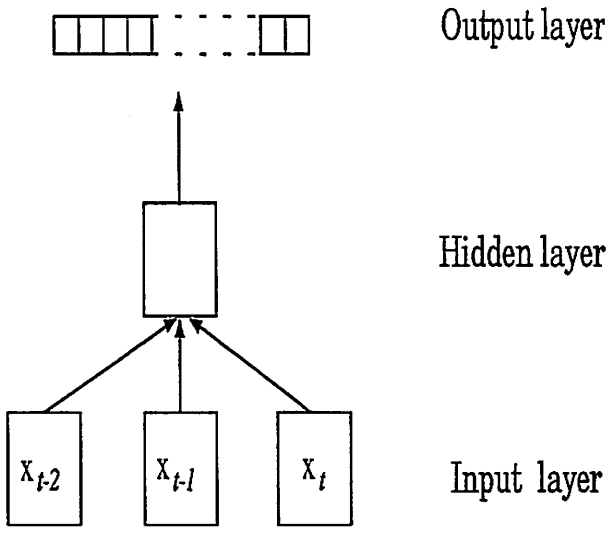
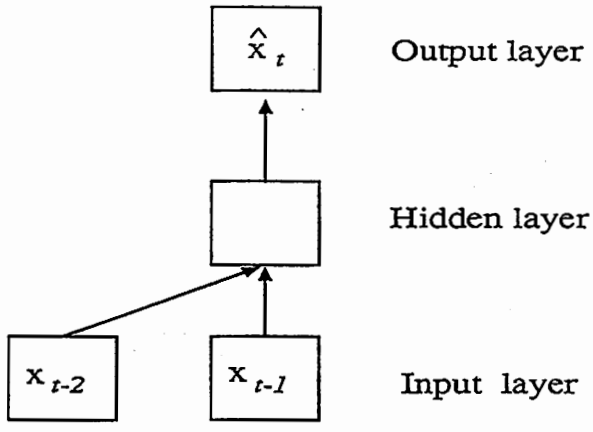
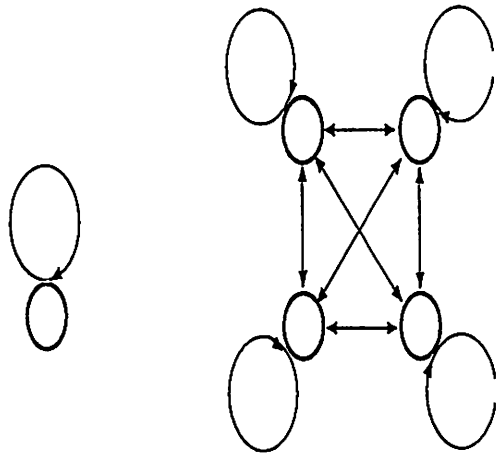


Fig. 1: A structure of a discriminative neural network



$$\text{prediction error} = \| x_t - \hat{x}_t \|^2$$

Fig. 2: A structure of a predictive neural network



a) 1-state model b) 4-state model

Fig. 3: Structures of speaker models

Table 1: The list of selected speakers

fvmh0-1	fmah1-7	fjwb0-2	fjsp0-1	fpab1-7	fbch0-6
fjlg0-3	fmju0-6	fnkl0-8	fdaw0-1	fear0-5	fbjl0-5
fpad0-6	fjre0-2	fkfb0-1	fjas0-2	fjsj0-8	fcmh1-8
fpas0-2	flkm0-4	fcrh0-4	fgmb0-5	feeh0-4	fsag0-5

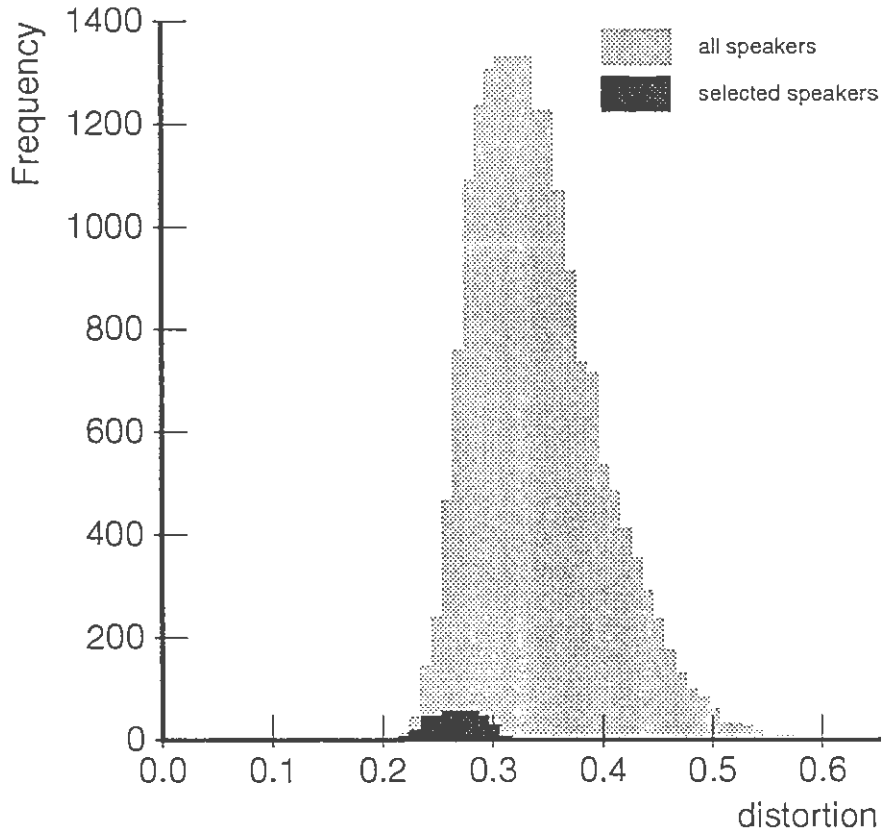


Fig. 4: Histogram of distortion between speakers

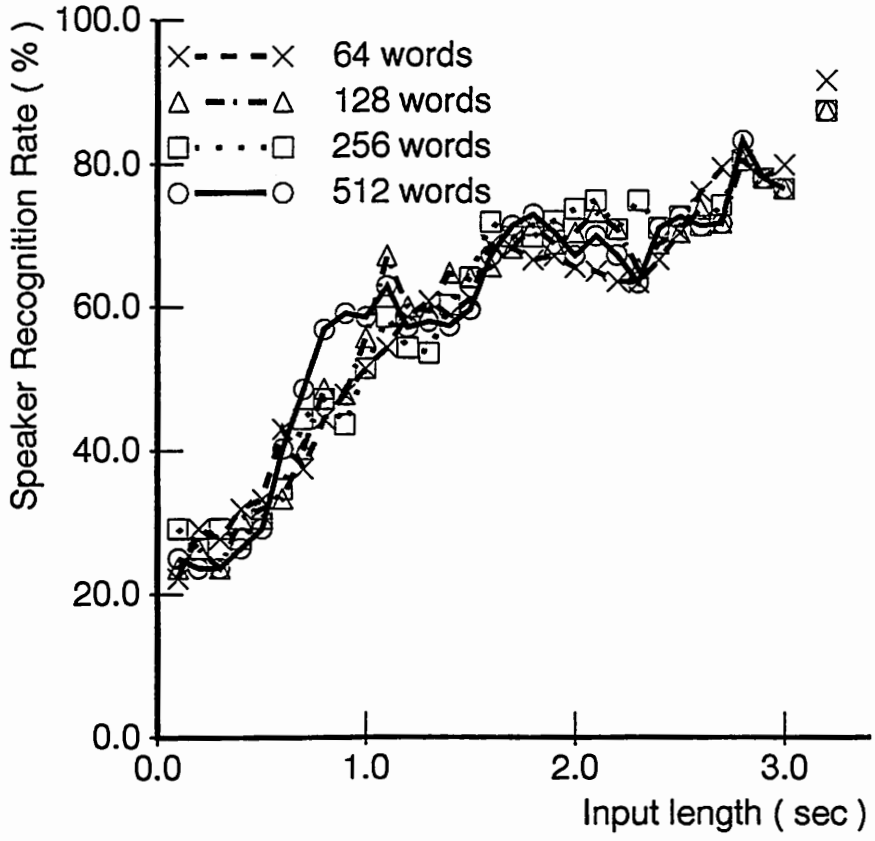


Fig. 5: Speaker recognition rates of the VQ based method

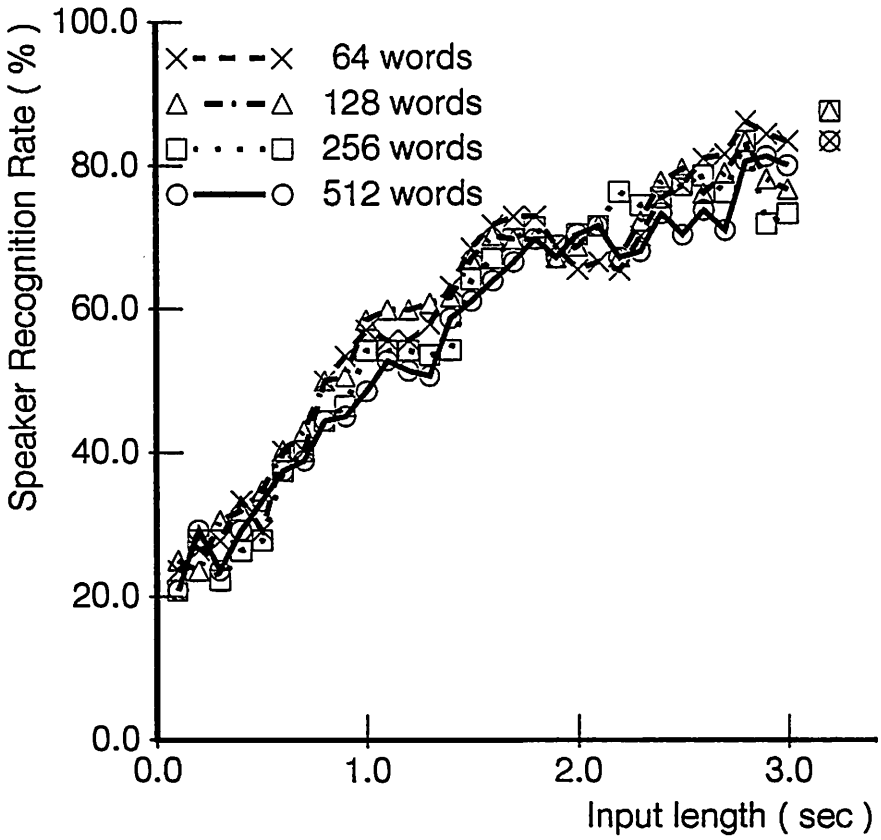


Fig. 6: Speaker recognition rates of the MQ based method

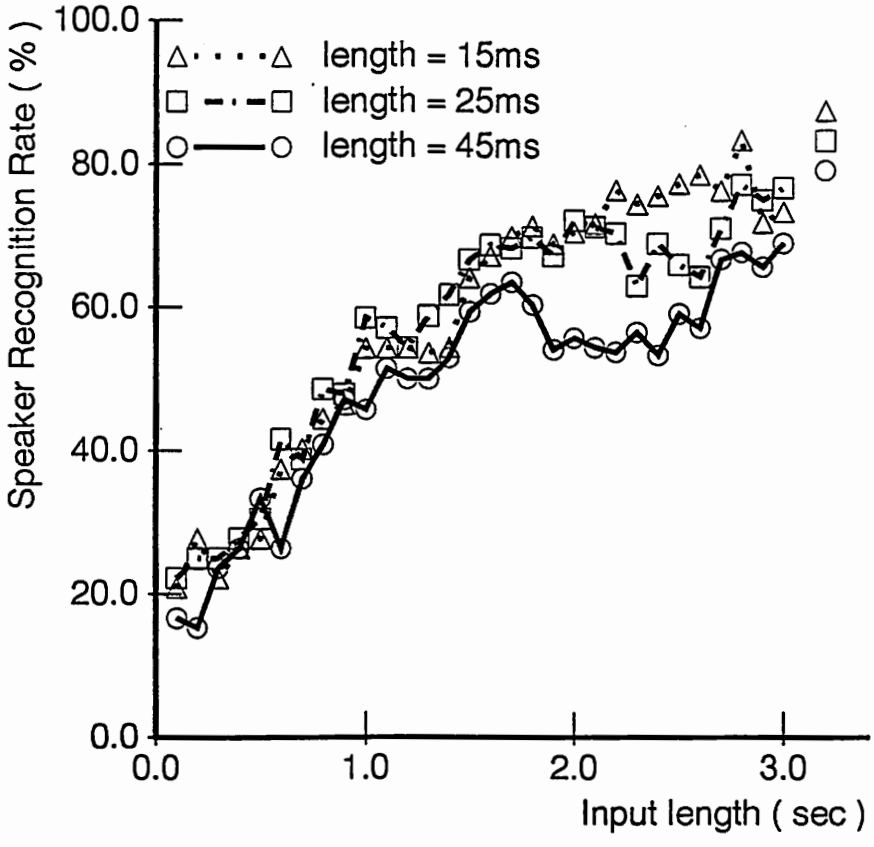


Fig. 7: Speaker recognition rates of the MQ based method with different matrix length

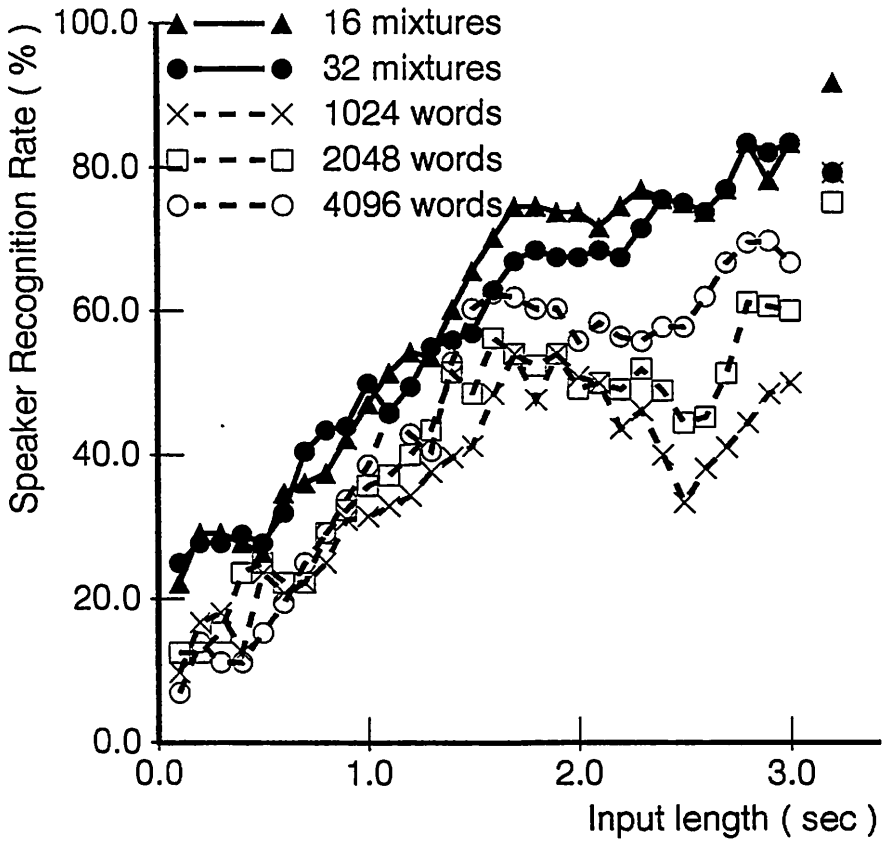


Fig. 8: Speaker recognition rates of the HMM based methods

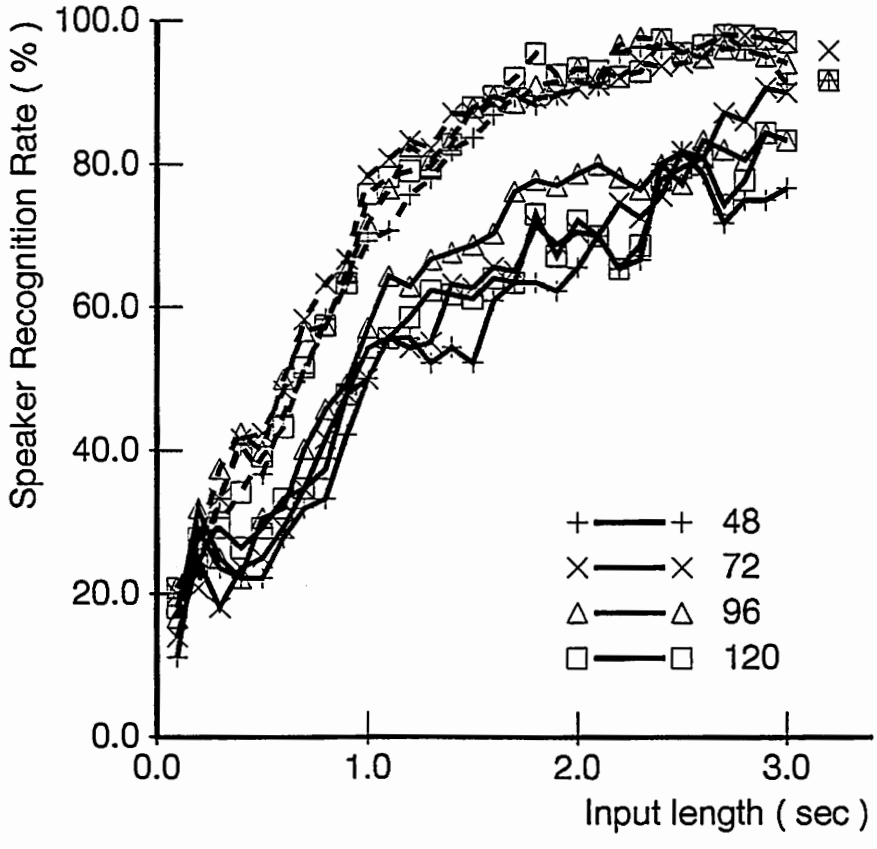


Fig. 9: Speaker recognition rates of the discriminative neural network

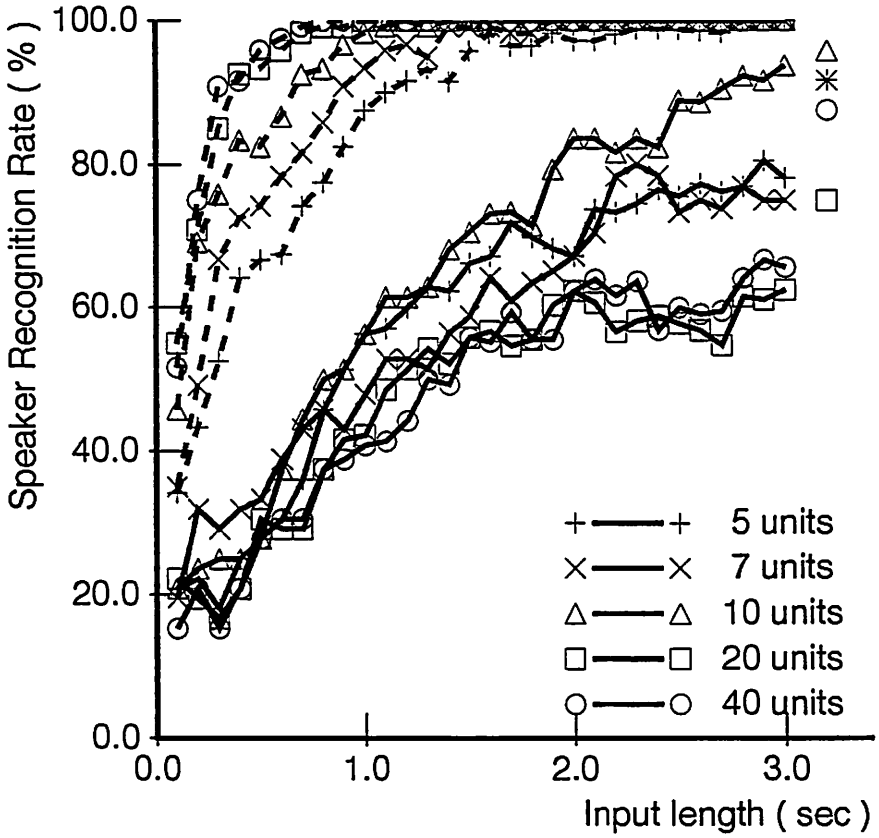


Fig. 10: Speaker recognition rates of the 1-state model

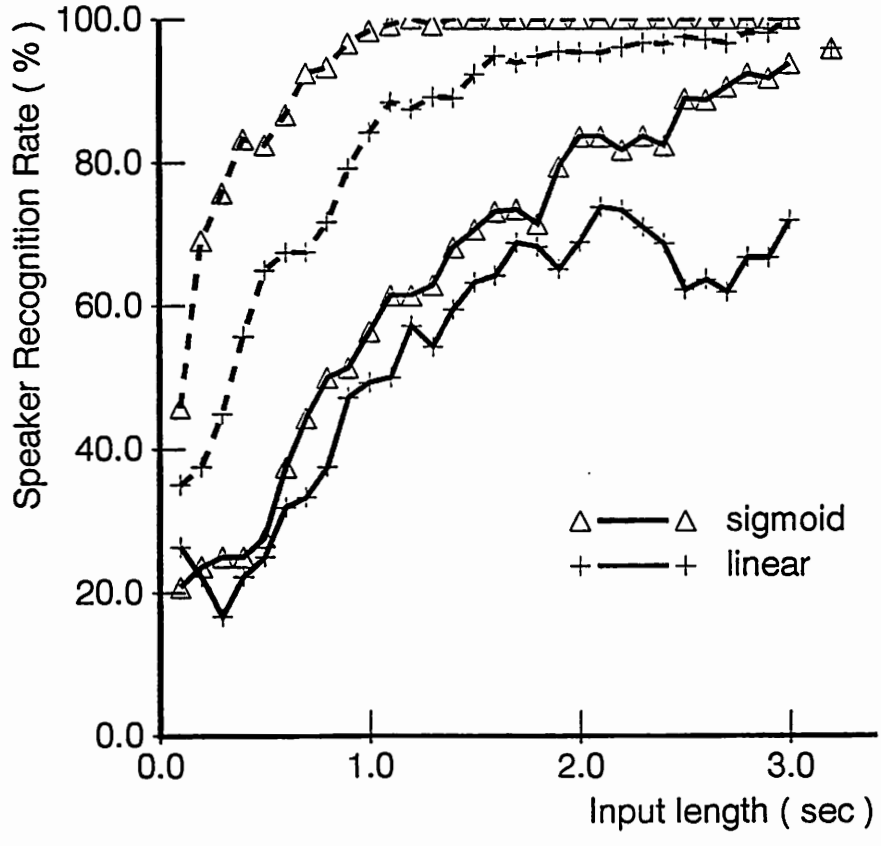


Fig. 11: The effect of sigmoid functions

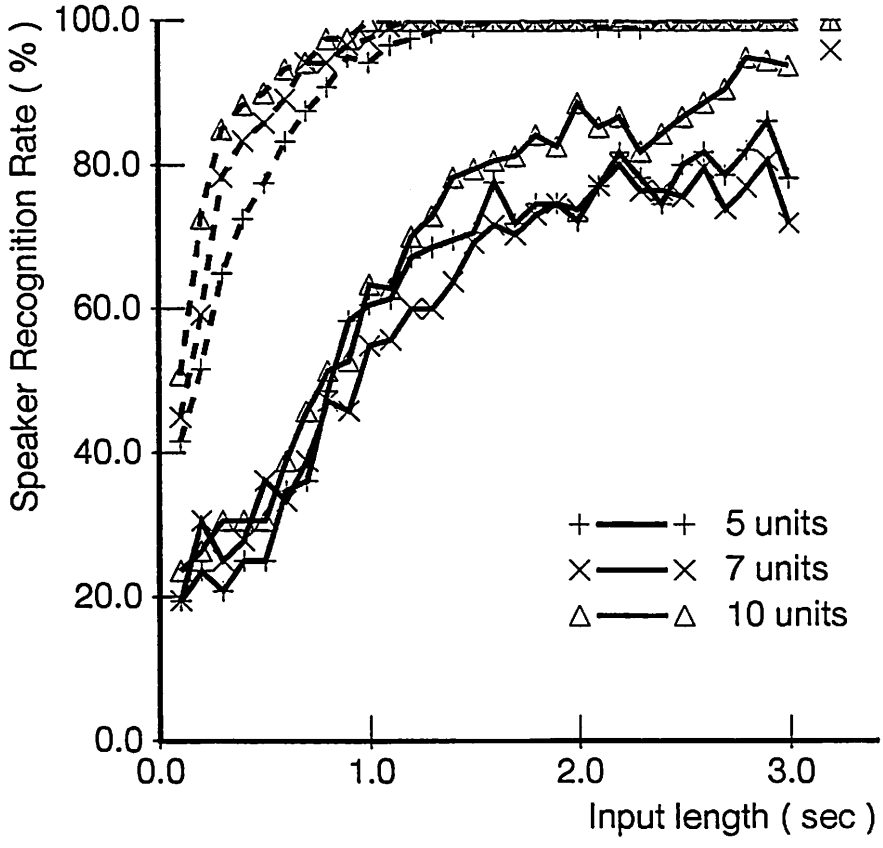
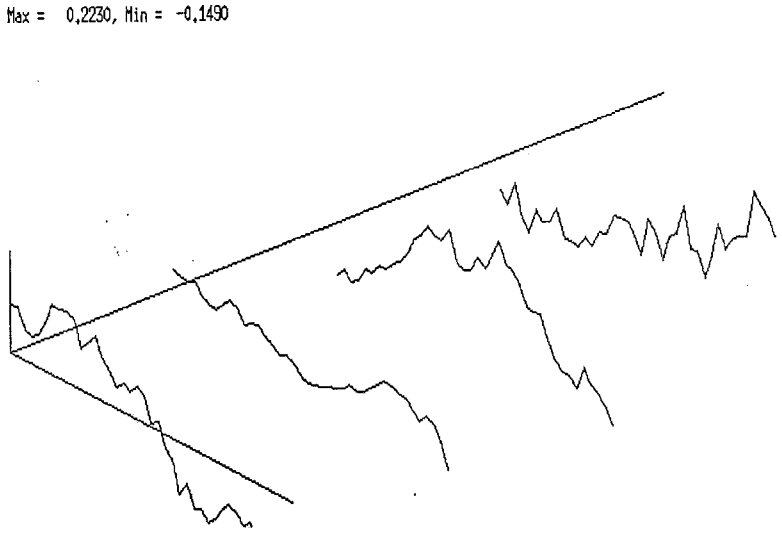
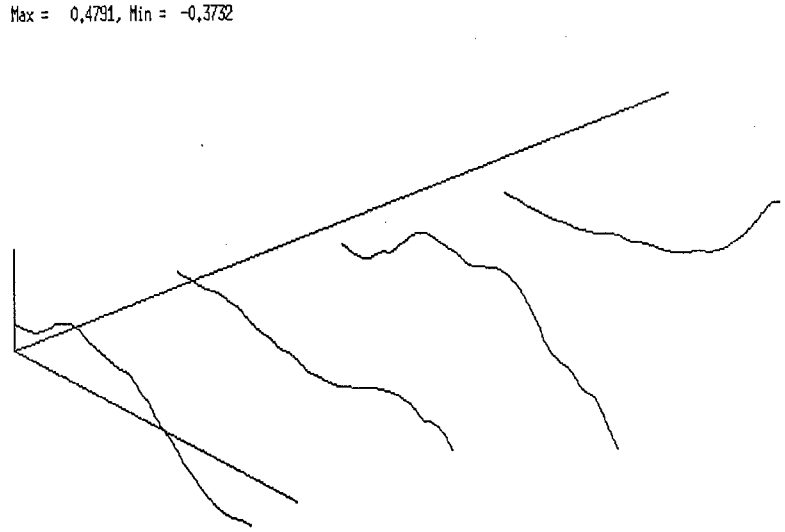


Fig. 12: Speaker recognition rates of the 4-state model



a) connections between the bias unit and the output units



b) average vectors of training samples

Fig. 13: Connections between the bias unit and the output units

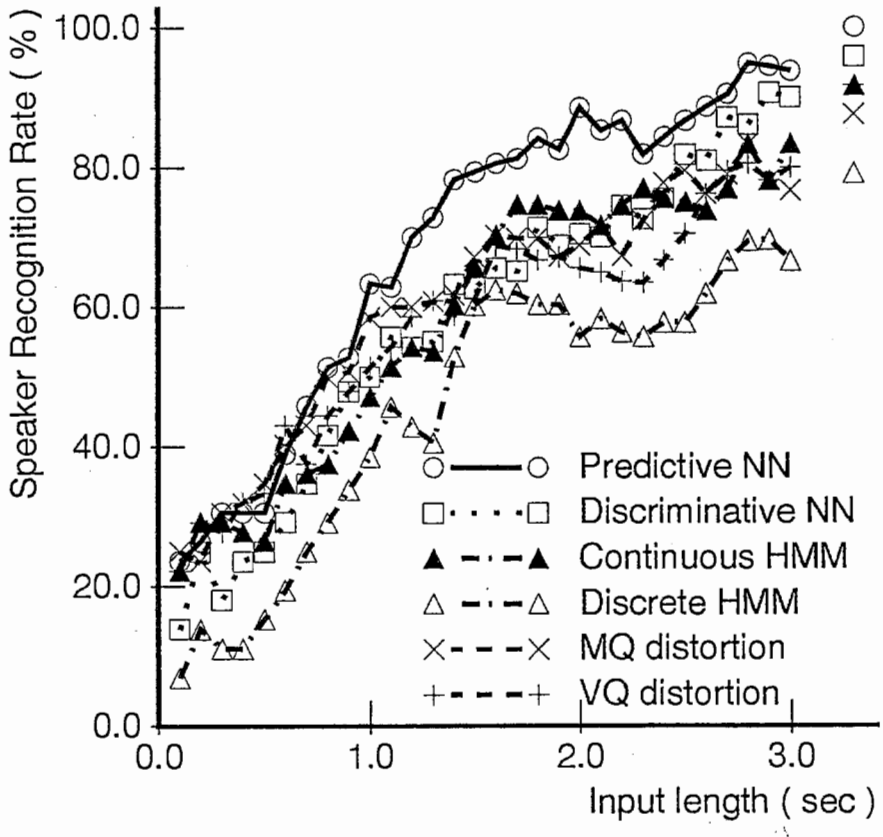


Fig. 14: Best results of the each method