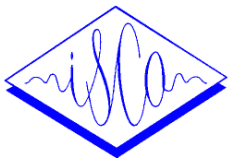

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SPEAKER RECOGNITION USING STATIC AND DYNAMIC CEPSTRAL FEATURE BY A LEARNING NEURAL NETWORK

Hujun Yin
Dept. of Electrical Engineering, Tongji University
1239 Sipin Rd., Shanghai 200092, P.R.China

Tong Zhou
Computer Laboritory, Shanghai TV University
527 Ming Xing Rd., Shanghai 200433, P.R.China

ABSTRACT

We have applied a Multi-layer learning neural network to speaker recognition. A set of LPC-based cepstral coefficients and their orthogonal polynomials were chosen as static and dynamic spectral feature of speaker's utterance. A number of experiments have been conducted to assess the performance of the network of both text-dependent and text-independent speaker recognition tasks. The results describe the relations between the recognition accuracy and the number of hidden units, the number of training set presentations, and the number of speaker database. In a critical condition, the network achieved a high recognition rate of 98.5 percent correct and 95.2 percent correct in text-dependent and text-independent tests respectively.

1. INTRODUCTION

Speaker recognition problem has long been amony the most interesting and challenging areas in speech research. In the past years many researches on this problem aiming at improving the recognition rate of automatic systems for speaker indentification and speaker verification have been reported [1]-[6]. In these researches much work was addressed on spectral feature extraction, reference template or VQ codebook construction, and matching scores or distance computation. The spectral information is usually extracted through LPC-based cepstral coefficients analysis. To characterize the nonlinearity and time-varibility of speaker's vocal tract, more and more researches have dealt with not only static (or instant) spectral information, but also dynamic (or transient) spectral information [3]-[5]. Orthogonal polynomial coefficient varying in time can be utilized to represent spectral variation in time. To achieve a high performance, many frames segmented from duration of speaker's utterance are needed to construct the reference template or codebook, therefor the number of frames and the size of codebook increases rapidly when speaker's database increases. However, by a learning neural network, it has been shown that this demand can be much reduced at the same recognition rate.

In recent years, neural network has stirred great interest in pattern recognition areas [7]-[9]. The inherent parallelism of the networks can allow very rapid parallel search and the best-match computation, and reduce much of the computation of match scores or distances, which is always needed in conventional nearest neighbor pattern recognition. The essence of computation of neural network is nonlinear logical operation and can produce emergently and spontaneously collective or convergent effects. The nonlinearity of the input-output relationship can yield nonlinear classification in state space more effectionaly. The more attractive reason for applying neural network is that they are trainable and adaptive and can provide more insights into nature of human interagence.

This paper addresses the application of a Multi-layer neural network to speaker recognition. A method of utilizing static and dynamic spectral information by the network is presented. Experiments on both text-dependent and text-independent recognition tasks have been conducted to asses that the network's recognition rates are affected by the number of hidden units and the number of training set presentations, or training circles. The performance of of the network has been discussed.

2. LAYERED NEURAL NETWORK

2.1 Architecture

The neural network used for the speaker's recognition experiments addressed below was composed of three layers of neurons as shown in Fig.1. The neurons of input layer are sensor neurons to detect the input signals. Neuron in output layer or hidden layer sums all its inputs and yields an output with a nonlinearity. The

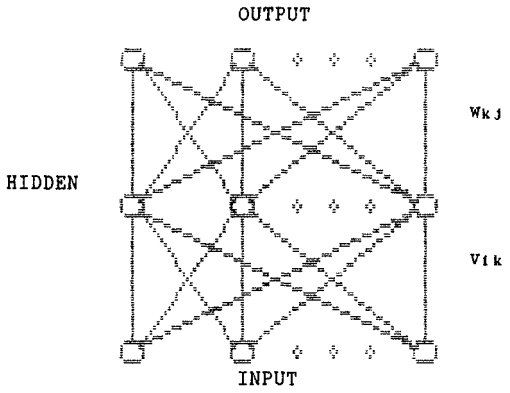

Fig.1 Structure of Three Layers Neural Network with Input x_i, Output y_j and Hidden z_k units

formulas are

$$y_j = f\left(\sum_{k=1}^{N_2} w_{kj} z_k - \Phi_j\right) \qquad 1 \le j \le M \qquad (1)$$

$$z_k = f\left(\sum_{i=1}^{N_1} v_{ik} x_i - \Theta_k\right) \qquad 1 \le k \le N_2 \qquad (2)$$

Where y_j and z_k are the output of neurons in ouput layer and hidden layer, Φ_j and Θ_k are the bias of these neurons, w_{kj} and v_{jk} are the connection weights between hidden and output layers and between input and hidden layers respectively, x_i are input signals, f() is a nonlinear function. A sigmoidal nonlinearity was used in our experiments, it is:

$$f(\alpha) = \frac{1}{1 + e^{-\alpha}} \qquad (3)$$

The number of input neurons (N_1) and the number of output eurons (M) depend on the dimension of feature patterns and the number of classes respectively. The number of hidden neurons (N_2) is usually selected empirically.

2.2 Network Learning Algorithm

The back-propagation learning algorithm has been found to perform well and can find good solutions in most recognition cases [7]. It is an iterative training process designed to minimize the mean square error between the actual output and the desired output. The error signal is backpropagated, and is used to modify connection weights. Weights w_{kj} are adjusted according to

$$w_{kj}(t+1) = w_{kj}(t) + \eta \delta_j z_k \qquad (4)$$

Where η is a gain term, δ_j is an error term for output neuron j, and is defined by

$$\delta_j = y_j(1-y_j)(d_j - y_j) \qquad (5)$$

Where d_j is the desired output, and y_j is actual output.

Weights v_{ik} are adjusted similarly by

$$v_{ik}(t+1) = v_{ik}(t) + \eta \delta'_k x_i + \alpha(w_{kj}(t+1) - w_{kj}(t)) \qquad (6)$$

Where $0<\alpha<1$, and error δ'_k is defined by

$$\delta'_k = z_k(1-z_k)\sum_j \delta_j w_{kj} \qquad (7)$$

Training sets will be presented to the input of the network until the error is reduced below a preseted limit.

3. SPECTRAL FEATURE

3.1 Static Cepstral Coefficient

Among several different spectral representation, Atal[2] found that LPC-based cepstral coefficients were the best for speaker recognition. These coefficients are refered to static spectral features, which can be obtained by minimizing the mean square prediction error between a speech sample and its linearly predicted value from the past p samples. p is the order of the LPC coefficients. Durbin's recursion is an effective method to solve for LPC coefficients. The process is

$$E(0) = R(0) \qquad (8)$$

$$a_i(i) = [R(i) - \sum_{j=1}^{i-1} a_j(i-1)R(i-j)]/E(i-1) \qquad 1 \le i \le p \qquad (9)$$

$$a_j(i) = a_j(i-1) - a_i(i)a_{i-j}(i-p) \qquad 1 \le j \le i-1 \qquad (10)$$

$$E(i) = [1-(a_i(i))^2]E(i-1) \qquad (11)$$

$$i = 1, 2, \ldots\ldots, p$$

$$a_j = a_j(p) \qquad (12)$$

Where a_j is the jth LPC coefficient. R(i) is autocorrelation function of speech samples. A rectangular or Hamming window is usually used in above calculation.

Then cepstral coefficients can be obtained by [1][2]

$$c_1 = a_1 \qquad (14)$$

$$c_n = \sum_{k=1}^{n-1}(1-k/n)a_n c_{n-k} + a_n \qquad 1 \le n \le p \qquad (15)$$

3.2 Dynamic Cepstral Feature

To represent cepstral variation in time, an orthogonal polynomial is used. A 1st-order polynomial or the generalized spectral slope, denoted $c'_m(t)$, is usually sufficient to characterize cepstral variation. $c'_m(t)$ can be approched by[5]

$$c'_m(t) = \frac{\sum_{k=-H/2}^{H/2} c_m(t+k)k}{\sum_{k=-H/2}^{H/2} k^2} \qquad (16)$$

Where H is the length of the duration for generalizing $c'_m(t)$.

4. EXPERIMENTS AND RESULTS

4.1 Text-Dependent Experiments

We used 20 adult speaker (10 male and 10 female) as test population in experiments. Each speaker spoke a designated Chinese word 15 times at different time. We randomly chose 10 of 15 utterances of each speaker as training sets, and the others as testing sets. There were totally 200 training sets and 100 tesing sets. Each utterance was band-limited from 100 Hz to 3.5KHz and sampled at 8 KHz. The digitized signals were then blocked into 40 ms frames every 10 ms. One frame had 321 samples. In conventional speaker recognition techniques, according to Euclidan or Mahalanobis distance measurement, an average distance was computed by averaging the distance from each frame to achieve the best recognition rate [2]. So the feature vector was actually composed of cepstral coefficients over all frames, i.e

$$\mathbf{V}=[v^T(1),v^T(2),.....,v^T(N)]^T \qquad (17)$$

Where

$$v(n)=[c_1(n),c_2(n),...,c_p(n)]^T \qquad (18)$$

,and N is the number of frames of an utterance.

In our experiments, we selected two particular frames instead. One was located at the energy maximum of the utterance, one was delayed a definited time. We extracted the first nine cepstral coefficients of these two frames as static cepstral feature. And we used (16), in which H was set to 3, to obtain a **v'** located at center of the two frames. Then we constracted a new vector by

$$\mathbf{V}=[v^T(n1),[v(n2)-\tfrac{1}{2}(n1-n2)v'(\tfrac{1}{2}(n1+n2))]^T]^T \qquad (19)$$

Where

$$v'(n)=[c'_1(n),c'_2(n),...,c'_p(n)]^T \qquad (20)$$

Each 18-dimensional vector from each training utterance was used to train a network with 18 input units, 20 output units. To determine an average performance. Ten trails with different division of training sets and testing sets were taken.

Several overall average indentification rates versus the number of training circles are shown in Fig.2.

4.2 Text-Independent Experiments

We used 10 speakers (5 male and 5 female). Each speaker spoke 10 different short (4-7 seconds) Chinese sentences in different time. Half of them were chosen randomly as training sets, half as testing sets. Each sentence utterance was divided into many energy segments according to that there was a local energy maximum in every segments. Then we took one of the first five segments to extract a feature vector using (19). There were 250 training feature vectors, 25/per-speaker. The network was with 18 input units and 10 output units.

When the network had been trained, a testing sentence utterance was forwared for recognizing. the utterance was also divided into many energy segments. Acculmulating outputs of each output neurons responding to the first five vectors derived from the first five segments, we chose one with the most times of the highest output. If the result was not clear, more segments would be forwared until the dicision was made. An average recognition performance (based on 10 trails) is shown in Fig.3.

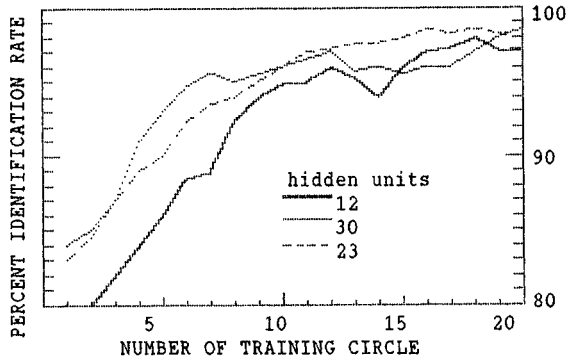

Fig.2 Indentification Rates versus The Number of Training Circles in Text-Dependent Speaker Recognition Test

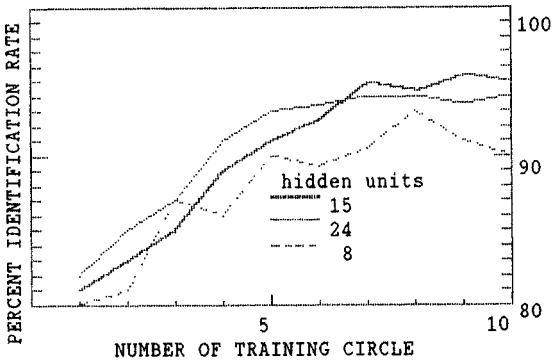

Fig.3 Indentification Rates versus The Number of Training Circles in Text-Independent Speaker Recognition Test

With the increase of the number of hidden units, the recognition rate is generilly improved, as the arbitrariness of decision regions is enhanced. And the training time also decreases, that is the less training sets are re-forwared.

In our another experiment, more frames

of an utterance were used to extract cepstral coefficients. Using (17) we constructed a long feature vector to represent the utterance. With ten utterance recorded at different time each speaker, A series of cepstral coefficients represented well with their variation in time. But more input units and connection weights were needed, and it took a very long time for our computer to simulate training process of the networks. The results are shown in Table 1.

Table 1. Test with Many Frames Used

	Text Dependent	Text Independent
Test Population	20	10
Training Utter.	10/per-speaker	10/per-speaker
Frames used	22	50
Input Unit	176	400
Hidden Unit	40	60
Connec. Weight	7,200	24,600
Training Circle	5	5
Average Recog. Rate	98.8%	97.9%

5. DISCUSSION

The experiment results of speaker recognition based on layered neural networks have shown that the network is very promising on speech research areas and can proform as accurately as nearest neighbor classifier. Its performance is well dependent on its structure, the number of hidden units and feature selection on a proposed problem.

In text-dependent experiment, a high recogniton rate of 98.5 percent was achieved, with 23 hidden units. In text-independent tests, the highest recognition rate was 95.2 percent with 15 hidden neurons. Although they is not very high comparing with many conventional techniques, such as VQ-codebooks, our tests utilized very limited feature of utterance, and it is much simple to test a testing utterance when the network has been trained. The networks are able to accept new training data for catching a slow change in a speaker'database. When more frames were used, recognition rate could reach a very high level, as shown in Table 1.

We also expect that a learning network may directly utilize the speech samples for recognition tasks. The feature extraction may be combined with its learning process, and the network may also learn transient factors of the features. A type of Time-Delay neural network [11] is a good example.

6. REFERENCES

[1] B.S.Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Amer., vol.55, pp.1304-1312, June,1974.

[2] B.S.Atal, "Automatic Recognition of Speakers from Their Voices," Proc. IEEE vol.64, No.4, pp.460-675, Apr. 1976.

[3] S.Furui, "Cepstral Analysis technique for Automatic Speaker Verfication," IEEE Trans. ASSP-29, No.2, pp.254-272, Apr. 1981.

[4] F.K.Soong, et al., "A Vector Quantization Approach to Speaker Recognition," AT&T Tech. J, vol.66, pp.14-26, 1987.

[5] F.K.Soong, et al., "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," IEEE trans. ASSP-36, No.6, June, 1988.

[6] J.M.Naik, et al.,"High Performance Speaker Verification Using Principal Spectral Components," ICASSP-86, vol.2, pp. 881-884, 1986

[7] R.P.Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Mag., pp.4-22, Apr.1987

[8] J.J.Hopfield, " Neural Network and Physical Systems with Emergent Collective Computational Abilities," Proc. Nat. Acad. Sci. USA, vol.79,pp.2554-2558, Apr.1982

[9] D.E.Rumelhart, et al., *"Parallel Distributed Processing"* Cambridge MA, M.I.T. Press, 1986

[10] D.J.Burr, "Experiments on Neural Net Recognition of Spoken and Written Text," IEEE Trans.ASSP-36, No.7, pp.1162-1168, July 1988

[11] A.Waibel, "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans.ASSP-37, No.3, pp.328-339, Mar.1989

[12] J.T.Buck, et al., "Text-Dependent Speaker Recognition Using Vector Quantization," Proc. ICASSP-85, vol.1, pp.391-394.