International Journal of Fuzzy Logic and Intelligent Systems, vol. 5, no. 1, March 2005 pp. 76-82

Speaker Identification Based on Incremental Learning Neural Network

Kwang-Seung Heo and Kwee-Bo Sim

School of Electrical and Electronic Engineering, Chung-Ang University

Abstract

Speech signal has various features of speakers. This feature is extracted from speech signal processing. The speaker is identified by the speaker identification system. In this paper, we propose the speaker identification system that uses the incremental learning based on neural network. Recorded speech signal through the microphone is blocked to the frame of 1024 speech samples. Energy is divided speech signal to voiced signal and unvoiced signal. The extracted 12 orders LPC cpestrum coefficients are used with input data for neural network. The speakers are identified with the speaker identification system using the neural network. The neural network has the structure of MLP which consists of 12 input nodes, 8 hidden nodes, and 4 output nodes. The number of output node means the identified speakers. The first output node is excited to the first speaker. Incremental learning begins when the new speaker is identified. Incremental learning is the learning algorithm that already learned weights are remembered and only the new weights that are created as adding new speaker are trained. It is learning algorithm that overcomes the fault of neural network. The neural network repeats the learning when the new speaker is entered to it. The architecture of neural network is extended with the number of speakers. Therefore, this system can learn without the restricted number of speakers.

Key words : LPC, LPCC, Neural Network, Back propagation, Incremental Learning

1. Introduction

The subject of speaker recognition can be divided into two main areas, that is, speaker verification and identification. Speaker verification is concerned with the verification whether a speaker is the person he claims to be or not, and involves a binary decision whether the test utterance matches the features of the claimed speaker. The purpose of a speaker identification system is to determine the identity of a speaker among several speakers of known speech characteristics, from sample of his or her voice. Speaker identification can be divided into two categories: text dependent and text independent. To reduce the complexity, speaker identification system may be confined to recognize chosen texts; such a system is called a text-dependent system. Text independent system identifies the speaker regardless of his utterance.

The speaker identification system divides by two parts. First is the feature extraction part from speech. Second is the speaker identification part. Two parts can divide the method of feature extraction in speech: Time domain and frequency domain. Time domain part analysis the speech and extract the feature through time domain. The time domain parameters consisted of linear prediction coefficients, reflections coefficients, log area ratio coefficients, and cepstral coefficients. The frequency domain parameters consisted of inverse filter spectral coefficients and speech spectrum parameters [1].

The learning method to identify speaker can classify by greatly 3. First is method to use euclidean distance. This method identifies speakers using approach that distinguish the data in nearest distance saving euclidean distance between reference data and test data [2]. But, this method requires much amounts memory and computation because it must remember all test and reference data. Second is method to use VQ (Vector Quantization). By make codebook with the speech feature that is extracted from speaker and pay distance value in the nearest codebook because saving codebook's centroid, it is method to look for speaker characteristic point that have fewest distance [3]. Third is method to have the neural network and achieve speaker identification. The neural network is efficient fairly in memory side because learn by fewer parameter that is less and simple computation and does not compose recognizer for each speaker. But, this has inconvenient point that speaker must study newly whenever change. If the neural network that is used in speaker identification, there are MLP (Multi Layer Perceptron), TDNN (Time Delay Neural Network), RBF (Radial Basis Function), LVQ (Learning Vector Quantization) [4]. Decision tree neural network is used beside above neural network [4].

The neural network has a fault that must study newly when new input data enters. The incremental learning method in neural network overcomes the fault of the conventional neural network [5].

Manuscript received Dec. 21, 2004; revised Feb. 2, 2005
This research was supported by the Brain Neuroinformatics Research Program sponsored by Korea Ministry of Science and Technology.

Incremental Learning has the structure that node is added if new input enters to the existent learned neural network. When new input enters, it studies about new input as is different from existent neural network that repeat learning [6]. This learning method can efficiently divide many speakers.

In this paper, the speaker identification system is divided two steps. First analyzes by Frame using speech signal that is recorded through computer as the feature extraction step and uses Cepstral Coefficients that get via LPC process by learning data. Second is gone by basic learning that use Bakcpropagation and is added by learning that achieve Incremental learning about new speaker. Main discourse presents the speech feature extraction and speaker identification model that use incremental Learning through an experiment.

2. Speaker's Feature Extraction

Speaker's feature extraction in the speech signal is important step to raise performance of the speaker identification system. Fig. 1 shows the whole process about speech signal processing. Analog speech signal is changed to digital signal by A/D converter. There is method to process such changed signal two method. One is method to process to single digital sample and the other is frame-processing method to process after store fixed quantity's sample [7]. Window's role is important as speech preprocessing in frame processing mode.

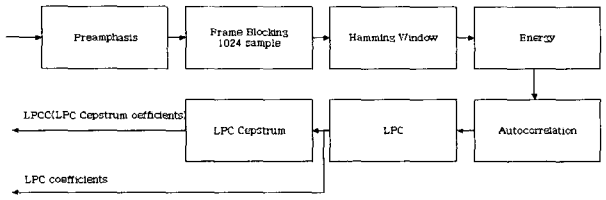

Fig 1. Speech Signal processing.

2.1 Hamming Window

The first exploration for treating objective functions separately in Evolutionary Algorithms (EAs) was launched by Schaffer[24] [25]. In his dissertation, Schaffer proposed his Vector Evaluated Genetic Algorithm (VEGA) for searching solution set to solve MOPs. He created VEGA to find and maintain multiple classification rules in a set-covering problem. VEGA tried to achieve this goal by selecting a fraction of the next generation using one of each of the attributes (e.g., cost, reliability)[14]. Other approaches that search solutions for MOPs include those of Fourman[10], Kursawe[17], and Hajela and Lin[13]. However, as none of them makes direct use of the actual definition of Pareto-optimality, different non-dominated individuals are generally assigned the different fitness values [8].

Window acts role that do as can see one part of signal when there is long signal. There are many possible windows (e.g. Rectangular, Hanning, Hamming, Blackman, Kaiser, etc), some of which are defined as follows [8].

The rectangular window has the highest frequency resolution due to the narrowest main lobe, but it has the largest frequency leakage. So it is not usually used in speech spectral analysis. A typical window used for the autocorrelation method of LPC (the method most widely used for recognition systems) is hamming window [9]. Hamming window adopts weighting that is fixed pattern's symmetry in interested part. In Fig. 2, the left figure shows sine wave. The right figure explains sine wave of hamming window.

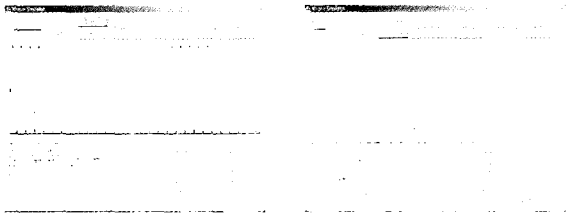

Fig 2. Hamming window of sine wave

Fig. 3 shows the speech signal using hamming window. About original signal, there is advantage that can get signal that there is good smoothing and no frequency leakage more than other window [1].

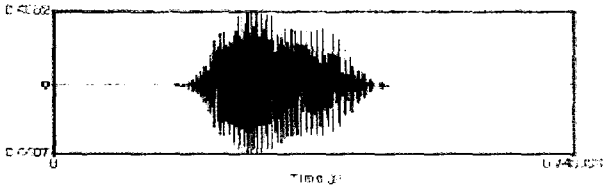

Fig 3. Speech Signal using Hamming Window

2.2 Energy

Speech signal's magnitude variety can appear using short-time energy. Change of magnitude in unvoiced speech is small and is the reverse in voiced speech [8]. The voiced speech and unvoiced speech can classify using energy. Energy is acquired with summing the square value of the sample value.

$$E(m) = \sum_{n=-\infty}^{\infty} s(n)^2 \quad (1)$$

It is approximated short-energy function that speech signals are divided with the frame acquiring N samples and each sample is squired and summed.

$$E(m) = \sum_{n=m-N+1}^{m} [s(n)w(m-n)]^2 \quad (2)$$

In equation (2), E(m) is the energy value and s(n) means each sample value in frame. w(m-n) means the window. The window's role is as follows.

$$w(m-n)=\begin{cases}1 & 0\le n\le m\\ 0 & otherwise\end{cases} \tag{3}$$

Energy is small in unvoiced speech and appears greatly in voiced speech.

Fig. 4 shows the energy about 5 vowels. The voiced speech and unvoiced speech are classified. The part that magnitude is big is the voiced speech and small part is the unvoiced speech. Energy can not distinguished unvoiced speech and silence speech. If it is added Zerocrossing rate to solve this, they are classified.

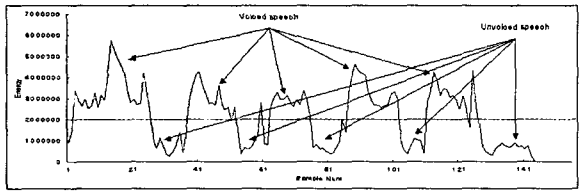

Fig 4. Voiced speech and unvoiced speech by energy.

2.3 LPC(Linear Predictive Coding)

Basic idea of LPC Model is that speech signal in the appointed Nth can approximate linear prediction of Pth speech signal [7]. This method is widely used because it is fast and simple, yet an effective way of estimating the main parameters of speech signals [10]. As shown in Fig. 5, an all pole filter with a sufficient number of poles is a good approximation for speech signal [10]. The filter H(z) could model as

$$H(z)=\frac{X(z)}{E(z)}=\frac{1}{1-\sum_{k=1}^{p}a_k z^{-k}}=\frac{1}{A(z)} \tag{4}$$

Where p is the order of the LPC analysis.

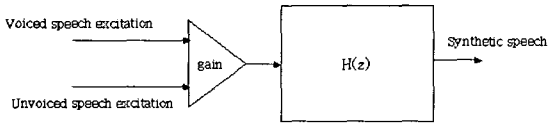

Fig 5. A mixed excitation source-filer model of speech.

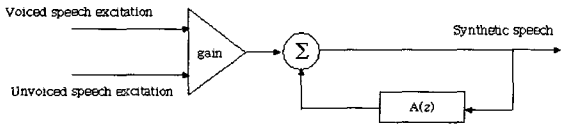

Fig 6. A mixed excitation source-filer model of speech using inverse filter.

The inverse filter A(z) is defined as

$$A(z)=1-\sum_{k=1}^{p}a_k z^{-k} \tag{5}$$

Taking inverse z-transform in eq. (5) results in

$$x[n]=\sum_{k=1}^{p}a_k x[n-k]+e[n] \tag{6}$$

The prediction error when using eq. (6)

$$e[n]=x[n]-\sum_{k=1}^{p}a_k x[n-k] \tag{7}$$

The short-term prediction error for that segment define as

$$E_m=\sum_{n}e_m^{\,2}[n] \tag{8}$$

So, given a signal , the LPC coefficients are estimated as those that minimize the total prediction error [10]. For LPC coefficients, The three different algorithms are presented: the covariance method, the autocorrelation method, the lattice method. In this paper, LPC coefficients are acquired with the autocorrelation method. $R_m[k]$ is the autocorrelation sequence of $x_m[n]$

$$R_m[k]=\sum_{n=0}^{N-1-k}x_m[n]x_m[n+k] \tag{9}$$

So, LPC equations are acquired with eq. (10)

$$\sum_{j=1}^{p}a_j R_m[|i-j|]=R_m[i] \tag{10}$$

Which corresponds to the Toeplitz matrix. Durbin's recursion exploits this fact resulting in a very efficient algorithm [8].

Durbin's recursion is as follows:

① Initialization

$$E^0=R[0] \tag{11}$$

② Iteration. For $i=1,\mathrm{K},p$ do the following recursion

$$k_i=(R[i]-\sum_{j}^{i-1}a_j^{i-1}R[i-j])/E^{i-j} \tag{12}$$

$$a_i^i=k_i \tag{13}$$

$$a_j^i=a_j^{i-1}-k_i a_{i-j}^{i-1},\quad 1\le j<i \tag{14}$$

$$E^i=(1-k_i^2)E^{i-1} \tag{15}$$

③ Final Solution

$$a_j=a_j^p,\quad 1\le j\le p \tag{16}$$

Where the coefficients called predictive coefficients or LPC coefficients. In the process of computing the predictor coefficients of order p, the recursion the finds the solution of the predictor coefficients for all orders less than p [10].

Fig. 7 shows the LPC coefficients of vowel a. 1 frame consists of 12 LPC coefficients. The identified pattern is repeated in the each frame. This special quality is used to the speaker identification system with LPCC.

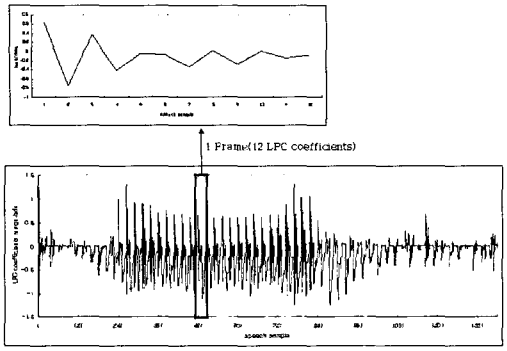

Fig. 7. LPC coefficients of vowel/a/.

2.4 LPC Cepstral Coefficients

The cepstrum c(n) is defined as the inverse z-transform of c(z).

$$C(Z)=\sum_{n} c(n)z^{n} \tag{17}$$

Given that all poles $z=z_i$ are inside the unit circle and the gain is 1, the causal LPC cepstral coefficient ($c_{lp}(n)$) is given by

$$c_{lp}(n)=\begin{cases}\dfrac{1}{n}\sum_{i=1}^{p} z_i^{n} & n>0\\ 0 & n\le 0\end{cases} \tag{18}$$

A recursive relation between the LPC cepstral coefficient and the predictor coefficients is given as [8]

$$c_{lp}(n)=\alpha_n+\sum_{i=1}^{n-1}\left(\frac{i}{n}\right)c_{lp}(i)\alpha_{n-i}\,,\quad 1<n<p \tag{19}$$

In equation (19), c is the LPC cepstral coefficients and α is the predictor coefficients. Also, p means order value. The small order value doesn't draws speaker's feature and The high order value does extract by feature in speech signal's noise. In this paper, the order value uses 12 degrees.

Fig. 8 shows the LPC cepstral coefficients in vowel a. In Fig. 8, transverse means speech sample and length displays cepstral coefficients magnitude.

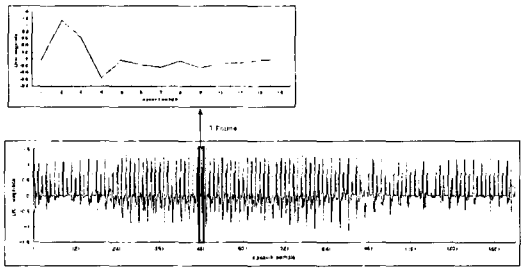

Fig. 8. LPC cepstral coefficients of vowel /a/.

3. Incremental Learning Neural Network

Many researches have presented incremental learning methods [13]. Park et al. have proposed an incremental learning system that predicts the change of the output to past input patterns and modifies parameters in such a way that the predicted change of the output becomes small [14]. Yoneda et al. have proposed a fine system that makes the network relearn a small number of compressed past input patterns stored in a buffer. The compression is realized by clustering techniques as discussed below. If the buffer becomes full, the system makes a space by merging the two most similar patterns in the buffer to one averaged pattern. Then, a new pattern is stored into the space in the buffer. However, there is no guarantee that the compressed patterns contain the interfered patterns. As a result, there are cases that some old memories are interfered with by new learning [15].

An incremental learning system updates its hypotheses as a new instance arrives without reexamining old instances. In other words, an incremental learning system learns Y based on X, and then learns Z based on Y, and so on. Such a learning strategy is both spatially and temporally economical since it need not store and reprocess old instances. It is especially crucial for a learning system that continually receives input and must process it in a real-time manner. Also, learning based on a single instance has been an important topic in machine learning [13].

In proposed system, incremental learning is learning that node about input is added if new input enters to learn neural network. When new input enters, it learns about the only new input. This learning method is different from conventional neural network that repeats learning.

Through backpropagation learning, the completed basic MLP identifies 4 speakers. If new speaker's speech data enters, learned MLP does not identify new speaker. Output node does not become excited. Conventional Neural Network repeats learning on the whole. Therefore, according as speaker number is increased, learning time takes much. If data about new speaker enters for effective learning, hidden node and output node increase two by one. Incremental Learning is gone through this learning process.

In the structure, added hidden nodes and output nodes are fully connected. The output in the added hidden node enters into the input of the output nodes in the basic neural network. For the added output node is excited, the setting of target value is important. When new speaker's data is entered target value is set by *, *, *, *, 1. * is don't care. * is increased as the number of new speaker. Because the maximum output node is the identified speaker.

Fig. 9 shows the incremental learning in MLP. Each speaker has two hidden node and one output node. In proposed system, incremental learning use MLP structue and backpropagation learning algorithm. This is the simple method and the computation parameter is lower than the other incremental learning algorithm.

Incremental learning is consisted of following 4 steps.

① Learning in the neural network.

② Testing in the neural network.

③ Learning in the incremental learning neural network.

④ Testing in the incremental learning neural network.

4 steps are processed in order.

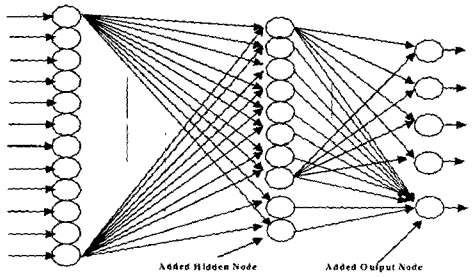

Fig. 9. Incremental Learning in MLP

① The first step : Learning in the neural network.

The feature from speech signal is entered to the neural network. Through the feed-forward and back-ward learning, the weights are updated. The error value is checked while the weights are updated. In the backward, the target value is set to the 1 in the activated output node and o in the other output node. The learning stop condition is that the error value is smaller than fixed error value. The next process is the testing in the neural network.

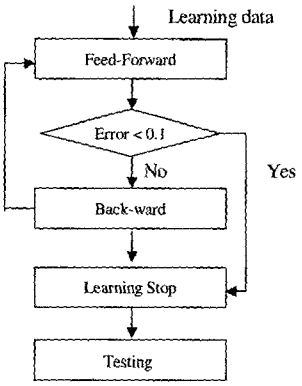

Fig. 10. Learning in the neural network

② The second step: testing in the neural network.

After the learning using backpropagation learning algorithm, the testing data(the feature of the speech signal) is entered to the neural network. Through the feed-forward, the output node is activated when the output of the one output node is larger than 0.9. The first output node means the first speaker. The second output node means the second speaker.

Incremental learning begins when the outputs of output nodes are smaller than 0.9.

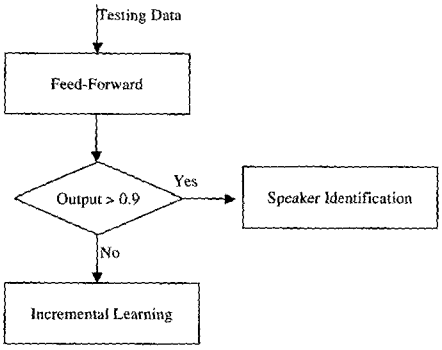

Fig. 11. Testing in the neural network

③ The third step: learning in the incremental learning.

2 Hidden nodes and 1 output nodes are added to the neural network. The added hidden nodes are fully connected with the input nodes but they are connected with only the added output. The hidden nodes of neural network are also connected with the added output, because the learned weights are affected to the output of the added output node.

The target value is set to the {*, *, *, *, 1}. The added nodes are learned and the other nodes are not learned. The added output node is activated to the new input learning data.

But, when the new learning data is increased with the new speaker, all of the added output nodes are activated. Because the learned weights are less influenced to the added nodes.

So, identified speaker learning data is entered to the neural network. Two learning data (identified speaker learning data, new speaker learning data) is learned through backpropagation learning algorithm. The learning stop condition is identical to the learning in the neural network.

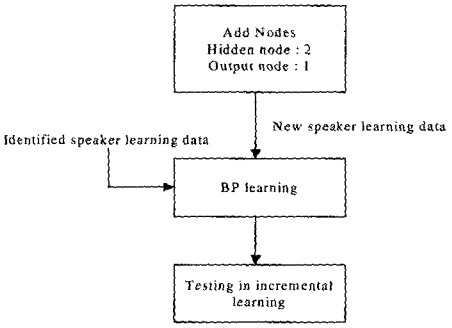

Fig. 12. Learning in the incremental learning

④ The forth step: testing in the incremental learning.

After incremental learning process, the new speaker testing data is entered to the neural network of the incremental learning. The output of added output node is excited when this step ends. If the output of output nodes is smaller than 0.9, the incremental learning is repeated until the learning process ends.

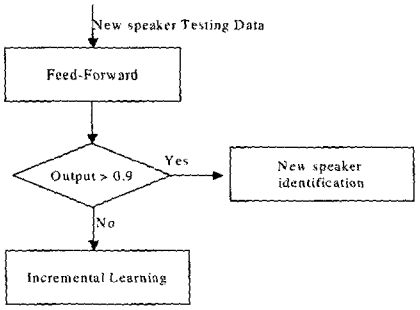

Fig. 13. Testing in the incremental learning

4. Simulation Results

4.1 Speech Signal Processing

The speech signal is recorded with 16bit, mono, 11.025Khz. Because speech signal can be band limited to 10Khz without significantly affecting the hearer's perception [10]. The number of speaker is 8. The each speaker says the 10 short sentences. They are repeated to 10. So, All speech data are 800. Speech length is different from each speaker. Although text is same, speech length is different in each speaker. So, frame is set by smallest frame in the extracted frame. The smallest is 20 frames. The each frame has the 12 order cepstral coefficients. The number of all feature data is 16000.

4.2 Incremental Learning Neural Network

MPL has the structure of 12 input nodes, 8 hidden nodes, and 4 output nodes.

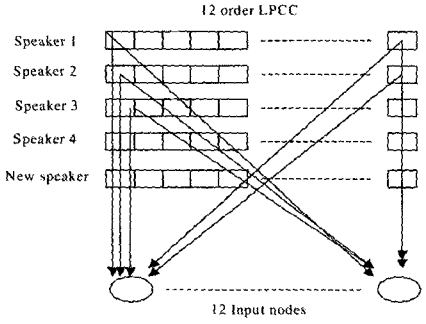

Fig. 14. The learning data input method

Fig. 14 explains the input method of learning data. Each frame of the speakers enters the neural network. 4 speakers are identified through the learning and testing. New speaker is identified through the incremental learning. Learning rate is set to 0.1. Error value is 0.1 when the learning stops.

Fig. 15 shows the result of identification rate of speech signal length. The identification rate of short speech signal(0.5 ~ 0.7sec) is lower than the identification rate of long speech signal, because the short speech signal is a similar speech signal. The identification rate of long speech signal (0.85sec) is 96.25%.

Fig. 15. The identification rate of speech signal length

Fig. 16 shows the result of identification rate of 8 speakers. Each of 8 speakers has the identification rate. The second speaker is the best identification rate (82%). The third speaker is the worst identification rate (50%).

The identification rate of this system is 65.2%. This is nearly equal the result of MLP. But, The learning time takes shortly more than MLP.

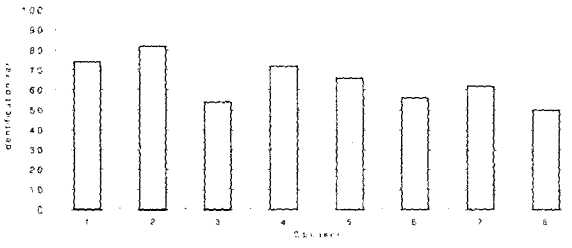

Fig. 16. The identification rate of 8 speakers

5. Conclusion

In this proposed system, the feature of speakers is LPCC that is the good identification rate in the speaker identification system. The speaker identification part uses the neural network that the learning algorithm is the incremental learning algorithm. The structure of neural network is MLP that has the 12 input nodes, 8 hidden nodes, and 4 output nodes. The incremental learning begins when the new speaker is entered to the system.

In the structure of neural network, 2 hidden nodes and 1 output node are added to the neural network. The incremental learning complements a defect of the neural network. The neural network repeats the learning process whenever the new input data enters to the neural network. Incremental learning begins when the new speaker is identified. Incremental learning is the learning algorithm that the learned weights are remembered and only the new weights that are created as adding new speaker are trained. The number of speakers is 8. The identification rate of this system is 65.2%. This is nearly

equal the result of MLP. But, The learning time takes shortly more than MLP. The architecture of neural network is extended with the number of speakers. So, this system can learn without the restricted number of speakers.

References

[1] N. Mohankrishnan, M. Shridhar, M.A. Sid-Ahmed "A Composite Scheme for Text-Independent Speaker Recognition," *Acoustic, Speech and Signal Processing, IEEE International Conference on'82*, vol. 7, pp. 1653-1656, 1982

[2] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *J. Acoustic. Soc. Amer*, vol. 35, pp. 354-358, Apr 1971

[3] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, B.H. Juang, "A vector quantization approach to speaker recognition," *in Proc. ICASSP*, pp. 387-390, 1985

[4] Kevin R.Farrell, Richard J.Mammone, Khaled T.Assaleh, "Speaker Recognition Using Neural Networks and Conventional Classifiers," *IEEE Transaction on speech and audio processing*, vol. 2, no. 1, pp. 194-205, January 1994

[5] K.Farrell, R.J.Mammone, A.L.Gorin, "Adaptive Language Acqusition Using Incremental Learning," *Acoustics, Speech, and Signal Processing, 1993, ICASSP-93, 1993, IEEE International conference on*, vol. 1, pp. 501-504, Apr 1993

[6] R.Poliker, L.Udpa, S.S.Udpa, V.Honavar, "Learn++: An Incremental Learning algorithm for Multilayer perceptron networks," *Acoustic, Speech and Signal Processing, 2000, ICASSP'00, Proceedings, 2000, IEEE International Conference on*, vol. 6, pp. 3414-3417, 2000

[7] Jin-soo Han, *Speech Signal Processing*, Osung Media, 2000.

[8] A.M. Kondoz, *Digital Speech coding for low bit rate communications systems*, John Wiley & Sons, 1994

[9] Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of speech recognition*, Prentice-Hall International Inc., 1993

[10] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Spoken Language Processing A guide to Theory, Algorithm, and System Development,

[11] Raul Rojas, *Neural Networks A systematic Introduction*, Springer, 1996

[12] Simon Haykin, *Adaptive Filter theory*, Prentice Hall Information And System Science Series, 2001

[13] Koichiro Yamauchi, Nobuhiko Yamaguchi, Naohiro Ishii, "Incremental Learning Methods with Retrieving of ..Interfered Patterns," *IEEE Transaction on Neural Network*, vol. 10, no. 6, pp. 1351-1365, November 1999

[14] D. C. Park, M. A. El-Sharkawi, R. J. Marks II, "An adaptively trained neural network," *IEEE Trans. Neural Network*, vol. 2, pp. 334-345, May 1991

[15] T. Yoneda, M. Yamanaka, Y. Kakazu, "Study on optimization of grinding conditions using neural networks-A method of additional learning," *J. Japan Soc. Precision Eng.*, vol. 58, no. 10, pp. 1707-1712, Oct 1992

Kwang-Seung Heo

Kwang-Seung Heo received his B.S. and M.S. degreed in the School of Electrical and Electronics Engineering from Chung-Ang University in 2002 and 2004 respectively. His areas of interest include artificial life, intelligent robot, intelligent system, Speaker Identification and Embeded System.

Kwee-Bo Sim

Kwee-Bo Sim received his B.S. and M.S. degrees in the Department of Electronic Engineering from Chung-Ang University in 1984 and 1986 respectively, and Ph.D. degree in the Department of Electrical Engineering from The University of Tokyo, Japan, in 1990. Since 1991, he has been a faculty member of the School of Electrical and Electronic Engineering at Chung-Ang University, where he is currently a Professor. His areas of interest include artificial life, neuro-fuzzy and soft computing, evolutionary computation, learning and adaptation algorithms, autonomous decentralized systems, intelligent control and robot systems, artificial immune systems, evolvable hardware, and artificial brain etc. He is a member of IEEE, SICE, RSJ, KITE, KIEE, ICASE, and KFIS.

Phone : +82-2-820-5319
Fax : +82-2-817-0553
E-mail : kbsim@cau.ac.kr