Text-Independent Speaker Identification System Based On Vowel And Incremental Learning Neural Networks

Kwang-Seung Heo, Dong-Wook Lee, and Kwee-Bo Sim

School of Electrical and Electronics Engineering, Chung-Ang University, Seoul, 156-756, KOREA
(Tel : +82-2-820-5319 ; Fax : +82-2-817-0553 ; E-mail : kbsim@cau.ac.kr)

Abstract: In this paper, we propose the speaker identification system that uses vowel that has speaker's characteristic. System is divided to speech feature extraction part and speaker identification part. Speech feature extraction part extracts speaker's feature. Voiced speech has the characteristic that divides speakers. For vowel extraction, formants are used in voiced speech through frequency analysis. Vowel-a that different formants is extracted in text. Pitch, formant, intensity, log area ratio, LP coefficients, cepstral coefficients are used by method to draw characteristic. The cpestral coefficients that show the best performance in speaker identification among several methods are used. Speaker identification part distinguishes speaker using Neural Network. 12 order cepstral coefficients are used learning input data. Neural Network's structure is MLP and learning algorithm is BP (Backpropagation). Hidden nodes and output nodes are incremented. The nodes in the incremental learning neural network are interconnected via weighted links and each node in a layer is generally connected to each node in the succeeding layer leaving the output node to provide output for the network. Though the vowel extract and incremental learning, the proposed system uses low learning data and reduces learning time and improves identification rate.

Keywords: Formant, Cepstral coefficients, MLP, Incremental Learning

1. INTRODUCTION

The subject of speaker recognition can be divided into two main areas, 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 recognition 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].

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 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 Neural Network and achieve speaker identification. Neural Network is efficient fairly in memoery 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 examine 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].

Neural Network has a fault that must study newly when new input Data enters. For this fault, method that applies Incremental Learning method in Neural Network is proposed [5].

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 divide many speakers efficiently.

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 IDENTIFICATION SYSTEM

Speaker's feature extraction in the speech signal is important step to raise performance of the speaker identification system.

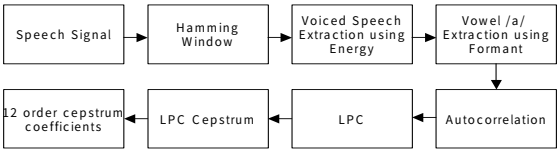

Fig. 1. Speech Signal processing

Fig. 1 shows 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.

2.1 Hamming Window & Energy

Window acts role that do as can see one part of signal when there is long signal. Hamming Window adopts weighting that is fixed pattern's symmetry in interested part [1].

$$W(n) = \begin{Bmatrix} 0.54 - 0.46\cos(2\pi n / M) & o \le n \le M \\ 0 & otherwise \end{Bmatrix} \quad (1)$$

About original signal, there is advantage that can get signal that there is good smoothing and no frequency leakage more than other Window.

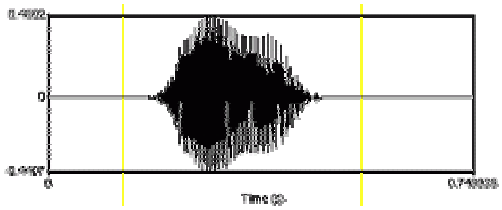

Fig. 2. Speech Signal using Hamming Window

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.

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

In equation (2), E (m) is the energy value and S (n) means each sample value in frame. Energy is small in unvoiced speech and appears greatly in voiced speech.

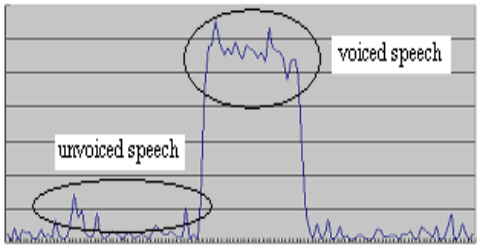

Fig. 3. Energy

Fig. 3 shows the energy about vowel /a/. 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.

2.2 Formant

In the FFT Spectrum, Formant frequency is detected. Formant Frequency is important part of speech signal processing. Formant Frequency is resonance frequency band created from vocal cords and nasal tract [7]. From lower formant, formant frequencies are added (Formant1, Formant2, Formant3…). When speaker tell "/a/", "/e/","/i/","/o/","/u/", Formant1 and Formant2 frequency divide "/a/", "/e/","/i/","/o/","/u/"

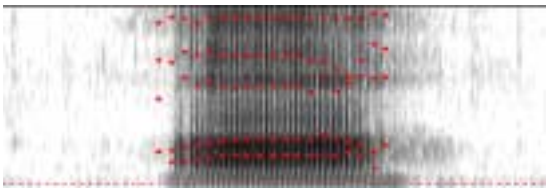

Fig. 4. Vowel /a/ Formant in FFT spectrum

Fig. 4 demonstrates vowel /a/ formant. From first line, Formant is increased to formant1, formant2 etc.

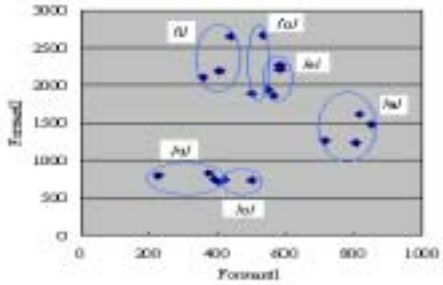

Fig. 5. Formant in vowels

Fig. 5 shows formant about 5 vowels. In Fig. 5, vowel/a / than other vowels better divide. Using this special quality vowel /a/ in voiced speech is extracted. In the next step, vowel /a / cpestrum coefficients through LPC is extracted.

2.3 LPC(Linear Predictive Coding)

The general feature-extraction step of interest here can be divided into two parts. First, LP analysis of speech is carried out to produce a set of predictor coefficients. Second, the predictor coefficients are transformed into feature vectors that are the Cepstrum coefficients.

LPC Model' basic idea is that speech signal in the appointed N can approximate linear prediction of Pth speech signal [7].

2.4 LPC Cepstral Coefficients

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

$$C(z) = \sum_{n} c(n) z^{-n} \quad (3)$$

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

$$c_{lp}(n) = \frac{1}{n}\sum_{i=1}^{p} z_i^n \qquad n>0$$
$$= 0 \qquad n \le 0 \quad (4)$$

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

$$c_{lp}(n) = \alpha_n + \sum_{i=1}^{n-1}(\frac{i}{n})c_{lp}(i)a_{n-i} \qquad 1<n<p \qquad (5)$$

In equation (5), 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.

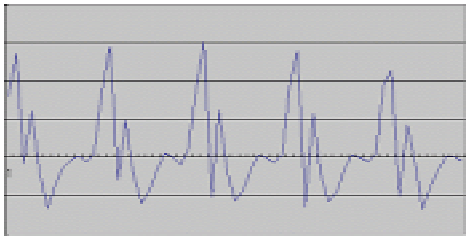

Fig. 6. LPC cepstral coefficients in vowel /a/

Fig. 6 shows the LPC cepstral coefficients in vowel /a/. In Fig. 6, transverse means Frame and length displays Cepstral Coefficients' Magnitude.

2.5 Incremental Learning Neural Network

Neural Network acts role that distinguish speaker using speaker's speech features. In this paper, we propose MLP structure that Backpropagation and Incremental Learning algorithm are used. MLP is composed of 3 layers: input layer, hidden layer, and output layer. Hidden layer and output layer are increased with the number of speakers. Two steps compose learning. First is Basic learning that uses BP and second is learning that uses incremental learning.

2.5.1 Backpropagation

The BP learning algorithm is used for basic learning. Because BP uses gradient descent, it is learning algorithm that finds smallest error value in weight [9]. The Neural Network consists of 12 input nodes, 8 hidden nodes and 4 output nodes. Hidden layer and output nodes increment as the number of speaker incremented. Initial weights are set up randomly and updated until the error value is lower than the fixed error.

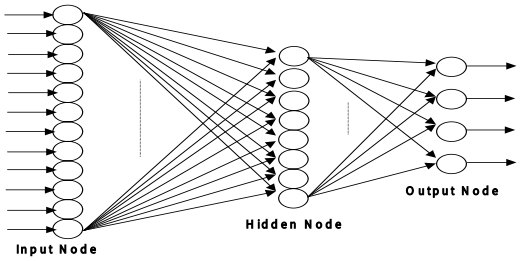

Fig. 7. MLP structure

Fig. 7 shows MLP's structure that applies Backpropagation learning algorithm. 4 Output Node means speaker's number. Input Data is entered by frame about 4 speakers. Error value finds using output of middle class and output layer and target value. Target value is established by 1, 0, 0, 0 about first speaker because is activated when output value is "1" in 4 output nodes and second is established by 0, 1, 0, 0. This Is equal at third and fourth speaker's case.

2.5.2 Incremental Learning

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 [8].

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 NN. 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.

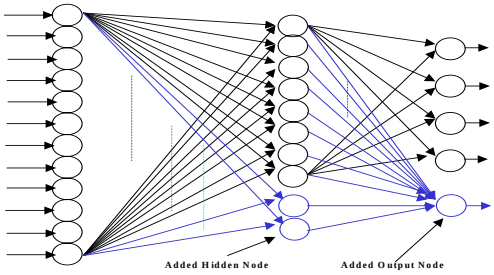

Fig. 8. Incremental Learning in MLP

Fig. 8 shows the incremental learning in MLP.

3. EXPERIMENTS

3.1 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 6. The each speaker says the 7 short sentences. So, All speech data are 42. 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. All feature data are 240.

3.2 Incremental learning Neural Network

Learning step is consists of 4 step. The fist part is learning in the basic NN using BP. The learning data is cepstral coefficients in vowel /a/. Vowel /a/ is recorded by computer from 4 speakers. If error value is low than 10%, learning is stopped. Each weight is memorized. The second part is testing in the basic NN using BP. Test uses 11 sentences. Test data is vowel /a/ cepstral coefficient that is extracted in each sentence. Output's order that the biggest value appears means speaker. First speaker is identified with the excited value of First output node. Also, the excited values of output nodes are above 0.8. If the outputs of all output nodes are below 0.8, incremental learning is begun. The third part is the incremental learning. If new speaker is entered, hidden nodes and output node are

increased with two by one. Added output node is connected with hidden node of the learned basic NN and added node is connected with only added output node. Weights of the learned NN are memorized and added weights are learned. The last step is test of the incremental learning.

The number of hidden node is set by experiment. In experiment, the best result is generated in 8 hidden nodes. Learning rate is 0.1. The fixed error value in incremental learning is set by 12.5% through experiment.

Table 1. Identification rate

Speaker	Identification rate (%)
Speaker 1	100
Speaker 2	100
Speaker 3	100
Speaker 4	100
New speaker 1	87
New speaker 2	71

Table 1 shows the identification rate of 6 speakers. Speaker is learned by the basic NN and is tested by the basic NN. New speaker is learned and tested by the incremental learning NN. The identification of the system is 93%.

Although the number of speaker is increased, the identification rate is lower. The reason is the lack of learning data. If learning data is increased, the identification rate is better than value of former rate.

4. CONCLUSION

Proposed speaker identification system uses vowel that extract in text. Each vowel of speakers is classified using Formant. Through Autocorrelation, LP coefficients are obtained and cepstral coefficients explain the features of speakers. The cepstral coefficients are entered the Neural Network with 12 input nodes, 8 hidden nodes, and output nodes. The output node is increased with the speakers. The whole text is learned in conventional speaker identification system. However conventional system has defects that identification rate is different by learning text. We propose the system that extracts vowel in text and then identify speakers. Though the proposed system uses low learning data, it reduces learning time and improves identification rate.

ACKNOWLEDGMENTS

This research was supported by grant No. N09-A08-4301-05(Development of the Key Technology for Autonomous Family Machine) from the project of Developing SIC(Super Intelligent Chip) and its Applications under the program of Next generation technologies in 2000 of Ministry of Commerce, Industry and Energy.

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] Ehab F.M.F.Badan, Hany Selim, "Speaker Recognition Using Artificial Neural Networks Based on Vowel phonemes", *Signal Processing Proceedings 2000, WCCC-ICSP 2000 5th International Conference on*, vol.2, pp. 796-802, 2000

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

[10] Xuedong huang, Alex acero, Hsiao-wuen hon, *Spoken Language Processing*