SOM과 LVQ에 의한 자음의 분류

이채봉* · 이창영**

Classification of Consonants by SOM and LVQ

Chai-bong Lee* · Chang-young Lee**

요 약

음성타자기의 구현에 접근하려는 노력의 일환으로서, 우리는 본 논문에서 자음의 분류에 대해 연구한다. 많은 자음들은 시간에 따른 주기적 거동을 보이지 않고 따라서 그들에 대한 푸리에 해석의 타당성에 확신을 갖기 어렵다. 그러므로, 우선 음성 신호로부터 추출되는 MFCC와 LPCC 특징벡터들이 자음에 대해 어느 정도의 의미가 있는지를 파악하기 위하여 LBG 클러스터링을 통한 벡터양자화를 수행한다. VQ의 실험적 결과는 자음에 대한 푸리에 해석의 타당성에 관해 분명한 결론을 내리는 것이 쉽지 않음을 보여주었다. 자음의 분류를 위해 SOM과 LVQ의 두 가지 신경망이 사용되었다. SOM의 결과는 몇 쌍의 자음들이 나뉘어 분류되지 않음을 보여주었다. LVQ에서는 본질적으로 이 문제가 사라지지만 자음의 분류 정확도는 낮은 수준이었다. 이로부터, LVQ에 의한 자음 분류에 있어서는 MFCC 및 다른 특징 벡터들이 함께 사용되어야 함이 사료된다. 하지만 본 연구에서 도입한 MFCC/LVQ의 결합은 기존의 언어모델을 기반으로 하는 음소 분류에 비해 그 결과가 나쁘지 않은 것으로 나타났다. 모든 경우에 LPCC 특징벡터는 MFCC에 비해 그 결과가 좋지 않았다.

ABSTRACT

In an effort to the practical realization of phonetic typewriter, we concentrate on the classification of consonants in this paper. Since many of consonants do not show periodic behavior in time domain and thus the validity for Fourier analysis of them are not convincing, vector quantization (VQ) via LBG clustering is first performed to check if the feature vectors of MFCC and LPCC are ever meaningful for consonants. Experimental results of VQ showed that it's not easy to draw a clear-cut conclusion as to the validity of Fourier analysis for consonants. For classification purpose, two kinds of neural networks are employed in our study: self organizing map (SOM) and learning vector quantization (LVQ). Results from SOM revealed that some pairs of phonemes are not resolved. Though LVQ is free from this difficulty inherently, the classification accuracy was found to be low. This suggests that, as long as consonant classification by LVQ is concerned, other types of feature vectors than MFCC should be deployed in parallel. However, the combination of MFCC/LVQ was not found to be inferior to the classification of phonemes by language-moded based approach. In all of our work, LPCC worked worse than MFCC.

키워드

Speech recognition, Phonetic Typewriter, Self Organizing Map (SOM), Learning Vector Quantization (LVQ)

* 동서대학교 전자공학과(lcb@dongseo.ac.kr) ** 교신저자 : 동서대학교 시스템경영공과(seewhy@dongseo.ac.kr)
접수일자 : 2010. 11. 21 심사(수정)일자 : 2011. 01. 03 게재확정일자 : 2011. 02. 09

I. Introduction

As a method of communication between man and machine, speech recognition provides a very effective interface. Speech input to a machine is about twice as fast as information entry by a skilled typist [1]. The need for and usefulness of speech-to-text transcription cannot be overe-stimated.

The earliest attempt to devise systems for automatic speech recognition by machine is traced back to 1952 when the researchers at Bell Laboratories built a system for isolated digit recognition for a single speaker [2]. As an application of the self-organizing map (SOM) [3], more specifically, Kohonen introduced the notion of a phonetic typewriter (PT) which is intended for transcribing human speech into text automatically. Since then, lots of endeavors have been made for over five decades in this field but practical applications are at the limited level [4-8].

One of the main difficulties that hinder the implementation of PT might be phrased in terms of consonants. In Korean language, especially, there are so many different endings and inflections that PT should construct its output in the form of small units such as syllables or phonemes. Though Waibel et al. achieved high accuracy of phoneme recognition by using time-delay neural networks [9], the experiment was done on three consonants.

Compared to vowels, consonants are characterized by low-energy and short length of time duration. In addition to these unhelpful features, the motivation for Fourier analysis, on which most speech processing technologies rely, is not convincing for consonants since they do not usually show periodic behavior in time domain.

Pattern classification proceeds largely in two stages, one for feature vector extraction from speech signal and the other for pattern classification (recognition) of the feature vectors through a scoring procedure. In this paper, we will consider two types of feature vectors: mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) [10]. Classification of the feature vectors will be performed by neural networks of SOM and learning vector quantization (LVQ), a supervised version of SOM [11]. In spite of many advantages of SOM and LVQ, the application of them to the field of speech recognition has not been active compared to the stochastic approach of hidden Markov models (HMM). Furthermore, investigation on the meaningfulness of MFCC, one of the widely accepted feature vectors in the field speech processing, is hardly found in conjunction with SOM and LVQ. Keeping this in mind, we investigate the following in this paper:

- The meaningfulness of Fourier and its subsequent cepstral analysis for consonants in conjunction with SOM and LVQ
- Classification of Korean consonants by SOM and LVQ

To this end and as preliminaries to the realization of PT, we study classification of 14 phonemes, 13 consonants and /a/ which plays the role of a representative vowel to be compared with consonants. Especially, our objective in this study is to classify all the Korean initial consonants except fortes by neural networks of SOM and LVQ. A distinguished feature of our study is that the output neurons will be arranged in a circular-linear form.

The organization of this paper is as follows. Section II provides overview on SOM and LVQ. Section III describes details on the experiments perpormed in our study. After expounding various results on the phoneme classification by SOM and LVQ with MFCC feature vectors for Korean consonants in section IV, concluding remarks are given in section V finally.

II. SOM and LVQ

One of the main functions of neural networks is pattern classification. Among lots of architectures employed in neural networks, the SOM introduced by Kohonen [1] has many interesting features. Fig. 1 shows its structure.

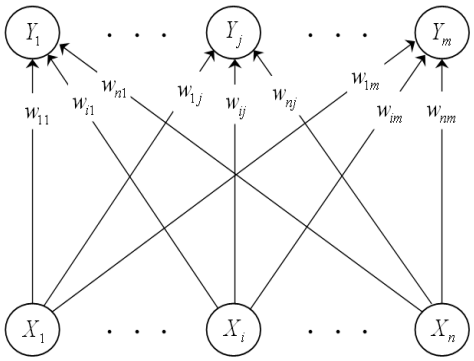

Fig. 1 SOM structure

The input layer and output layer, labeled by X and Y respectively, are connected by weights w. A winning output neuron is determined through a scoring procedure. A distinguished feature of SOM is that the neighbors of the winner have the position of runner-ups, meaning that the winning neuron and its adjacent neurons are eligible to learn the input pattern. As a result, the geometric features of input vector space are naturally mapped onto the space of output neurons.

In this paper, we consider one-dimensional array of output neurons. If the array is naively chosen as a linear form, then the neurons at both ends have only a single neighbor to be trained together. To circumvent this 'unfairness' problem, we arrange the neurons in a closed form, i.e. a rectangular ring. There are then no ends and all the neurons have two nearest neighbors. Learning is then performed on the winning neuron and the adjacent two runner-up neurons, with different learning parameters of necessity.

LVQ, a supervised form of SOM and also developed by Kohonen, has many great features: it requires relatively less training time and creates prototypes that are easy to interpret for experts in the field. The architecture of LVQ is the same as that of SOM. It differs from SOM, however, in that the training proceeds in supervised mode. For this purpose, target value is assigned to each input vector. Since there are no geometrical feature assumed in LVQ, training is performed in winner-takes-all mode. There are many versions of LVQ [12] to enhance the performance but we adopt the basic one [13] in this paper.

III. Experiments

Hundred male speakers pronounced 14 syllables given in Table 1, arranged in Korean alphabetical order. In Korean, glide /r/ and liquid /l/ [14-15] are not classified distinctly and hence the phoneme labelled as /r/ in this paper actually represent both phonemes /r/ and /l/.

Table 1. Fourteen syllables and phonemes

Syllable	Phoneme	Syllable	Phoneme
/ga/	/g/	/ja/	/J/
/na/	/n/	/cha/	/C/
/da/	/d/	/ka/	/k/
/ra/	/r/	/ta/	/t/
/ma/	/m/	/pa/	/p/
/ba/	/b/	/ha/	/h/
/sa/	/s/	/ah/	/a/

Speech utterances were sampled at 16 kHz and quantized by 16 bits. The beginning of speech signal was detected from analysis of energy and confirmed manually. For FFT, 512 data points corresponding to 32 ms of time duration were taken to be a speech frame for short-term analysis. The first frame of each syllable was used as a

phoneme in our study. To each phoneme, Hanning window was applied after pre-emphasis for spectral flattening. MFCC feature vectors of order 13 were then obtained for subsequent processing. Among the 100 tokens for each phoneme, 80 and 20 were used for training and classification test in neural networks, respectively.

Another feature vector of LPCC was also studied for comparison with MFCC. For this purpose, Hamming window over the pre-emphasized speech signal was first applied. After post-processing of bandpass liftering, LPCC of order 18 was obtained by the Durbin's algorithm [16].

In order to see the meaningfulness of MFCC and LPCC for consonants, vector quantization (VQ) was performed by clustering 1,400 feature vectors (14 phonemes from 100 speakers each) into 16 clusters by LBG clustering procedure.

For the training of neural networks, we set the initial learning parameter for the winning neuron and its two adjacent neurons be 0.9 and 0.45, respectively. As training proceeds, the learning parameters were reduced by a factor of 0.99 on each epoch of training.

As for the initial weights, bipolar random values were assigned for SOM, while initial weight vectors were taken from the first tokens of feature vectors in LVQ.

IV. Results and Discussion

4.1 Classification of phonemes by VQ

To see whether MFCC is ever meaningful for consonants or not, we first investigated clustering of the feature vectors. If the Fourier and subsequent cepstral analyses have validity for consonants, most of which do not show periodic behavior in time, then the result of clustering would manifest such feature naturally.

Table 2 shows VQ result for MFCC, which grouped 14 phonemes into 16 clusters on the basis of Euclidean distance measure. For each phoneme, 100 vectors pronounced by 100 speakers each are distributed over 16 clusters.

Table 2. VQ result for MFCC

Cluster	/g/	/n/	/d/	/r/	/m/	/b/	/s/	/J/	/C/	/k/	/t/	/p/	/h/	/a/
C16	3	2	19	5	0	38	0	0	0	0	0	2	0	9
C15	0	1	1	7	6	28	0	0	0	0	0	0	0	12
C14	0	0	0	35	0	1	0	0	0	0	0	0	0	0
C13	0	0	0	1	0	6	0	0	0	0	0	1	1	76
C12	0	19	2	12	9	0	0	0	0	0	1	1	0	0
C11	0	25	0	3	44	1	0	0	0	0	0	0	0	0
C10	0	7	0	27	6	2	0	0	0	0	0	0	0	0
C9	0	46	0	3	35	0	0	0	0	0	0	0	0	0
C8	3	0	42	4	0	20	1	0	0	1	9	38	8	0
C7	74	0	1	0	0	0	0	0	0	74	1	1	1	3
C6	10	0	33	2	0	1	1	0	1	9	79	14	0	0
C5	7	0	1	1	0	3	0	0	0	6	7	43	90	0
C4	3	0	1	0	0	0	34	5	3	10	3	0	0	0
C3	0	0	0	0	0	0	52	18	10	0	0	0	0	0
C2	0	0	0	0	0	0	10	39	33	0	0	0	0	0
C1	0	0	0	0	0	0	2	38	53	0	0	0	0	0

Phoneme

The most crowded cluster for each column of phoneme is marked by small-dotted pattern and is identified as representing that phoneme. For phoneme /t/, as an example, 79 among 100 vectors fall into the 6-th cluster 'C6'. Thus the cluster C6 is associated with phoneme /t/.

Fig. 2 shows the array of 16 clusters and their associated phonemes. The box of the bottom left corner represents the first cluster C1 and the cluster indices increase counterclockwise. The labels within the boxes denote the corresponding phonemes according to Table 2. For /s/, e.g., 52 vectors among 100 are gathered at the 3rd cluster C3 and hence the 3rd box has label /s/ in it.

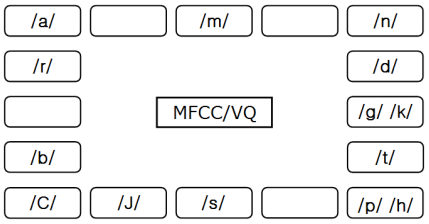

Fig. 2 Array of 16 clusters and associated phonemes formed by LBG/VQ for MFCC

We see from Fig. 2 (and Table 2, equivalently) that the voiced stop /g/ and the unvoiced stop /k/ fall into the same cluster C7 and so do the unvoiced stop /p/ and the whisper /h/ into C5. A minor excuse is that there's a tendency for many speakers to pronounce the stops /g/ and /k/ quite similarly in our experiments due to dialectal reason.

It is not possible either to separate the two pairs of phonemes by considering winner and runner-ups together. For example, as can be seen from Table 2, if we consider the runner-up for the phoneme /p/, then the cluster C8 has two phonemes /d/ and /p/ in association. We also tried to resolve the pairs of phonemes by increasing the number of clusters but failed.

The probability of correct classification for consonants (with the exclusion of the vowel /a/) was found to be around 45%. This result was based on the loose criterion that /g/ and /k/ share the same single cluster, and so do /p/ and /h/.

Since 10 phonemes excluding the two pairs [/g/ and /k/] and [/p/ and /h/] are classified as distinct and the recognition accuracy is about half, it is quite a bit difficult to draw a clear-cut conclusion on the meaningfulness of MFCC for consonants. We might say in a fuzzy fashion: MFCC for consonants is half meaningful and half not.

For LPCC, the same procedure yielded the result of Fig. 3. This time, a little worse result was obtained in that, in addition to the two pairs of the MFCC case, another pair of phonemes, i.e., voiced stop /d/ and unvoiced stop /t/ are not distinguishable by LBG/VQ scheme.

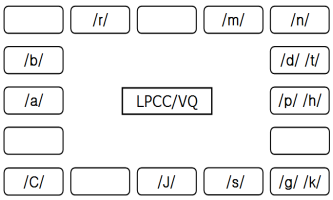

Fig. 3 Array of 16 clusters and associated phonemes formed by LBG/VQ for LPCC

Since LBG/VQ is usually not intended for recognition through iterative training, we content ourselves with finding that two or three pairs of consonants are not resolved by LBG/VQ and the overall error for classification is about 50% for both MFCC and LPCC. From the results, we infer that the meaningfulness of the two types of feature vectors is somewhat doubtful.

4.2 Classification of phonemes by SOM

Fig. 4 shows weight changes of SOM as training proceeds. The weights converge so rapidly that 30 epochs were found to be sufficient for settling.

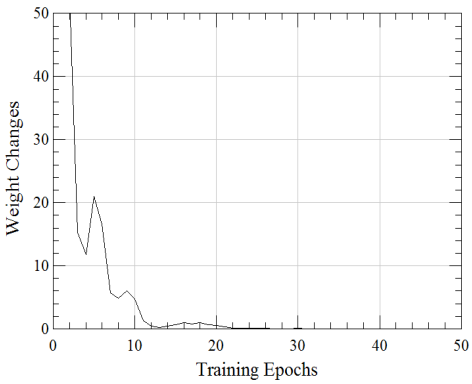

Fig. 4 Weight changes of SOM in training

Fig. 5 shows the training result of phoneme classification by SOM. Darkness of the box denotes relative crowdedness of members belonging to that class. The darker, the more crowded. The darkest neuron, that is the most crowded one, on each column of phoneme label is distinguished by diagonal cross-hatching. For example, the most crowded neuron index for the phoneme /g/ is the 13-th one and it is marked by cross-hatching. The phoneme /g/ is thus assigned to the 13-th output neuron.

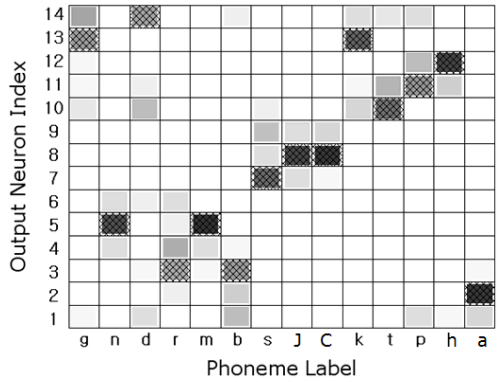

Fig. 5 Training result of phoneme classification by SOM

Since SOM works in unsupervised mode, the phoneme assignment to each output neuron is not given before training. After training is completed, the most crowded neuron for each phoneme is associated with that phoneme. In this way, we set the output neuron index of the most frequent members as its class. Fig. 6 shows the arrangement of the output neurons with the assigned phonemes. The box of the bottom left corner represents the first neuron and the indices increase counterclockwise. This sort of phonotopic map was also obtained for Finnish phonemes by Kohonen et al [6]. Our result is distinguished in that our mapping is one-dimensional with both ends connected and semi-complete for Korean consonants.

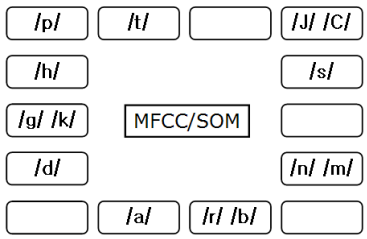

Fig. 6 Arrangement of the 14 phonemes over SOM output neurons with MFCC input vectors

In case of SOM/MFCC, four pairs of phonemes share the same classes and hence are not resolved by this scheme. The results are enumerated in Table 3.

Table 3. Pairs of phonemes sharing the same neuron

Pair of Phonemes	Acoustic Nature	Neuron Index
/r/ and /b/	Glide & Stop	3
/n/ and /m/	Nasals	5
/J/ and /C/	Affricates	8
/g/ and /k/	Stops	13

Fig. 7 is the result for LPCC. We see that a little worse situation occurs for LPCC compared to MFCC, in that 5 pairs are not resolved this time.

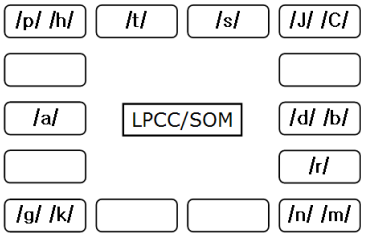

Fig. 7 Arrangement of the 14 phonemes over SOM output neurons with LPCC input vectors

A plausible strategy for resolving the troublesome pairs might be to increase the number of output neurons. Table 4 shows the number of unresolved pairs of phonemes as we vary the number of output neurons. We see that this prescription does not work for resolving the pairs.

Table 4. Number of unresolved pairs of phonemes for LPCC/SOM vs. the number of output neurons

Number of Output Neurons	Number of Unresolved Pairs of Phonemes
14	4
15	5
16	3
17	5
18	3
19	4
20	3

We conclude that SOM/MFCC and SOM/LPCC schemes are not satisfactory in classifying consonants and turn to LVQ, a supervised version with the same architecture as SOM.

4.3 Classification of phonemes by LVQ

LVQ shares many features with SOM. The main difference is that, in LVQ, the target value is assigned to each input vector and the training is performed in supervised mode. The input vector and target value for each phoneme are MFCC and the phoneme index, respectively. Therefore, each output neuron is assigned to a phoneme in one-to-one correspondence beforehand. Fig. 8 shows the output neuron array, arranged in Korean alphabetical order, which is set irrespective of feature vector types.

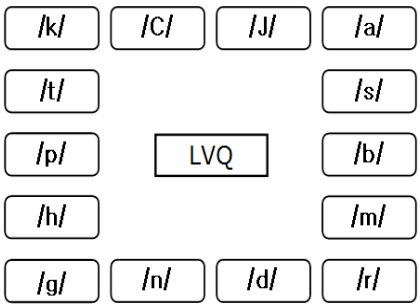

Fig. 8 Array of output neurons for LVQ

Fig. 9 shows LVQ training result for the neuron weights with input vectors of MFCC and constant learning parameter of $\alpha = 0.007$.

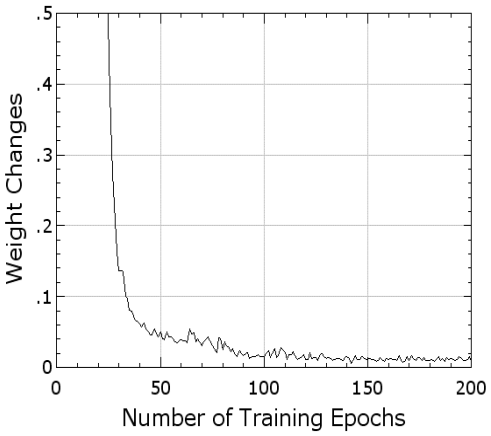

Fig. 9 Training result for MFCC/LVQ

After training of the net by speeches from 80 people is completed, we tested classification accuracy for 20 people. The result is given in Table 5.

Table 5. Classification accuracy of LVQ

Feature Type	Classification Accuracy
MFCC	66%
LPCC	50%

This result is reminiscent of the result of Yamada et. al. [7] who used language source models for phoneme recognition. Considering that their experiment was performed in speaker-dependent mode on a single male speaker for all the phonemes while our MFCC/LVQ result was obtained in speaker-independent mode for 100 people for consonants, the recognition accuracy of ours, 66% for MFCC/LVQ, is not inferior to their 73.2% of phoneme recognition rate without the syllable trigram.

For the case of MFCC/LVQ, dominant errors are given in Table 6.

Table 6. Dominant classification errors for MFCC/LVQ

Tested Phoneme	Recognized Phoneme	Frequency of Error	Partial Sum
/J/	/C/	10	19
/C/	/s/	4	
/J/	/s/	3	
/s/	/C/	2	
/g/	/k/	8	13
/k/	/g/	5	
/d/	/b/	6	11
/b/	/d/	5	
/h/	/p/	4	7
/p/	/h/	3	
/n/	/m/	1	5
/m/	/n/	4	

The following groups incur the dominant errors: [/s/ /J/ /C/] [/g/ /k/] [/d/ /b/] [/p/ /h/] [/n/ /m/] This is in accord with the result of SOM in Fig. 6.

All of the results from our experiment might be summarized as follows: the following phonemes were not found to be easy to classify by neural networks of SOM or LVQ with feature vectors of MFCC or LPCC. This implies that auxiliary information such as zero-crossing-rate (ZCR) should be used in parallel with MFCC in realization of phonetic typewriter that adopts LVQ, like the work by Lin et al [17].

Table 7. Hard-to-resolve pairs of phonemes

Pair of Phonemes	Acoustic Nature
/g/ and /k/	Stops
/n/ and /m/	Nasals
/J/ and /C/	Affricates

V. Conclusion

Classification of consonants was investigated by vector quantization via clustering and neural networks of SOM and LVQ. In order to see whether Fourier and cepstral analyses are meaningful for consonants, vector quantization via LBG clustering was first performed. From the result, it was found to be difficult to draw a clear-cut conclusion about the meaningfulness of MFCC and LPCC for consonants.

In classifying the phonemes by SOM, we found that some pairs of phonemes are not resolved. Our result is distinguishable from the earlier works in that all the Korean consonants except fortes were mapped phonotopically by an one-dimensional array.

As another neural network, we considered LVQ which is trained in supervised mode. The accuracies of classification were only 66% and 50% for MFCC and LPCC, respectively, signifying that other types of feature vectors than MFCC are required as long as consonant classification by LVQ is concerned. This result might be compared to the earlier work by Yamada et al who used language source models for phoneme recognition. Considering that their experiment was performed in speaker-dependent mode on a single male speaker for all the phonemes while our MFCC/LVQ result was obtained in speaker-independent mode for 100 people for consonants, our recognition accuracy for MFCC/LVQ might not be considered to be inferior to their performance without the syllable trigram.

Most of earlier works have provided recognition error rates but our result showed which phonemes are hard to resolve pairwise. This information is useful in that auxiliary speech feature such as zero-crossing-rate (ZCR) might be employed in a restrictive fashion to discriminate only such pairs.

The present work was performed on the first speech frame of 32ms time duration. Our next research will be done on the more speech frames encompassing the whole consonant portion of speech.

References

[1] Kaplan, G. "Words Into Action I," IEEE

Spectrum, Vol. 17, pp. 22-26, 1980.
[2] Davis, K. H., Biddulph, R., and Balashek, S., “Automatic Recognition of Spoken Digits,” J. Acoust. Soc. Am., Vol. 24, No. 6, pp. 637-642, 1952.
[3] Kohonen, T. Self-organization and Associative Memory, 3rd ed., Springer-Verlag, Berlin, 1989..
[4] Olson, H. F. and Belar, H., "Phonetic Typewriter," ITE Trans. on Audio, Vol. 5, No. 4, pp. 90-95, 1957.
[5] Kohonen, T. "The Neural Phonetic Typewriter," Computer, Vol. 21, No. 3, pp. 11-22, 1988.
[6] Kohonen, T. et al, "Phonetic Typewriter for Finnish and Japanese," ICASSP-88, Vol. 1, pp. 607-610, 1988.
[7] Yamada, T., Hanazawa, T., and Kawabata, T. . "Phonetic Typewriter Based on Phoneme Source Modeling," ICASSP-91, Vol. 1, pp. 169-172, 1991.
[8] Kohonen, T.,"Workstation-Based Phonetic Typewriter," Neural Networks for Signal Processing, pp. 279-288, 1991.
[9] Waibel, A. et al,. "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 37, No. 3, pp. 328-339, 1989.
[10] Picone, J. W., "Signal Modeling Techniques in Speech Recognition." Proc. IEEE, Vol. 81, No. 9, pp. 1215-1247, 1993.
[11] Haykin, S. (1999). Neural Networks (2nd Ed.), Prentice Hall, pp. 443-479, 1999.
[12] Kohonen, T., "Improved Versions of Learning Vector Quantization," International Joint Conference on Neural Networks, Vol. 1, pp. 545-550, 1990.
[13] Fausett, L., Fundamentals of Neural Networks, Prentice Hall, pp. 187-194, 1994.
[14] Rabiner, L. & Juang, B., undamentals of Speech Recognition, Prentice Hall, pp. 20-37, 1993.
[15] Deller, J. R., Proakis, J. G., & Hansen, J. H. L. Discrete-Time Processing of Speech Signals, Macmillan, pp. 117-137, 1993.
[16] Durbin, J., "The Fitting of Time Series Models," Review of the Institute for International Statistics, Vol. 28, pp. 233-243, 1960.
[17] Lin, H. & Ou, Z. "Switching Auxiliary Chains for Speech Recognition," IEEE Signal Processing Letters, Vol. 14, No. 8, pp. 568-571, 2007.

저자 소개

이채봉(Chai-bong Lee)

1985년 동아대학교 전자공학과 졸업(공학사)
1988년 동북대학교 대학원 전기통신공학과 졸업(공학석사)
1992년 동북대학교 대학원 전기통신공학과 졸업 (공학박사)
1993～현재 동서대학교 전자공학과 교수
※ 관심분야 : 신호처리, 음향공학

이창영(Chang-young Lee)

1982년 서울대학교 물리교육학과 졸업(이학사)
1984년 한국과학기술원 물리학과 졸업(이학석사)
1992년 뉴욕주립대학교(버펄로) 물리학과 졸업 (이학박사)
1993년～현재 동서대학교 시스템경영공학과 교수
※ 관심분야 : 음성인식, 화자인식