음성인식에서 중복성의 저감에 대한 연구

이창영[*]

A Study on the Redundancy Reduction in Speech Recognition

Chang-Young Lee[*]

요 약

음성 신호의 특성은 인접한 프레임에서 크게 변화하지 않는다. 따라서 비슷한 특징벡터들에 내재된 중복성을 줄이는 것이 바람직하다. 본 논문의 목적은 음성인식에 있어서 음성 특징벡터가 최소의 중복성과 최대의 유효한 정보를 갖는 조건을 찾는 것이다. 이를 이하여 우리는 하나의 감시 파라미터를 통하여 중복성 저감을 실현하고, 그 결과가 FVQ/HMM을 사용한 화자독립 음성인식에 미치는 영향을 조사하였다. 실험 결과, 인식률을 저하시키지 않고 특징벡터의 수를 30% 줄일 수 있음을 확인하였다.

ABSTRACT

The characteristic features of speech signal do not vary significantly from frame to frame. Therefore, it is advisable to reduce the redundancy involved in the similar feature vectors. The objective of this paper is to search for the optimal condition of minimum redundancy and maximum relevancy of the speech feature vectors in speech recognition. For this purpose, we realize redundancy reduction by way of a vigilance parameter and investigate the resultant effect on the speaker-independent speech recognition of isolated words by using FVQ/HMM. Experimental results showed that the number of feature vectors might be reduced by 30% without deteriorating the speech recognition accuracy.

키워드

Speech Recognition, Redundancy Reduction, MFCC, Hidden Markov Model
음성인식, 중복성 저감, MFCC, 은닉 마르코브 모델

I. Introduction

The state of the art in the field of speech recognition has now reached such a level of performance and robustness, even in the noisy environment, that permits lots of applications. As a result, we are now living in a world of various devices which deploy the relevant technology [1-2].

Redundancy is unavoidably inherent in any kind of information. There are four major forms of redundancy: hardware redundancy, information redundancy, time redundancy, and software redundancy. In some applications, redundancy is utilized to improve the system performance [3-6]. For example, differences in perception between natural speech and high-quality synthetic speech is

* 동서대학교 시스템경영공학과(seewhy@dongseo.ac.kr)

접수일자 : 2012. 04. 18 심사(수정)일자 : 2012. 04. 30 게재확정일자 : 2012. 06. 07

inferred to be due to the redundancy of the acoustic-phonetic information encoded in the speech signal and hence redundancy might not be better removed for natural speech [7]. There are other cases, on the other hand, that redundancy reduction enhances the relevant performance [8]. The ultimate goal in regards to the redundancy is to pursue both maximum relevance and minimum redundancy (MRMR) at the same time [9].

There are several reasons for the redundancy reduction, which might be enumerated as follows. Firstly, redundancy reduction can help to decrease the search space for goal-directed learning procedures in pattern recognition. It can sometimes help to achieve enormous learning speed-ups. Secondly, redundancy reduction promises to simplify statistical classifiers. Thirdly, redundancy reduction allows for data compression and storage reduction. As an example, in distributed speech recognition (DSR) [10-11], the need for redundancy reduction is demanding for an efficient way of translating automatic speech recognition technologies to lightweight mobile device such as PDA.

The subject of redundancy reduction can be found in all the fields of pattern recognition including face recognition [12-13], shape recognition. speaker recognition [14-15], image and 3D video compression [16], speech emotion recognition [17-18].

As for the field of speech recognition, there are many sources of redundancy and as many routes to redundancy reduction have been developed. One approach is focused on the recognition tool such as neural network [19-20] and hidden Markov model (HMM) [21-22]. Another method delves into developments of algorithms such as subspace Gaussian mixture [23], Gabor transform [24], nonlinear resampling transformation [25], utilization of probability density function (PDF) [26], independent component analysis (ICA) [27], to name a few. However, above all, the most common and usual approach for the redundancy reduction is performed in conjunction with the signal processing [28-30].

The phonetic contents of speech signal overlap more or less, and independent treatment of the constituent sounds would be inevitably redundant regardless of their identifiable similarities. The usual procedure is to reduce or remove redundancy in the front-end processing. For MRMR, the digitized speech is compressed before transmitting into the back-end processing. Specifically, feature frames are chosen to have minimum redundancy within selected feature frames but keeping maximum relevancy [31-32].

The adjacent frame vectors usually show similarity in the feature space because of the slow movements of the articulators. Hence efficient frame selection techniques to select non-redundant frames in the preprocessing stage will be very effective in real time application of the recognition system. The motivation of this paper is that the redundancy in vowel is relatively strong. From this observation, we select a new smaller sequence of frames from the original frames, thereby achieving the redundancy reduction between consecutive frames. The aim is not only to reduce the number of frames for feature extraction but also to maintain the recognition accuracy reasonably high by selecting suitable frames. In this way, our method not only reduces the number of frames but also prevents from deteriorating the recognition accuracy.

The organization of this paper is as follows. Section Ⅱ provides description of our method that selectively choose the feature vectors. Section Ⅲ describes details of the experiments performed in our study. In section Ⅳ, experimental results are presented to demonstrate the efficacy of the proposed method. Concluding remarks are given in section Ⅴ finally.

II. Elimination of Redundancy

Figure 1 shows a partial waveform of a phoneme /a/ pronounced by a young female speaker. Three frames labelled are each of length of 512 data points corresponding to 32ms of time duration. The adjacent frames are overlapped by 50%, which is necessary in order not to lose any information contents during coarticulation. The three frames look similar barring shifts of data in time relative to each other.

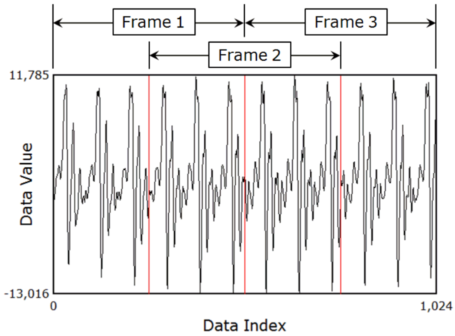

Fig. 1 A partial waveform of a phoneme /a/ pronounced by a young female speaker.

Figure 2 shows mel-frequency cepstral coefficients (MFCC) of order 13 without cepstral mean subtraction (CMS) [33], which are extracted for the three consecutive frames of Figure 1. Considering relative data shifts of the three frames of Figure 1, the close features of the result reflects the shift-invariant nature and strength of MFCC [34]. Anyhow, the results for the three consecutive frames are so close that they are practically undiscernible. This result is suggestive of the necessity that the three frames might not be treated distinctly in subsequent speech processing. Rather, it is preferable to consider only one of the three frames as a representative one. This is the motivation of our study in this paper.

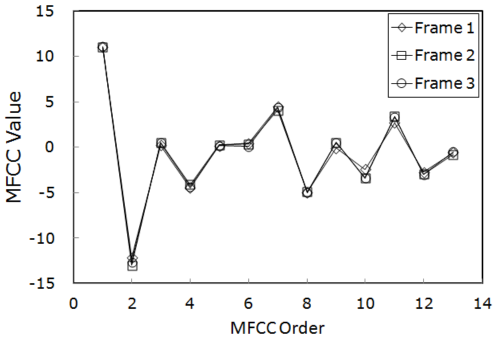

Fig. 2 MFCC components for the three consecutive frames of figure 1. the three feature vectors are so close that they are almost undiscernible.

The question is how to detect similar frames and determine whether to discard some of the similar frames. For this end, we first consider the distance of the two MFCC feature vectors

$$d_{ij} \equiv \| \vec{v_i} - \vec{v_j} \| \tag{1}$$

where $\| \cdot \|$ denotes the norm of a vector. An intuitive approach is to consider the criterion

$$d_{ij} < \epsilon \tag{2}$$

where ϵ is some constant. If this criterion is met, then we consider the two vectors as similar and discard one of the two vectors.

The problem in this prescription is that the consonants, compared to the vowels, have very low energy that the magnitudes of the feature vectors are correspondingly small. Therefore it is not easy to choose ϵ in such a way to perform well for vowels and consonants simultaneously.

In order to remedy this problem, we consider a different approach by introducing a relative distance of the consecutive two vectors. Instead of (2), we examine the condition

$$\frac{\| \vec{v}_{j+1} - \vec{v_j} \|}{\| \vec{v_j} \|} < \alpha \tag{3}$$

If this condition is true, the vectors $\vec{v}_{j+1}$ is considered as similar to its preceding vector $\vec{v}_j$ and will be discarded for redundancy reduction. This concept is illustrated in Figure 3. The vector $\overrightarrow{OA}$ is the reference vector and the radii of the two circles are $\| \overrightarrow{OA} \|$ and $\alpha \| \overrightarrow{OA} \|$, respectively. The vector $\overrightarrow{OB}$ is determined to be similar to the vector $\overrightarrow{OA}$ while the vector $\overrightarrow{OC}$ is not.

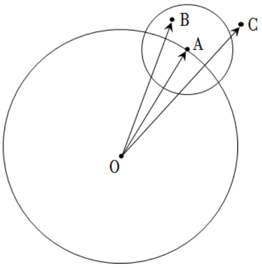

Fig. 3 Illustration of the concept of similarity of the vectors proposed in this paper.

Figure 4 shows the flowchart of selecting and discarding similar vectors following the extraction of feature vectors. By investigating the condition (3) for the two consecutive vectors, similar vectors are excluded in the subsequent processing. Even if a vector is discarded, it acts as a reference vector for the next consecutive vector.

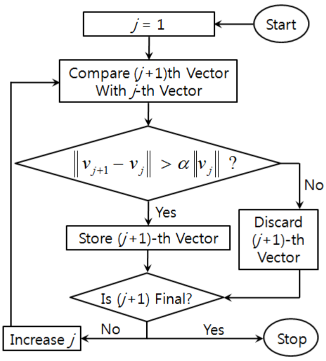
Start
j = 1
Compare (j+1)th Vector With j-th Vector
$\|v_{j+1} - v_j\| > \alpha\|v_j\|$?
Yes
No
Store (j+1)-th Vector
Discard (j+1)-th Vector
Is (j+1) Final?
No
Yes
Increase j
Stop

Fig. 4 The flowchart of the algorithm proposed in this paper.

Large α means considering more vectors as similar and hence there are less distinct vectors. In this sense, vigilance parameter α might be called as similarity factor. Then, the optimum features are selected according to the minimum redundancy and maximum relevance (MRMR) strategy.

III. Experiment

Our experiments were performed on a set of phone-balanced 300 Korean words. To see the effect of vocabulary size also, we divided the words into three sets as in Table 1. The sets A and B are disjoint each other and C is the union of them.

Table 1. Three sets of speech data divided for studying the effect of the vocabulary size.

Word Set	Number of Words
A	100
B	200
C	300

Forty people including 20 males and 20 females participated in speech production. Speech utterances of them were divided into three disjoint groups as in Table 2.

Table 2. Division of the 40 people's speech production into three groups.

Speaker Group	Number of People
I	32
II	4
III	4

Thirty-two people's speech tokens of the group I were used in generating codebook of size 512, whose centroids serve for fuzzy vector quantization (FVQ) of all the speeches of 40 people. HMM parameters were updated on each iteration of training. In order to choose which values of parameters to use in the final test of speech recognition, some test speeches are necessary. The parameters that yield the best performance on the group Ⅱ were stored and used for the test on the group Ⅲ to obtain the final performance of the speaker-independent speech recognition system. This prescription prevents the system from falling too deep into the local minimum driven by the training samples of the group I and hence becoming less robust against the speaker-independence when applied to the group Ⅲ [35].

The speech utterances were sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32ms of time duration were taken to be a speech frame for short-term analysis. The vigilance parameter α were varied from 0 to 0.5 in steps of 0.02. The value of $\alpha = 0.5$ means that, for a given reference vector of radius R, the next vector within hypersphere of radius $0.5R$ with center at the tip of the reference vector, is treated as similar and hence discarded.

To each frame, Hanning window was applied after pre-emphasis for spectral flattening. MFCC feature vectors of order 13 were obtained and then cepstral mean subtraction were applied on utterance basis to endow robustness against various adverse effects such as system dependence and noisy environment. The similarity of the vectors were looked for and redundancy was removed by discarding similar vectors.

Codebooks of 512 clusters were generated by the Linde-Buzo-Gray clustering algorithm on the MFCC feature vectors obtained from the speeches of the group I of Table 2. The distances between the vectors and the codebook centroids were calculated and sorted. Appropriately normalized fuzzy membership values were assigned to the nearest two clusters and a train of two doublets (cluster index / fuzzy membership) were fed into HMM for speech recognition processing.

For the HMM, a non-ergodic left-right (or Bakis) model was adopted. The number of states that is set separately for each class (word) was made proportional to the average number of frames of the training samples in that class [36]. Initial estimation of HMM parameters $\lambda = (\pi, A, B)$ was obtained by K-means segmental clustering after the first training. By this procedure, convergence of the parameters became so fast that enough convergence was reached mostly after several epochs of training iterations.

Backward state transitions were prohibited by suppressing the state transition probabilities a_{ij} with $i > j$ to a very small value but skipping of states was allowed. The last frame was restricted to end up with the final state associated with the word being scored within a tolerance of 3.

Parameter reestimation was performed by Baum-Welch reestimation formula with scaled multiple observation sequences to avoid machine-errors that might be caused by repetitive multiplication of small numbers. After each iteration, the event observation probabilities $b_i(j)$ were boosted above a small value.

Three features were monitored while training the HMM parameters: (1) the recognition error rate for the group Ⅱ of Table 2, (2) the total probability likelihood of events summed over all the words of the training set according to the trained model, and (3) the event observation probabilities for the first state of the first word in the vocabulary list. Training was terminated when the convergences for these three features were thought to be enough. The parameter values of $\lambda = (\pi, A, B)$ that give the best result for the group Ⅱ were stored and used in speech recognition test on the group Ⅲ of Table 2.

Ⅳ. Results and Discussion

Figure 5 shows the ratio of the number of feature vectors $N(\alpha)/N(0)$ for the set B of Table 1 (words 200). For α below around 0.1, $N(\alpha) \approx N(0)$ which means that almost all the feature vectors are treated as distinct. For $\alpha > 0.1$ $N(\alpha)/N(0)$ decreases monotonically with a little curvature. This result is almost the same in the cases of the set B and C of Table 1

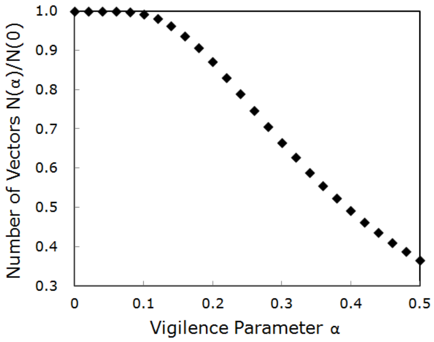

Fig. 5 The ratio of the number of feature vectors $N(\alpha)/N(0)$ for the set B of table 1 vs. the vigilance parameter α.

Figure 6 is the speech recognition result as the vigilance parameter α is varied. The general behavior might be phrased in terms of two stages, one for little change (with minor fluctuations) and the other for the approximately linear increase in the recognition error rate. This feature were found to pervade all the cases under our study.

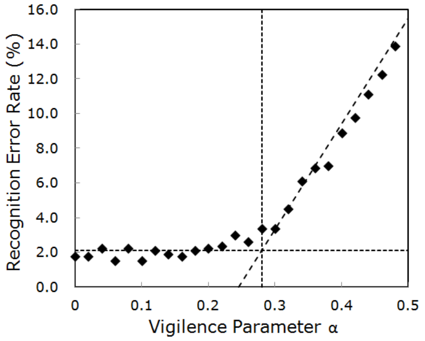

Fig. 6 Recognition error rate vs. the vigilance parameter α.

By two separate curve-fittings on the two regions, the optimal vigilance parameter was located as the abscissa coordinate of the intersection of the two fitted lines. Table 3 shows the summary of the results.

Table 3. Optimal values of the vigilance parameter α and redundancy reduction rate $N(\alpha)/N(0)$ for three vocabulary sizes.

# of Words	Optimal Condition	
	Vigilance Parameter α	$N(\alpha)/N(0)$
100	0.29	≈0.7
200	0.28	≈0.7
300	0.24	≈0.8

By an appropriate choice of the vigilance parameter, we might be able to find the features with the highest amount of information relevant for the recognition task, and at the same having minimal redundancy [37].

Our result might be compared with the result of Weng et al. [22] who reported a compression factor of 1.5 without any loss in recognition accuracy. also showed that the acoustic model size can be reduced by 8% with almost the same performance as the standard acoustic modeling [23]. Our result is superior to the result of Bouallegue et al. and comparable to the work of Weng et al. in regards to the degree of redundancy reduction.

V. Conclusion

In this paper, experimental search for the optimal condition of maximum relevancy and minimum redundancy of the speech signal was performed through a vigilance parameter. The aim is to reduce the redundancy that would not give rise to signifiant adverse effect on the speech recognition. Experiments were performed on 100/200/300 isolated words produced by 40 people.

The concept of relative distance of two vectors was introduced and a criterion was established which determines similarity of the vectors. As the value of the vigilance parameter is increased, experimental results showed that the number of frames to process decreases monotonically with a little curvature

Speech recognition performance showed largely two stages of changes as the value of the vigilance parameter is increased: one is minor fluctuation in the small value of the parameter and the other is roughly linear increase in recognition error rate in the large value of the vigilance parameter. Optimal value of the vigilance parameter was located by the point of intersection for curve-fittings in the two regions.

Corresponding optimal values of vigilance parameter were found to be around 0.3 and the corresponding number of vectors for processing were reduced by around 30%. The dependence of the results on the vocabulary size was minor.

References

[1] Y. Chang, S. Hung, N. Wang, & B. Lin, "CSR: A Cloud-assisted speech recognition service for personal mobile device", International Conference on Parallel Processing (ICPP), pp. 305-314. 2011.

[2] 김범준, “와이브로 네트워크를 통한 음성 서비스의 측정 기반 품질 기준 수립”, 한국전자통신학회논문지, 6권, 6호, pp. 823-829, 2011.

[3] Spiro, G. Taylor, G. Williams, & C. Bregler, “Hands by hand: Crowd-sourced motion tracking for gesture annotation”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 17-24. 2010.

[4] W. Sun, Z. Wu, H. Hu, & Y. Zeng, "Multi-band maximum a posteriori multi-transformation algorithm based on the discriminative combination", International Conference on Machine Learning and Cybernetics, Vol. 8, pp. 4876-4880. 2005.

[5] H. R. Tohidypour, S. A. Seyyedsalehi, H. Roshandel, & H. Behbood, "Speech recognition using three channel redundant wavelet filterbank", 2nd International Conference on Industrial Mechatronics and Automation (ICIMA), Vol. 2, pp. 325 - 328. 2010.

[6] M. Paulik & A. Waibel, "Spoken language translation from parallel speech audio: Simultaneous interpretation as SLT training data", IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5210-5213. 2010.

[7] D. B. Pisoni, H. C. Nusbaum, & B. G. Greene, "Perception of synthetic speech generated by rule", Proceedings of the IEEE, Vol. 73, No. 11, pp. 1665-1676. 1985.

[8] S. Alizadeh, R. Boostani, & V. Asadpour, "Lip feature extraction and reduction for HMM-based visual speech recognition systems", 9th International Conference on Signal Processing (ICSP), pp. 561-564. 2008.

[9] V. Estellers, M. Gurban, & J. P. Thiran, "Selecting relevant visual features for speechreading", IEEE International Conference on Image Processing (ICIP), pp. 1433 - 1436. 2009.

[10] Z. Tan, P. Dalsgaard, & B. Lindberg, "Adaptive Multi-Frame-Rate Scheme for Distributed Speech Recognition Based on a Half Frame-Rate Front-End", IEEE 7th Workshop on Multimedia Signal Processing, pp. 1-4. 2005.

[11] V. Sanchez, A. M. Peinado, J. L. Perez-Cordoba, "Low complexity channel error mitigation for distributed speech recognition over wireless channels", EEE International Conference on Communications, Vol. 5, pp. 3619-3623. 2003.

[12] S. M. Lajevardi & Z. M. Hussain, "Contourlet structural similarity for facial expression recognition", IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 1118-1121. 2010.

[13] T. Kim, H. Kim, W. Hwang, S. Kee, & J. Kittler, "Independent component analysis in a facial local residue space", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 579-586. 2003.

[14] S. van Vuuren, "Comparison of text-independent speaker recognition methods on telephone spe-

ech with acoustic mismatch", Fourth International Conference on Spoken Language, Vol. 3, pp. 1788-1791. 1996.

[15] C. Jung, M. Kim, & H. Kang, "Normalized minimum-redundancy and maximum-relevancy based feature selection for speaker verification systems", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4549 - 4552. 2009.

[16] L. Granai, T. Vlachos, M. Hamouz, J. R. Tena, & T. Davies, "Model-Based Coding of 3D Head Sequences", 3DTV Conference, pp. 1-4. 2007.

[17] T. S. Tabatabaei & S. Krishnan, "Towards robust speech-based emotion recognition", IEEE International Conference on Systems Man and Cybernetics (SMC), pp. 608-611. 2010.

[18] L. Xu, M. Xu, & D. Yang, "Factor Analysis and Majority Voting Based Speech Emotion Recognition", International Conference on Intelligent System Design and Engineering Application (ISDEA), Vol. 1, pp. 716-720. 2010.

[19] M. D. Emmerson, & R. I. Damper, "Relations between fault tolerance and internal representations for multi-layer perceptrons", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, pp. 281-284. 1992.

[20] 최재승, “신경회로망에 의한 음성 및 잡음 인식 시스템”, 한국전자통신학회논문지, 5권, 4호, pp. 357-362, 2010.

[21] P. Nguyen, L. Rigazio, C. Wellekens, & J.-C. Junqua, "Construction of model-space constraints", IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 69-72, 2001.

[22] J. Weng & X. Jia, "A Memory-Efficient Graph Structured Composite-State Network for Embedded Speech Recognition", Fifth International Conference on Natural Computation (ICNC), Vol. 3, pp. 570-573. 2009.

[23] M. Bouallegue, D. Matrouf, & G. Linares, "A simplified Subspace Gaussian Mixture to compact acoustic models for speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4896-4899. 2011.

[24] P. Min & S. Yihe, "ASIC design of Gabor transform for speech processing", 4th International Conference on ASIC, pp. 401-404. 2001.

[25] Y. D. Liu, Y. C. Lee, H. H. Chen, & G. Z. Sun, "Nonlinear resampling transformation for automatic speech recognition", Neural Networks for Signal Processing, pp. 319-326. 1991.

[26] G. Sarkar & G. Saha, "Efficient pre-quantization techniques based on probability density for speaker recognition system", IEEE Region 10 Conference (TENCON), pp. 1-6. 2009.

[27] H. Hsieh, J. Chien. K. Shinoda, & S. Furui, "Independent component analysis for noisy speech recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4369-4372. 2009.

[28] T. Lee & G. Jang, " The statistical structures of male and female speech signals", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, pp. 105-108. 2001.

[29] X. Zhao, P. Yang, & L. Zhang, "Research on the low rate representations for speech signals", 11th IEEE Singapore International Conference on Communication Systems (ICCS), pp. 188-192. 2008.

[30] Y. Yangrui, Y. Hongzhi, & L. Yonghong, "The Design of Continuous Speech Corpus Based on Half-Syllable Tibetan", International Conference on Computational Intelligence and Software Engineering, pp. 1-4. 2009.

[31] C. Jung, M. Kim, & H. Kang, "Selecting Feature Frames for Automatic Speaker Recognition Using Mutual Information", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 6, pp. 1332-1340. 2010.

[32] J. Song, M. Lyu, J. Hwang, & M. Cai, "PVCAIS: a personal videoconference archive indexing system", International Conference on Multimedia and Expo (ICME), Vol. 2, pp. 673-676. 2003.

[33] S. Sadjadi & J. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions", IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP), pp. 5448-5451. 2011.

[34] D. Dimitriadis, P. Maragos, & A. Potamianos, "On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 6, pp. 1504-1516. 2011.

[35] L. Fausett, "Fundamentals of Neural Networks", Prentice-Hall, New Jersey, p. 298. 1994.

[36] M. Dehghan, K. Faez, M. Ahmadi, & M. Shridhar, "Unconstrained Farsi Handwritten Word Recognition Using Fuzzy Vector Quantization and Hidden Markov models," Pattern Recognition Letters, Vol. 22, pp. 209-214. 2001.

[37] T. Drugman, M. Gurban, & J.-P. Thiran, "Relevant Feature Selection for Audio-Visual Speech Recognition", IEEE 9th Workshop on Multimedia Signal Processing (MMSP), pp. 179-182. 2007.

저자 소개

이창영(Chang-Young Lee)

1982년 2월 서울대학교 물리교육학과 졸업(이학사)

1984년 2월 한국과학기술원 물리학과 졸업(이학석사)

1992년 8월 뉴욕주립대학교 (버펄로) 물리학과 졸업(이학박사)

1993년~현재 동서대학교 시스템경영공학과 교수

※ 관심분야 : 음성인식, 화자인식, 신호처리