Journal of KIIT. Vol. 16, No. 3, pp. 11-16, Mar. 31, 2018. pISSN 1598-8619, eISSN 2093-7571 11
http://dx.doi.org/10.14801/jkiit.2018.16.3.11

잡음 중의 화자분류를 위한 켑스트럼 잡음모델에 의한 음성 특징 강조기법

최 재 승*

A Speech Feature Enhancement Technique by Cepstral Noise Model for Noisy Speaker Identification

Jae-Seung Choi*

요 약

최근 배경잡음 환경 하에서 잡음에 강인한 음성의 화자인식의 중요성이 강조되고 있다. 실제 환경에서 배경잡음의 영향이 많은 경우에 음성인식의 성능을 상당히 저하시키기 때문에 음성인식의 실용화를 위해서 잡음에 강인한 음성인식 시스템의 구현이 필요하다. 따라서 본 논문에서는 주파수 영역에서 잡음을 억제하는 위너필터와 스펙트럼 평활화 방법을 사용한 켑스트럼 잡음모델에 의한 음성특징 강조기법을 제안한다. 제안하는 음성특징 강조기법은 잡음이 억제된 MFCC와 LPC에 의한 음성특징량을 추출하여, 오차역전파 신경회로망을 사용하여 잡음으로 오염된 음성의 화자를 분류한다. 본 실험에서는 백색잡음 및 자동차 잡음으로 오염된 음성신호를 사용하여 음성의 화자분류 실험을 수행하였으며, 기존 방법과 비교하여 백색잡음에 대하여는 최대 14.86%, 자동차 잡음에 대해서는 최대 6.36%의 화자 분류율을 개선하였다.

Abstract

Recently, it is emphasized the importance of robust speech recognition and speaker recognition under noisy background environments. In real environments, the performance of the speech recognition is significantly degraded because of the effect of heavy background noises. Thus it is necessary to implement a noise-robust speech recognition system for the practical use of the speech recognition. Therefore, this paper proposes a speech feature enhancement technique by a cepstrum noise model using Wiener filter that suppresses noise in the frequency domain, and a spectrum smoothing method. The proposed technique extracts speech features with noise suppression by MFCC and LPC, and classifies the speakers of noise-contaminated speech by using error backpropagation neural network. In this experiment, speaker classification rates were improved about 14.86% for white noise and 6.36% for car noise compared with conventional methods, after speech classification experiments are performed using speech signals contaminated with white noise and car noise.

Keywords
speech feature enhancement, cepstral noise model, noisy speaker identification, noise subtraction

* 신라대학교 스마트전기전자공학부 교수
- ORCID: http://orcid.org/0000-0002-5699-9701

· Received: Jan. 18, 2018, Revised: Feb. 21, 2018, Accepted: Feb. 24, 2018
· Corresponding Author: Jae-Seung Choi
Div. of Smart Electrical and Electronic Engineering, Silla University, 140 Baegyang-daero(Blvd), 700beon-gil(Rd), Sasang-gu, Busan, 46958 Korea,
Tel.: +82-51-999-5608, Email: jschoi@silla.ac.kr

Ⅰ. 서 론

음성은 정보전달 속도가 빠르고 특별히 도구를 사용할 필요가 없다는 이점을 가진 자연스럽고 중요한 인간의 통신수단이라고 할 수 있다. 따라서 인간과 기계, 특히 컴퓨터와의 대화수단으로써 음성의 정보처리가 주목되고 있으며 그 기대가 크다고 할 수 있다. 이러한 정보처리의 하나로써 인간으로부터 컴퓨터에의 통신수단으로 가능한 것이 음성인식이다. 음성인식의 궁극적인 목표는 일반적인 말들을 화자에 상관없이 바로 인식 가능하게 하며, 이러한 목표대로 음성인식 시스템에 저비용으로 실현된다면 상당히 많은 분야에 응용될 것으로 본다[1].

최근에 여러 방면에서 음성인식 수법이 제안되고 있으며, 음성인식 시스템을 장착한 소프트웨어, 가전제품 등이 상품화되어 음성인식의 실용화가 진행되고 있다. 그러나 이러한 대부분의 음성인식 수법들은 비교적 조용한 환경에서 성능을 테스트하고 있는 경우가 많으며, 실제 환경에서 배경잡음의 영향이 큰 경우에 음성인식 성능이 극단적으로 저하되는 문제가 있기 때문에 완전하게 실용화하는 경우에 문제가 발생하고 있는 상황이다. 이러한 잡음환경 하에서의 음성인식의 수법으로는 은닉마프코프 모델(HMM, Hidden Markov Model), 인공 신경회로망(ANN, Artificial Neural Network), 위너필터(Wiener Filter) 등이 제안되어 그 유효성이 보고되어 있으며, 또한 배경잡음에 강인한 음성인식 시스템을 확립하여 음성인식 시스템의 실용화를 실현하기 위하여 여러 연구가 진행되고 있다[2]-[5].

따라서 본 논문에서는 배경잡음 환경 하에서 잡음에 강인한 음성인식 시스템의 구현을 위하여, 잡음이 중첩된 음성으로부터 잡음성분을 제거하여 깨끗한 음성만을 추출하는 특징 강조기법 및 음성인식 방법을 제안한다. 구체적으로는 주파수 영역에서 잡음을 억제하는 위너필터와 스펙트럼 평활화 방법을 사용한 켑스트럼 잡음모델에 의한 특징 강조 기법을 사용하여, 멜 필터뱅크분석에 기초한 멜 주파수 켑스트럼 계수(MFCC, Mel Frequency Cepstral Coefficient)와 선형예측분석에 의한 선형예측부호화(LPC, Linear Predictive Coding) 켑스트럼 계수의 음성특징 파라미터를 추출한다. 추출된 잡음이 억제된 MFCC와 LPC의 음성특징 파라미터는 오차역전파 인공신경회로망(BPANN, Backpropagation Artificial Neural Network)의 입력으로 사용한다. 본 논문에서는 잡음억제에 의한 음성인식을 구현하기 위하여 BPANN을 사용하여 종래의 선형사상을 사용한 분리수법보다도 높은 성능을 실현할 수 있도록 잡음으로 오염된 음성의 화자를 분류한다[6][7]. 음성의 화자인식 시스템의 성능 평가를 위하여 종래 수법과의 비교를 통하여 본 논문에서 제안한 알고리즘에 대한 성능 비교를 수행한다.

Ⅱ. 제안한 분류 알고리즘

청각 시스템의 지각을 모의하기 위한 일반적인 접근방식은 왜곡된 주파수 척도에서 일률적으로 사상하는 비선형 필터 뱅크를 사용하는 것이다. 이러한 비선형 필터의 대표적인 예로는 MFCC라고 할 수 있다[7][8]. 그림 1은 일반적인 MFCC 계수에 의한 음성 특징량 추출 블록도를 나타낸다. 본 논문에서의 입력 음성신호는 각 프레임에 있어서 인접한 프레임들의 연속성을 증가시키기 위한 식 (1)과 같은 해밍창을 씌운다.

$$W_H(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right) \qquad (1)$$

여기에서 $n = 0,1,\cdots,N-1$ 이며, N은 창의 크기를 나타낸다. 이 후에 각 프레임의 256 샘플(16ms)을 고속 푸리에 변환(FFT, Fast Fourier Transform)을 사용하여 시간영역을 주파수 영역으로 변환한다. 다음으로 FFT에 의해 나타난 스펙트럼의 진폭은 실제 주파수 스케일(Hz)과 지각된 주파수 스케일(Mel) 사이를 맵핑하는 멜 스케일에 의해서 변형된다. 이 후에 MFCC 계수를 계산하기 위하여 멜 스펙트럼 계수에 대하여 로그를 취하면 로그 멜 스펙트럼이 출력되며, 이 결과를 이산 코사인 변환(DCT, Discrete Cosine Transform)하면 멜 주파수 켑스트럼 계수를 구할 수 있다.

MFCC는 잡음이 없는 깨끗한 음성에 대해서 효과적인 특징량이지만 잡음으로 오염된 음성의 경우

에는 음성인식률을 저하시킨다고 알려져 있고, LPC 켑스트럼 계수는 음성신호의 샘플값에 서로 상관관계가 있다는 것을 전제로 하여 음성의 특징량을 추출하는 방법이다[7]-[9].

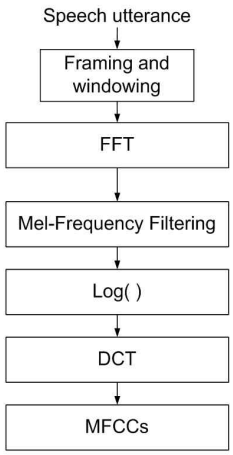

그림 1. 일반적인 MFCC 음성특징량 추출
Fig. 1. Block diagram of the general MFCC feature extraction

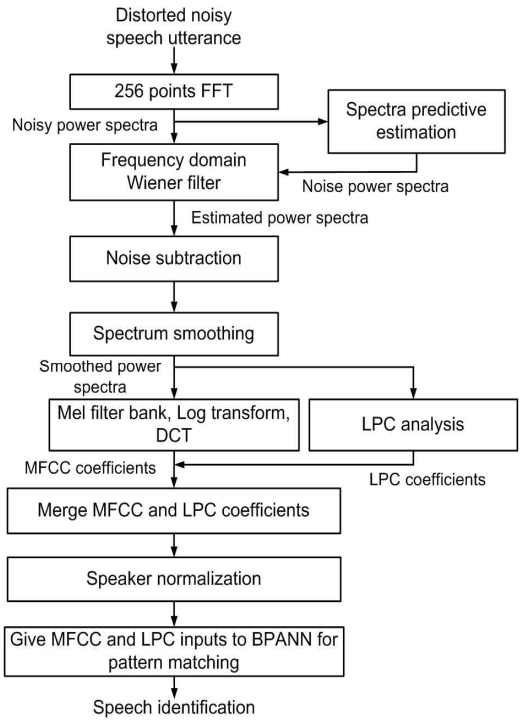

그림 2. 제안한 잡음 억제 특징 추출 방법
Fig. 2. Block diagram of the proposed noise subtraction feature extraction

그림 2에 나타낸 바와 같이 본 논문에서 제안하는 잡음억제 MFCC와 LPC의 특징량은 위너필터에서 추정된 잡음 스펙트럼을 사용하여 각 프레임에서 잡음스펙트럼 억제한 원 음성에 가까운 깨끗한 음성스펙트럼의 특징량을 추출한다. 다음으로 이동평균을 통한 평활화 알고리즘을 적용한다. 이 후에 멜 필터 뱅크, 대수 변환, DCT 과정을 거친 다음에 12차의 MFCC 계수를 구하며, 또한 LPC 분석을 통하여 12차의 LPC 계수를 구한다. 마지막으로 BPANN에서 사용할 수 있도록 정규화 과정을 거치며, MFCC 및 LPC를 통합하여 총 24차의 계수를 BPANN의 입력 패턴으로 하여 음성의 화자인식을 수행한다. MFCC 및 LPCC 계수는 신경회로망에 입력되기 전에 +1과 -1 사이의 값으로 정규화된다. 따라서 이 MFCC 및 LPCC 켑스트럼 계수는 BPANN의 입력층으로 입력되고, 연결강도를 입력층, 중간층, 출력층 순서로 개입시켜 접속시킴으로써 각 유닛의 출력이 계산된다. BPANN 네트워크의 구조는 24-30-3의 네트워크를 사용하여 학습 및 테스트를 수행하였다. 출력으로는 최종적으로 분류할 3명의 화자에 대한 출력을 나타낸다.

본 논문에서 BPANN의 학습은 오차역전파 학습 알고리즘에 기초한 경사하강법을 사용하여 가중치를 조정함으로써 수행된다[6]. 이 학습 알고리즘은 식 (2)와 같이 목표로 하는 출력(D_o)과 뉴런의 실제 출력(Y_o) 사이의 제곱오차의 합이 최소화되도록 연결강도가 적용되어 학습된다. 여기에서 O는 출력 뉴런들의 수를 나타내며, 네트워크의 각 연결강도는 가능한 빨리 오차 E가 줄어들도록 조정된다.

$$E = \frac{1}{2}\sum_{o=1}^{O}(D_o - Y_o)^2 \tag{2}$$

그림 3은 본 논문에서 제안한 화자 분류 알고리즘으로 그림 2의 잡음 억제 음성 특징 추출 알고리즘과 BPANN을 사용하여 화자 분류를 실행하는 블록도를 나타낸다. 그림 3에서 각 프레임에 걸쳐서 256 샘플의 해밍창과 FFT를 거친 후에 위너필터에 의한 잡음억압 및 3 프레임분의 스펙트럼 평활화 과정을 처리한다.

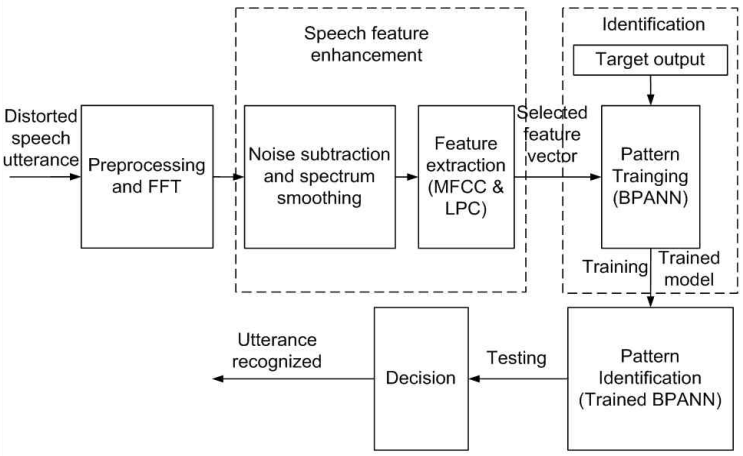

그림 3. 제안한 화자 분류 알고리즘
Fig. 3. Proposed speaker identification algorithm

이 후에 12차의 MFCC 계수와 12차의 LPC 계수를 추출하여 총 24차 계수들을 BPANN 네트워크에 입력하여 목표로 하는 화자를 출력하도록 식 (2)와 같이 제곱 합이 최소가 되도록 24-30-3의 네트워크의 BPANN을 학습시킨다. BPANN의 학습이 완료된 후에 학습과정에서 구한 연결 강도를 이용하여 미지의 24차 켑스트럼 계수의 입력이 들어왔을 때 최종적으로 분류할 3명의 화자가 분류 가능하도록 음성인식을 실시한다.

III. 실험 환경 및 실험 결과

본 장에서는 실험에서 사용한 음성의 데이터베이스와 잡음데이터에 대하여 기술한다. 실험에 사용하는 Aurora2 데이터베이스에는 잡음이 중첩된 음성이 포함되어 있으며, 미국 LDC(Linguistic Data Consortium)로부터 배포된 TI-DIGITS(연속 영어숫자 음성) 데이터베이스에 여러 종류의 잡음을 계산기상에서 중첩함으로써 생성하였다. Aurora2 음성데이터의 샘플링은 8kHz이며 이러한 깨끗한 음성데이터에 다른 잡음신호가 인위적으로 추가되었다[10].

본 실험에서는 Aurora2 데이터베이스의 셋 A에 포함된 20명의 화자에 의해 발성된 음성신호를 사용하여 BPANN에 의하여 학습 및 테스트를 수행하였으며, 음성신호에 중첩되는 배경잡음은 컴퓨터에서 시뮬레이션하여 작성한 백색잡음(White Noise)와 Aurora2 데이터베이스에 포함된 자동차 잡음(Car Noise)를 사용하였다. 또한 입력 신호대잡음비는 20dB에서 5dB까지의 잡음으로 오염된 음성신호를 사용하였다.

표 1은 제안한 알고리즘을 사용하여 백색잡음 환경 하에서 화자인식 실험을 실시한 음성의 화자 인식률의 실험결과를 나타낸다. 표 1에서 “GMM & HMM”은 클러스터 기반 손실특징 복원을 위한 독립적인 대역의 마스크추정 방법이며 일반적인 MFCC를 이용한 가우시안 혼합 모델(GMMs, Gaussian Mixture Models)에 의한 음성인식모델이다. 여기에서는 Aurora2의 셋 A를 이용하여 HMM을 사용하여 음성인식의 학습을 하며, 클러스터 기반 손실 특징 복원을 위하여 GMM을 사용하여 학습한다[11]. 표 1의 결과로부터 본 논문에서 제안한 알고리즘이 “GMM & HMM” 방법보다 음성의 화자 분류율이 최대 14.86% 개선됨을 알 수 있다.

표 1. 백색잡음에 대한 잡음음성 분류 결과
Table 1. Noisy speech identification results for white noise

Algorithm	Noisy speech identification results[%]				
	Clean	20dB	15dB	10dB	5dB
GMM & HMM	-	96.60	89.50	84.70	74.90
Proposed Alg.	99.27	99.24	98.97	96.93	89.76
Impr.	-	2.64	9.47	12.23	14.86

표 2는 자동차 잡음에 대하여 기존의 방법(GMM & HMM[11], NS & VTSA[12], SPLICE & HMM[13], SPARK & PBS[14])과 본 논문에서 제안한 방법(Proposed Alg.)을 비교한 화자의 음성 분류율을 나타낸다. 기존의 "NS & VTSA"[12]는 잡음억제(NS, Noise Suppression)와 벡터 테일러급수 적응(VTSA, Vector Taylor Series Adaptation) 방법을 이용한 음성인식기이다.

표 2. 자동차 잡음에 대한 잡음음성 분류 결과
Table 2. Noisy speech identification results for car noise

Algorithm		Noisy speech identification results[%]				
		Clean	20dB	15dB	10dB	5dB
NS & VTSA		-	97.80	96.40	93.90	87.80
SPARK & PBS		99.19	98.69	98.03	95.47	88.76
SPLICE & HMM		98.27	97.49	96.48	94.18	88.52
GMM & HMM		-	95.00	95.50	92.00	83.00
Proposed Alg.		99.18	99.05	98.69	96.51	89.36
Impr.	SPLICE & HMM	0.91	1.56	2.21	2.33	0.84
	GMM & HMM	-	4.05	3.19	4.51	6.36

"SPLICE & HMM"은 HMM 기반의 음성합성방법을 사용하여 스테레오 채널로부터 이상적인 깨끗한 음성특징을 대체할 수 있는 특징을 생성하는 스테레오 기반 구분적 선형 환경 보상(SPLICE, Stereo-based Piecewise Linear Compensation for Environments) 방법이다[13]. "SPARK & PBS"는 성긴청각 재생커널(SPARK, Sparse Auditory Reproducing Kernel) 특징과 전력 바이어스차감(PBS, Power Bias Subtraction) 방법을 이용한 음성인식 모델이다[14]. 위에서 기술한 종래의 방법들은 모두 Aurora2의 테스트 셋 A를 사용하여 HMM을 학습시킨다. 표 2의 자동차 잡음의 실험결과로부터 "SPLICE & HMM"에 대해서는 최대 2.33%, "GMM & HMM"에 대해서는 최대 6.36%로 음성의 화자 분류율이 개선된 것을 알 수 있다. 특히 잡음이 많이 부가된 음성의 경우에 화자 분류율이 더욱 더 개선됨을 확인할 수 있었다. 따라서 본 논문에서 제안한 화자분류 알고리즘의 성능이 종래의 방법과 비교하여 양호하였으며, 특히 잡음 하에서 본 논문에서 제안한 알고리즘이 유효하다는 것을 실험 결과로부터 확인할 수 있었다.

Ⅳ. 결 론

본 논문에서는 음성의 화자분류를 위하여 잡음으로 오염된 음성으로부터 잡음성분을 억제하여 깨끗한 음성만을 추출하는 음성특징 강조기법 및 음성의 화자분류 기법을 제안하였다. 제안한 주파수 영역의 위너필터와 스펙트럼 평활화 방법을 사용한 켑스트럼 잡음모델의 음성특징 강조기법을 사용하여 잡음이 억제된 MFCC와 LPC에 의한 음성특징량을 추출하였으며, 오차역전파 신경회로망을 사용하여 잡음으로 오염된 음성의 화자를 분류하였다. 제안한 알고리즘을 사용하여 백색잡음 및 자동차 잡음으로 오염된 음성신호에 대하여 음성의 화자분류 실험을 수행하였으며, 백색잡음에 대하여는 최대 14.86%, 자동차 잡음에 대해서는 최대 6.36%의 화자 분류율을 구하였다. 따라서 본 논문에서 제안한 알고리즘은 낮은 신호대잡음 환경에서도 음성의 화자분류율이 상당히 개선되는 것을 알 수 있었으며, 이러한 점을 고려한다면 본 논문에서 제안한 알고리즘이 백색잡음 및 자동차 잡음에 대하여 비교적 유효하다는 것을 실험으로 확인할 수 있었다.

References

[1] A. Maurya and R. K. Aggarwal, "Speaker recognition for noisy speech in telephonic channel", 2nd International Conference on Applied and Theoretical Computing and Communication Technology, pp. 451-456, Jul. 2016.

[2] J. L. Carmona, J. Barker, A. M. Gómez, and Ning Ma, "Speech Spectral Envelope Enhancement by HMM-Based Analysis/Resynthesis", IEEE Signal Processing Letters, Vol. 20, No. 6, pp. 563-566, Jun. 2013.

[3] B. Li and K. C. Sim, "A Spectral Masking Approach to Noise-Robust Speech Recognition Using Deep Neural Networks", IEEE/ACM

Transactions on Audio, Speech, and Language Processing, pp. 1296-1305, Vol. 22, No. 8, Jun. 2014.

[4] M. K. Lim, D. H. Lee, H. S. Park, and J. H. Kim, "Audio Event Detection Using Deep Neural Networks", Journal of Digital Contents Society, Vol. 18, No. 1, pp. 183-190, Feb. 2017.

[5] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 4, pp. 1218-1234, Jul. 2006.

[6] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagation errors", Nature, Vol. 323, pp. 533-536, Oct. 1986.

[7] J. S. Choi, "Speech-dependent Speaker Identification Using Mel Frequency Cepstrum Coefficients for Continuous Speech Recognition", Journal of KIIT, Vol. 14, No. 10, pp. 67-72, Oct. 2016.

[8] B. L. Sturm, M. Morvidone, and L. Daudet, "Musical Instrument Indentification Using Multiscale Mel-frequency Cepstral Coefficients", 2010 18th European Signal Processing Conference, pp. 477-481, Apr. 2010.

[9] H. K. Kim, S. H. Choi, and H. S. Lee, "On approximating line spectral frequencies to LPC cepstral coefficients", IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 2, pp. 195-199, Mar. 2000.

[10] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", in Proc. ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, Oct. 2000.

[11] W. L. Kim and R. M. Stern, "Band-Independent Mask Estimation for Missing-Feature Reconstruction in the Presence of Unknown Background Noise", IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Vol. 1, pp. I-305-I-308, May 2006.

[12] S. Komeiji, T. Arakawa, and T. Koshinaka, "A noise-robust speech recognition method composed of weak noise suppression and weak Vector Taylor Series Adaptation", 2012 IEEE Spoken Language Technology Workshop (SLT), pp. 103-106, Dec. 2012.

[13] J. Du, Y. Hu, L. R. Dai, and R. H. Wang, "HMM-based pseudo-clean speech synthesis for splice algorithm", 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4570-4573, Mar. 2010.

[14] A. Fazel and Shantanu Chakrabartty, "Sparse Auditory Reproducing Kernel (SPARK) Features for Noise-Robust Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1362-1371, Dec. 2012.

저자소개

최 재 승 (Jae-Seung Choi)

1989년 : 조선대학교 전자공학과 공학사

1995년 : 일본 오사카시립대학 전자정보공학부 공학석사

1999년 : 일본 오사카시립대학 전자정보공학부 공학박사

2000년 ~ 2001년 : 일본 마쯔시타 전기산업주식회사 (현, 파나소닉 주식회사) AVC사 연구원

2002년 ~ 2007년 : 경북대학교 디지털기술연구소 책임연구원

2007년 ~ 현재 : 신라대학교 스마트전기전자공학부 교수

관심분야 : 음성신호처리, 신경회로망, 적응필터와 잡음제거, 디지털 TV 및 멀티미디어 등