JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 21, No. 6 : 1149~1154 Jun. 2017

위너필터법이 적용된 MFCC의 파라미터 추출에 기초한 화자독립 인식알고리즘

최재승*

Speaker Independent Recognition Algorithm based on Parameter Extraction by MFCC applied Wiener Filter Method

Jae-Seung Choi*

Division of Smart Electrical and Electronic Engineering, Silla University, Busan 46958, Korea

요 약

배경잡음 하에서 음성인식 시스템의 우수한 인식성능을 얻기 위해서 적절한 음성의 특징 파라미터를 선택하는 것이 매우 중요하다. 본 논문에서 사용한 특징 파라미터는 위너필터 방법이 적용된 인간의 청각 특성을 이용한 멜 주파수 켑스트럼 계수(Mel frequency cepstral coefficient, MFCC)를 사용한다. 즉, 본 논문에서 제안하는 특징 파라미터는 배경잡음을 제거한 후에 깨끗한 음성신호의 파라미터를 추출하는 새로운 방법이다. 제안한 수정된 MFCC 특징 파라미터를 다층 퍼셉트론 네트워크에 입력하여 학습시킴으로써 화자인식을 구현한다. 본 실험에서는 14차의 MFCC 특징 파라미터를 사용하여 화자독립 인식실험을 실시하였으며, 백색잡음이 혼합된 경우의 음성의 화자독립 인식률은 평균 94.48%로 효과적인 결과를 구할 수 있었다. 본 논문에서 제안한 방법과 기존의 방법들을 비교하였을 때 본 논문에서 제안한 화자인식 성능이 수정된 MFCC 특징 파라미터를 사용함으로써 향상되었다.

ABSTRACT

To obtain good recognition performance of speech recognition system under background noise, it is very important to select appropriate feature parameters of speech. The feature parameter used in this paper is Mel frequency cepstral coefficient (MFCC) with the human auditory characteristics applied to Wiener filter method. That is, the feature parameter proposed in this paper is a new method to extract the parameter of clean speech signal after removing background noise. The proposed method implements the speaker recognition by inputting the proposed modified MFCC feature parameter into a multi-layer perceptron network. In this experiments, the speaker independent recognition experiments were performed using the MFCC feature parameter of the 14th order. The average recognition rates of the speaker independent in the case of the noisy speech added white noise are 94.48%, which is an effective result. Comparing the proposed method with the existing methods, the performance of the proposed speaker recognition is improved by using the modified MFCC feature parameter.

키워드 : 멜 주파수 켑스트럼 계수, 위너필터, 화자인식, 화자독립, 특징 파라미터

Key word : Mel frequency cepstral coefficient, Wiener filter, speaker recognition, speaker independent, feature parameter

Received 01 February 2017, **Revised** 06 February 2017, **Accepted** 20 February 2017

* **Corresponding Author** Jae-Seung Choi (E-mail: jschoi@silla.ac.kr, Tel: +82-51-999-5608)
Division of Smart Electrical and Electronic Engineering, Silla University, Busan 46958, Korea

Open Access https://doi.org/10.6109/jkiice.2017.21.6.1149 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근의 컴퓨터기술의 발전과 더불어 현대 사회에서 이용 가능한 음성대화 시스템을 염두에 둔 음성인식 시스템 개발이 진행되고 있다[1]. 이러한 음성대화 시스템은 입력 음성으로부터 단순히 음성의 언어정보만을 추출하지 않고 음성의 개인 화자성 및 인간의 감정 등의 화자의 정적 혹은 동적 특징을 정확히 추출하여 효율적인 대화를 수행하려고 하는 연구들이다[2, 3]. 또한 이러한 시스템은 화자의 언어내용을 인식하여, 그 결과로부터 사용자의 의도에 따라서 정보제공을 하려는 시스템이다. 음성대화 시스템을 구축할 때 고려해야 할 사항으로는 화자의 정확한 검출, 배경잡음에 의한 음성시스템의 오동작, 감정인식, 모국어의 추정 등이 있다.

일반적으로 실제 환경에서 발생하는 주변 잡음으로 인하여 음성인식 시스템의 성능이 저하되는 문제점이 주로 발생된다[4]. 이러한 문제점의 원인으로는 학습된 음성인식 시스템의 환경과 실제로 구현된 음성인식 시스템의 환경에서 발생하는 음성신호의 특성 차이 때문이라고 할 수 있다. 여러 가지 잡음에 대한 대처방안으로서 음성인식 시스템의 성능을 향상시키기 위하여 음성강조[5], 위너필터[6], 최소통계모델에 기반을 둔 최소평균 제곱오차(Minimum Mean Square Error, MMSE) 방법[7] 등이 제안되고 있다. 음성인식 시스템의 인식률 향상 및 시스템의 우수한 음성신호 판별 성능을 얻기 위해서는 적절한 특징 파라미터를 선택하는 것이 매우 중요하다. 음성특징 파라미터를 추출하는 방법으로 선형예측부호화(Linear Predictive Coding, LPC) 분석을 통하여 파라미터를 추출하는 LPC 켑스트럼[8], 인간의 청각 특성을 이용한 MFCC[9, 10], 지각선형 예측분석(Perceptual Linear Prediction, PLP)[11]에 의한 방법 등이 있다.

본 논문에서는 잡음에 강인하고 음성인식에 효과적인 위너필터를 적용한 MFCC의 파라미터를 이용한 다층 퍼셉트론 신경회로망 기반의 음성의 화자 분류 알고리즘을 제안한다. 제안한 알고리즘을 기초로 하여 연속음성 데이터베이스를 사용하여 깨끗한 음성에 배경잡음을 혼합한 후 화자인식을 실시한다. 본 논문에서 제안한 MFCC 파라미터를 사용하여 기존의 방법들과 비교 실험한 결과 백색잡음이 혼합된 음성신호에 대하여 최대 7.81%의 인식률이 향상된 것을 확인할 수 있었다.

Ⅱ. 제안한 알고리즘

음성인식에 사용하는 특징 파라미터 값으로 푸리에 고속 변환(Fast Fourier Transform, FFT)에 의한 켑스트럼 계수, LPC 등과 같은 방법이 존재하지만, 음성인식에 우수한 성능을 가지는 MFCC는 스펙트럼 기반을 특징으로 하며 인간의 귀가 가지는 비선형적인 주파수 특성을 이용하기 때문에 유효한 특징 파라미터 값으로 알려져 있다[9, 10]. 그림 1은 일반적인 MFCC 계수를 추출하는 과정을 나타낸다. 음성신호를 프레임별로 구분하여 Pre-emphasis 과정을 수행하며 Hamming window를 적용한다. FFT 분석에 의하여 power spectrum을 구한 후에 주파수에 인지 특성을 반영한 Log 함수를 취한다. 이산코사인변환(Discrete Cosine Transform, DCT) 연산을 통하여 켑스트럼을 추출하여 MFCC 계수를 이용한 특징 파라미터를 추출한다.

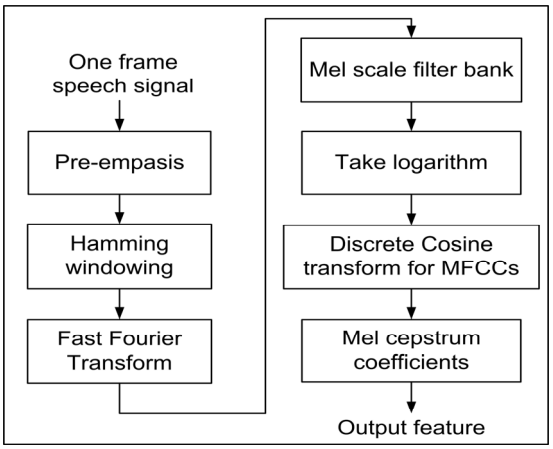

Fig. 1 General MFCC feature extraction

본 논문에서는 음성신호로부터 특징 파라미터 추출은 인간의 목소리의 특징을 추출하기 위해 사용하고 있는 MFCC 계수를 사용 하였다. 본 논문에서는 이러한 MFCC 계수 방법에 더하여 위너필터 방법을 적용하여 배경잡음을 제거한 후에 깨끗한 음성신호를 추출하는 그림 2와 같은 새로운 방법을 제안한다. 그림 2는 각 프레임에서 음성신호로부터 잡음의 스펙트럼 추정하여 위너필터에 의하여 잡음을 제거하는 위너필터 블록이 추가되었다. 따라서 본 논문에서는 잡음의 스펙트럼을 차감하기 위하여 잡음이 섞인 음성 중에서 각 프레임의

잡음을 추정한 후에 잡음이 제거된 음성의 스펙트럼을 구한다. 구해진 음성의 스펙트럼은 그림 2와 같이 위너필터 방법을 적용하여 수정된 MFCC 특징벡터를 구한다. 여기에서 위너필터는 잡음이 중첩된 원 신호와 필터링된 출력 신호와의 차이를 최소화하도록 한다.

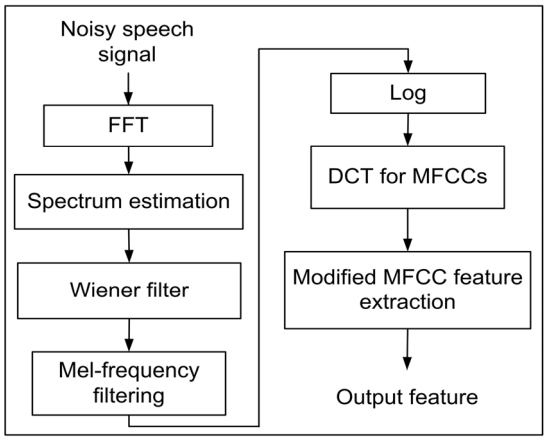

Fig. 2 Block diagram of proposed MFCC feature extraction using Wiener filter method

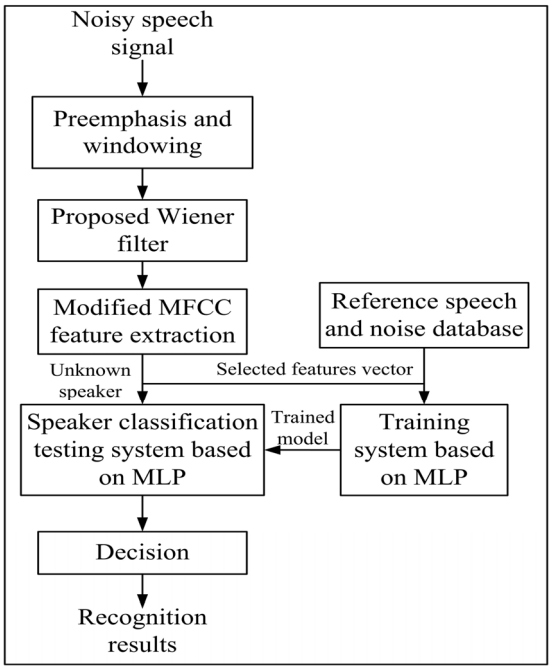

Fig. 3 Proposed speaker recognition algorithm using MFCC applied Wiener filter in noisy environment

일반적으로 음성인식 분야에 있어서 주변의 배경잡음으로 인하여 음성인식시스템의 성능이 저하되는 문제점이 많이 발생된다[4-7, 12]. 특히 마이크의 특성, 주변의 배경잡음, 마이크와의 거리 문제 등 여러 요소들이 음성인식 성능을 저하시키는 요인이 된다. 따라서 본 논문에서는 이러한 문제를 해결하기 위하여 위너필터 방법에 기초한 그림 2의 수정된 MFCC 계수를 추출하는 방법을 사용하여, 잡음이 섞인 환경에서 음성 향상기법을 적용한 그림 3의 음성인식 알고리즘을 제안한다. 화자의 인식은 다층 퍼셉트론 신경회로망(Multi-Perceptron Neural Network, MLP)[13]을 이용하여 화자인식 시스템을 구축하였다. 본 논문에서는 14차의 MFCC 특징벡터를 사용하여 원음 및 백색잡음 하에서 연속음성에 대하여 화자를 인식하는 실험을 한다. 여기에서 멜 주파수 켑스트럼(Mel-frequency cepstrum, MFC)은 단구간 신호의 전력스펙트럼을 나타내며, 본 논문에서 사용하는 14차의 MFCC 특징벡터는 여러 MFC들을 모아 놓은 계수들을 나타낸다. 또한 MFCC는 주파수 대역이 멜 스케일에서 균등하게 나누어지기 때문에 음성 등을 잘 표현할 수 있는 장점을 가지고 있다.

Ⅲ. 실험환경 및 실험결과

본 실험에서는 16 bits/sample, 8 kHz로 샘플링된 일본어 입력 음성신호를 사용하며, 256 샘플(32ms)을 한 프레임으로 사용하여 MLP 네트워크의 학습 단계와 테스트 단계의 두 단계로 실험을 진행하였다. 이 일본어 음성 데이터베이스는 일본 음성정보처리 개발협회에서 배포한 ATR 연구용 연속음성 데이터베이스 중에서 성인남성 화자와 성인여성 화자에 의한 문장을 임의적으로 선택하였다[14]. 본 실험에서 사용한 연속음성데이터의 시간은 최소 3초 정도에서 최대 약 7초 정도의 길이를 가지며, 평균 5.5초 정도의 시간 길이를 가진다. 본 실험에서 사용한 잡음데이터는 가우스 백색잡음의 배경잡음을 사용하여 평가하였다.

본 실험에서는 임의적으로 선택한 연속음성의 문장 중에서 Speaker 1, Speaker 2, Speaker 3의 3 종류의 화자를 사용하였으며, 각 화자에는 각각 5개의 문장들로 구성되며 각 화자 간에는 서로 다른 화자에 의해서 발성된 음성이다. 본 실험에서는 백색잡음이 부가된 경우

의 식 (1)의 입력 신호대잡음비(Signal-to-Noise Ratio, SNR)는, N=256에 대하여 Speaker 1의 입력 SNR은 9dB, Speaker 2의 입력 SNR은 1dB, Speaker 3의 입력 SNR은 –5dB을 각각 나타낸다. 여기에서 N은 한 프레임의 샘플수이다. 본 실험에서는 임의적으로 음성신호에 잡음신호를 중첩시켜 9dB, 1dB, -5dB의 입력 SNR 값을 정의한다.

$$SNR = 10\log_{10}\left(\frac{\sum_{n=1}^{N} S(n)^2}{\sum_{n=1}^{N} N(n)^2}\right) \quad (1)$$

본 실험에서는 화자인식 알고리즘의 성능을 평가하기 위하여 MLP 네트워크를 사용한 화자독립 인식 실험을 실시하였다. 본 논문에서는 입력층, 출력층 및 은닉층의 3층의 네트워크로 구성된 MLP 네트워크[13]에 의하여 화자인식을 구현하였으며, 출력층으로부터 입력층으로 오차를 전파하는 특징을 가지는 역전파 학습 알고리즘(Back-propagation Learning Algorithm)[15]에 의하여 화자인식을 구현한다. MLP 네트워크의 입력으로는 그림 2에서 구해진 수정된 14차의 MFCC 계수이다. 신경회로망의 학습 단계에서는 잡음이 중첩된 음성신호로부터 위너필터에 의하여 음성이 향상된 MFCC 계수를 MLP에 입력하여 학습시키는 MLP 학습 모델을 구현하였다. 테스트 단계에서는 학습된 MLP 모델을 사용하여 화자를 분류한다. 본 실험에서는 MFCC 계수는 14개의 특징벡터를 사용하며 화자 분류는 각 프레임에서 MLP에 의해서 분류된다. 본 논문에서는 제안한 MFCC 특징 파라미터를 사용한 MLP 기반의 화자인식 알고리즘을 이용하여 음성의 화자인식 성능을 평가하였으며, 특히 잡음 환경 하에서 인식률 실험을 하였다.

본 논문에서 제안한 위너필터에 의한 효과를 나타내기 위하여 한 프레임의 음성신호(제 12프레임)에 대하여 그림 4의 FFT 스펙트럼과 그림 5의 MFCC 계수의 입력과 출력의 변화 모양을 나타낸다. 그림 4(a)는 백색잡음이 중첩된 음성신호의 입력 FFT 스펙트럼을 나타내며, 그림 4(b)는 위너필터링 처리를 한 후의 잡음이 제거된 FFT 스펙트럼을 나타낸다. 그림 5(a)는 백색잡음이 중첩된 음성신호에 대하여 잡음이 섞인 MFCC 계수를 나타내며, 그림 5(b)는 위너필터링 처리를 하여 수정된 MFCC 계수를 나타낸다. 그림 5의 결과로부터 위너필터에 의하여 수정된 MFCC 계수와 원래의 MFCC 계수를 비교해보면, 저차의 계수들에서 진폭의 차이를 가지는 것을 알 수 있다. 따라서 본 논문에서 제안하는 MLP 네트워크에 입력되는 수정된 MFCC 계수를 이용하여 화자 인식률이 향상되는 것을 확인할 수 있었다.

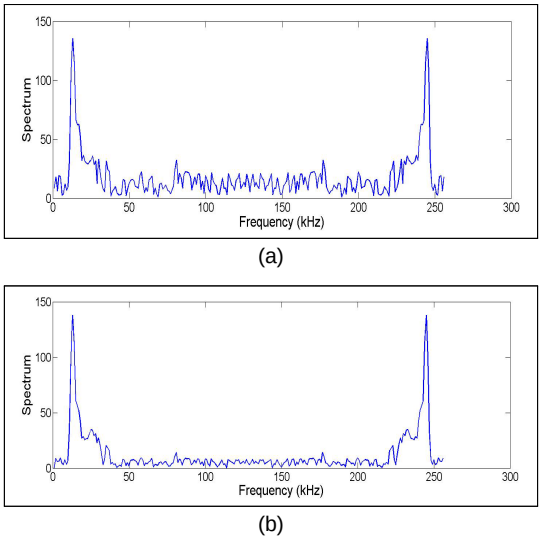

Fig. 4 FFT spectrums in the case of white noise (a) FFT spectrum of noisy speech signal (b) FFT spectrum after Wiener filtering

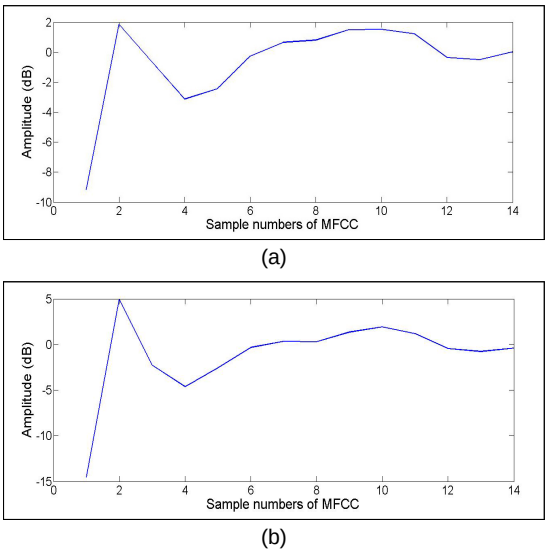

Fig. 5 MFCCs in the case of white noise (a) MFCC of noisy speech signal (b) MFCC after Wiener filtering

본 실험에서는 네트워크의 입력층으로는 14개의 유닛(14차의 MFCC)을 사용하며, 중간층으로는 30개의 유닛을 사용한다. 출력층은 3개 그룹의 화자를 식별하기 위하여 3개의 출력층 유닛으로 구성된다. 즉, 14-30-3의 네트워크로 구성된다. 또한 MLP 네트워크에 백색잡음이 부가된 경우에 대하여, 음성신호의 MFCC 계수를 네트워크에 입력하여 MLP 네트워크를 학습시켜 테스트를 수행하였다. 네트워크를 학습시킬 때의 최대학습 횟수는 20,000회로 하였다.

본 논문에서 제안한 MLP 네트워크는 식 (2)와 같은 비선형 함수인 시그모이드 함수를 사용하며, 오차함수는 목표 출력값과 실제 출력값 간의 차이를 제곱합하여 계산한다. 여기에서 y_j는 상위 서브넷의 유닛 j의 출력, x_i는 하위 서브넷의 유닛 i의 출력, w_{ji}는 연결강도, θ는 각 유닛의 임계값을 각각 나타낸다.

$$y_j = \frac{2.0}{1.0 + \exp(-\sum_i w_{ji} + \theta)} - 1.0 \qquad (2)$$

본 실험에서는 배경잡음 환경에서 화자인식 실험을 실시하여 인식률 실험결과를 표 1에 나타낸다. 표 1의 화자의존의 경우 평균 화자 인식률이 99.06%로 가장 양호하였으며, 화자독립 인식의 경우는 평균 94.48%의 인식률을 나타냈다. 표 2는 Multitaper MFCC와 Gaussian Model을 사용한 화자식별(참고문헌[10])에 의한 기존방법과 본 논문에서 제안한 방법에 대한 평균 화자인식률의 결과를 비교하였다. 표 2에서 'MFCC in Ref[10]'은 참고문헌[10]의 MFCC에 의한 방법을 나타내며, 또한 'Multitaper in Ref[10]'은 참고문헌[10]의 Multitaper에 의한 방법을 나타낸다. 이 Multitaper 방법은 다수의 시간영역 창함수나 여러 개의 taper를 사용하여 스펙트럼 추정의 변화를 줄이는 방법이다. 표 1과 표 2에서 "Correct rates"는 정확하게 식별된 화자인식률을 나타내며, "Error rates"는 잘못 식별된 화자인식 오류율을 나타낸다.

기존방법과 비교에 있어서 참고문헌[10]에서 사용된 데이터와 제안한 방법에서 사용된 데이터가 서로 다르며, 또한 실험환경에 대한 조건이 다르기 때문에 성능을 비교하는데 어려움이 있었다. 그러나 참고문헌[10]의 방법과 동일한 환경에서 실험은 수행되지 않았지만, 간접적인 성능비교를 통하여 본 논문에서 제안한 알고리즘은 'MFCC in Ref[10]' 방법보다도 7.81% 인식률이 향상되었으며, 'Multitaper in Ref[10]' 방법보다도 인식률이 4.48% 향상된 것을 알 수 있었다. 따라서 본 논문에서 제안한 방법의 화자인식 알고리즘의 성능이 기존의 방법보다 우수할 것으로 예상되었다.

Table. 1 Recognition rates by MFCC for white noise

Test speaker	Speaker dependent		Speaker independent	
	Correct rates[%]	Error rates[%]	Correct rates[%]	Error rates[%]
S1	98.79	1.21	93.68	6.32
S2	98.38	1.62	94.19	5.81
S3	100.0	0.00	95.57	4.43
Average	99.06	0.94	94.48	5.52

Table. 2 Average recognition rates of conventional and proposed algorithm

Feature		Speaker independent	
		Correct rates[%]	Error rates[%]
MFCC in Ref[10]		86.67	13.33
Multitaper in Ref[10]		90.00	10.00
Proposed Alg.		94.48	5.52
Impr.	MFCC	7.81	-
	Multitaper	4.48	-

Ⅳ. 결 론

본 논문에서는 배경잡음으로 인하여 음성인식 시스템의 성능이 크게 저하되는 문제점을 해결하기 위하여 위너필터 방법이 적용된 MFCC를 사용하여 배경잡음을 제거한 후에 깨끗한 음성신호를 추출하는 새로운 방법을 제안하였다. 추출한 MFCC 특징 파라미터를 이용하여 MLP 네트워크 기반의 독립적인 음성의 화자인식 알고리즘을 제안하여 잡음으로 오염된 음성신호에 대하여 화자인식 실험을 실시하였다. 실험결과로부터 본 논문에서 제안한 방법과 기존의 방법을 비교하였을 때 본 논문에서 제안한 방법의 화자 인식률이 더 효과적이라는 것을 확인할 수 있었다. 그러나 향후의 연구에서는 화자인식 방법의 객관적인 성능을 위하여, SNR 비율의 인식률 변화 특성 및 통계적 교차검정 등의 실험이 필요하다고 본다.

REFERENCES

[1] L. R. Gottlieb and G. Friedland, “On the Use of Artificial Conversation Data for Speaker Recognition in Cars,” *IEEE International Conference on Semantic Computing*, pp. 124-128, Sept. 2009.

[2] P. Day and A. K. Nandi, “Robust Text-Independent Speaker Verification Using Genetic Programming,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 285-295, January 2007.

[3] P. Song, Y. Jin, C. Zha and L. Zhao, “Speech emotion recognition method based on hidden factor analysis,” *Electronics Letters*, vol. 51, no. 1, pp. 112-114, Jan. 2015.

[4] T. Yamada, M. Kumakura and N. Kitawaki, “Performance Estimation of Speech Recognition System Under Noise Conditions Using Objective Quality Measures and Artificial Voice,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2006-2013, October 2006.

[5] 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, June 2013.

[6] 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, July 2006.

[7] M. Krawczyk-Becker and T. Gerkmann, “On MMSE-Based Estimation of Amplitude and Complex Speech Spectral Coefficients Under Phase-Uncertainty,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2251-2262, December 2016.

[8] 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, March 2000.

[9] W. W. Hung and H. C. Wang, “On the use of weighted filter bank analysis for the derivation of robust MFCCs,” *IEEE Signal Processing Letters*, vol. 8, no. 3, pp. 70-73, Mar. 2001.

[10] K. V. Veena and M. Dominic, “Speaker Identification and Verification of Noisy Speech Using Multitaper MFCC and Gaussian Models,” *IEEE International Conference on Power, Instrumentation, Control and Computing*, pp. 1-4, Dec. 2015.

[11] M. Holmberg, D. Gelbart and W. Hemmert, “Automatic speech recognition with an adaptation model motivated by auditory processing,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 43-49, Jan. 2006.

[12] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol.27, no.2, pp. 113-120, April 1979.

[13] S. K. Pal and S. Mitra, “Multilayer perceptron, fuzzy sets, and classification,” *IEEE Transaction on Neural Networks*, vol. 3, no. 5, pp. 683-697, Sep. 1992.

[14] A. Kurematsu, K.Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp.357-363, 1990.

[15] D. Rumelhart, G. Hinton and R. Williams, “Learning representations by back-propagation errors,” *Nature*, vol. 323, pp. 533-536, Oct. 1986.

최재승(Jae-Seung Choi)

1989년 조선대학교 전자공학과 공학사
1995년 일본 오사카시립대학교 전자정보공학부 공학석사
1999년 일본 오사카시립대학교 전자정보공학부 공학박사
2000년~2001년 일본 마쯔시타 전기산업주식회사 (현, 파나소닉 주식회사) AVC사 연구원
2002년~2007 경북대학교 디지털기술연구소 책임연구원
2007년~현재 신라대학교 스마트전기전자공학부 교수
※ 관심분야 : 음성신호처리, 신경회로망, 잡음제거 등