한국음향학회지 제27권 제4호 pp. 191~199 (2008)

음성 개선 기반의 모델 보상 기법을 이용한 강인한 잡음 음성 인식

A Noise Robust Speech Recognition Method Using Model Compensation Based on Speech Enhancement

신 광 호*, 정 호 열*, 정 현 열*
(Guanghu Shen*, Ho-Youl Jung*, Hyun-Yeol Chung*)

*영남대학교 정보통신공학과
(접수일자: 2008년 3월 7일; 수정일자: 2008년 4월 22일; 채택일자: 2008년 5월 8일)

본 논문에서는 잡음 환경하의 음성 인식을 위해 전처리 단계에서 Mel-warped Wiener Filtering (MWF) 기법을 이용하여 입력 음성을 개선하고 후처리 단계에서 PMC (Parallel Model Combination) 기법을 이용하여 인식 모델을 보상하는 MWF-PMC 잡음 처리 기법을 제안한다. PMC 기법은 전처리 단계에서 개선된 음성의 묵음 구간으로부터 잔류 잡음을 취하여 깨끗한 음성을 이용하여 작성한 인식 모델을 보상함으로써 잡음 환경하의 음성 인식 성능을 향상시킬 수 있다. 인식 실험을 위한 음성 데이터는 국어공학연구소 (KLE)에서 작성한 PBW (Phoneme Balanced Words) 452 단어 음성 데이터를 8 kHz로 다운 샘플링한 후 Subway, Car 및 Exhibition 잡음을 5단계의 신호 대 잡음비 (SNR)를 0, 5, 10, 15, 20 dB로 부가하여 구성하였다. 인식 실험 결과, 본 논문에서 제안한 MWF-PMC 기법이 기존의 결합된 기법보다 전반적으로 향상된 인식 성능을 얻어 그 유효성을 확인할 수 있었다.
핵심용어: 음성 인식, 음성 개선, Mel-warped wiener filtering, 모델 보상, PMC
투고분야: 음성처리 분야 (2.5)

In this paper, we propose a MWF-PMC noise processing method which enhances the input speech by using Mel-warped Wiener Filtering (MWF) at pre-processing stage and compensates the recognition model by using PMC (Parallel Model Combination) at post-processing stage for speech recognition in noisy environments. The PMC uses the residual noise extracted from the silence region of enhanced speech at pre-processing stage to compensate the clean speech model and thus this method is considered to improve the performance of speech recognition in noisy environments. For recognition experiments we down-sampled KLE PBW (Phoneme Balanced Words) 452 word speech data to 8 kHz and made 5 different SNR levels of noisy speech, i.e., 0 dB, 5 dB, 10 dB, 15 dB and 20 dB, by adding Subway, Car and Exhibition noise to clean speech. From the recognition results, we could confirm the effectiveness of the proposed MWF-PMC method by obtaining the improved recognition performances over all compared with the existing combined methods.
Keywords: Speech recognition, Speech enhancement, Mel-warped wiener filtering, Model compensation, PMC
ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

음성 인식 기술은 사람이 기계를 가장 손쉽게 사용할 수 있게 해주는 수단으로 오랫동안 실용화를 위한 많은 연구가 계속되고 있다. 음성 인식 기술의 최종 목표는 임의의 화자가 발성한 연속적인 음성을 실시간에 높은 인식률로 인지하는 인식 시스템의 개발이라 할 수 있다. 그러나 현재까지 개발된 많은 음성 인식 시스템은 훈련 및 인식 과정이 잡음이 거의 없는 환경에서 수행되고 있어 배경 잡음이 존재하는 실제 환경에서는 인식 성능이 현저히 저하되는 문제가 발생한다. 일반적으로 이러한 성능 저하는 학습 환경과 인식 환경의 불일치 (mismatch)에서 비롯된다.

잡음에 강인한 음성 인식을 위해 일반적으로 많이 사용되는 잡음 처리 방식은 잡음에 강인한 특징 파라미터

책임저자: 정 현 열 (hychung@ynu.ac.kr)
712-749 경북 경산시 대동 214-1 영남대학교 전자정보공학부
(전화: 053-810-2496; 팩스: 053-810-4742)

추출 방식 (robust feature extraction method), 음성 개선 방식 (speech enhancement method) 및 모델 보상 방식 (model compensation method)등의 세 가지로 구분할 수 있다 [1]. 먼저, 잡음 환경에 강인한 음성 특징 파라미터로는 MFCC (Mel Frequency Cepstral Coefficient)와 PLP (Perceptual Linear Prediction) [2] 등이 알려져 있으며 음성 개선 방식은 대표적으로 Spectral Subtraction (SS) [1][3], MMSE-STSA (Minimum Mean Square Error -Short Time Spectral Amplitude) [4-5], Wiener Filtering (WF) [3][6] 및 Mel-warped Wiener Filtering (MWF) [7-8] 기법 등이 있다. 이러한 기법들은 일반적으로 우수한 인식 성능을 보여주고 있지만 신호 대 잡음비 (SNR: Signal to Noise Ratio)의 저하에 따라 잡음 스펙트럼의 추정 오차가 증대함으로 인해 음성 왜곡이 발생하여 음성 인식 성능을 현저히 저하시키는 문제가 발생한다 [15]. 현재 표준 잡음 제거 기법으로 Aurora 프로젝트 [7]에서는 Mel-warped Wiener Filtering 기법을 적용하고 있다.

이와 달리 모델 보상 방식은 HMM (Hidden Markov Models) 파라미터의 변형을 이용하여 수행됨으로 음성 개선 방식에서 발생하는 음성 신호의 왜곡을 근본적으로 방지할 수 있다. 대표적인 기법으로 PMC (Parallel Model Combination) [9-10], VTS (Vector Taylor Series) [11] 및 JA (Jacobian Adaptation) [4][17] 등이 있다. 그 중에 VTS 기법은 PMC 기법에 비해 성능 면에서 조금 우수하나 구현이 다소 번거롭고 계산량이 많은 단점이 있다 [11]. 반면에 JA 기법은 PMC 기법에 비해 계산량은 적으나 PMC 기법보다 인식 성능이 약간 저하 된다는 단점이 있다 [12].

잡음 환경 하에서의 대부분의 음성 인식에 관한 연구들은 음성 개선 방식과 모델 보상 방식으로 구분하여 독립적으로 수행해 오고 있다 [4]. 이 두 방법을 결합할 경우, 보다 더 나은 인식성능을 기대할 수 있으며 현재까지 이 두 방법을 결합한 방법으로는 SS-PMC [13], CSS-PMC [14], MMSESTSA-PMC/JA [4] 및 WF-NOVO [15] 등의 기법들이 발표 되었는데 여기서 NOVO (Noise and Voice Composition) 기법은 F. Martin [16] 등에 의하여 제안되었으며 PMC 기법과 거의 동일하다 [17].

본 논문에서는 동일한 실험 조건 하에서 다양한 인식 실험을 수행하여 기존의 각 기법들에 대하여 한국어 단어 음성 인식의 인식 성능을 비교 평가한 다음 향상된 인식 성능을 얻기 위하여 위에서 열거한 음성 개선 기법과 모델 보상 기법들을 결합하는 기법을 이용하여 효과적인 음성인식 성능 향상 기법을 강구하고자 한다. 특히 잡음 환경 하에서 한국어 단어 음성 인식 성능을 향상시키기 위해 음성 개선 기법으로 잘 알려진 MWF 기법과 모델 보상 기법으로 잘 알려진 PMC 기법과의 결합에 대해 중점적으로 검토하여 타 방법과의 성능비교를 실시한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존의 잡음 처리 방식에 대하여 음성 개선 방식과 모델 보상 방식으로 구분하여 각 방식의 대표적인 기법에 대하여 간략히 설명한다. 3장에서는 본 논문에서 제안하는 MWF-PMC 기법에 대하여 소개하고, 4장에서는 다양한 잡음 환경 하에서 인식 실험을 수행하여 기존의 기법들과 인식 성능을 비교 및 분석한다. 그리고 마지막으로 5장에서는 본 논문의 결론을 맺는다.

II. 잡음 처리 방식 (noise processing method)

2.1. 음성 개선 방식 (speech enhancement method)

음성 개선은 부가 잡음에 의해 왜곡된 입력 음성 신호로부터 잡음을 제거하고 원 음성 신호를 추정하는 기술이다. 따라서 음성 개선 기술은 음성 인식은 물론 음성 부호화 (speech coding) 등의 분야에서 오래 전부터 널리 사용되고 있는 기술이다. 음성 개선 방식의 대표적인 기법들로는 Spectral Subtraction (SS) [1][3], MMSE-STSA [4-5], Wiener Filtering (WF) [3][6] 및 Mel-warped Wiener Filtering (MWF) [7-8] 등이 있다. 이하 이 기법들을 간략히 분석한다.

2.1.1. Spectral Subtraction (SS)

Spectral Subtraction (SS) 기법은 음성 신호와 부가 잡음이 서로 독립이라는 가정 하에서 왜곡된 입력 음성 신호로부터 개선된 음성 신호를 얻는 기법이다. 실제 환경에서 음성 신호는 다양한 채널 특성 때문에 부가 잡음 및 왜곡이 발생한다. 여기서 채널 왜곡을 고려하지 않으면 입력 음성 신호에 부가되는 배경 잡음으로 인하여 대부분의 문제가 발생한다.

부가 잡음에 왜곡된 입력 음성 신호 $y(t)$는 식 (1)과 같이 표현할 수 있다.

$$y(t) = x(t) + d(t) \tag{1}$$

여기서 $x(t)$, $d(t)$는 각각 원 음성 신호 및 잡음 신호를 나타낸다. 식 (1)의 각 신호에 대해 푸리에 변환 (fourier

transform)을 수행하여 주파수 영역으로 표현하면 식 (2)와 같다.

$$\begin{aligned} X_k(t) &= A_k(t)e^{j\Phi_k} \\ D_k(t) &= N_k(t)e^{j\psi_k} \\ Y_k(t) &= R_k(t)e^{j\eta_k} \end{aligned} \tag{2}$$

여기서 t, k는 각각 시간과 주파수 스펙트럼 빈 (bin)을 나타내고, Φ_k, ψ_k, η_k는 각 주파수 스펙트럼의 위상을 나타내며, $A_k(t)$, $N_k(t)$ 및 $R_k(t)$는 각 신호의 스펙트럼 크기를 나타낸다.

원 음성 신호와 잡음 신호가 서로 상관성이 없다는 가정을 할 때 주파수 영역에서 각 신호의 스펙트럼과 파워 스펙트럼은 식 (3)과 식 (4)로 나타낼 수 있다.

$$Y_k(t) = X_k(t) + D_k(t) \tag{3}$$

$$P_{yk}(t) = P_{xk}(t) + P_{dk}(t) \tag{4}$$

정확한 잡음 제거를 위하여 현재 t번째 프레임의 예측 잡음 파워스펙트럼 $\widehat{P_{dk}}(t)$은 바로 전 $t-1$번째 프레임에서 예측된 잡음 파워 스펙트럼 $\widehat{P_{dk}}(t-1)$과 현재 t번째 프레임의 왜곡된 입력 음성 신호의 파워 스펙트럼 $P_{yk}(t)$을 이용하여 식 (5)와 같이 갱신한다.

$$\widehat{P_{dk}}(t) = (1-\gamma)\widehat{P_{dk}}(t-1) + \gamma P_{yk}(t) \tag{5}$$

따라서 잡음이 제거된 음성 신호의 파워 스펙트럼은 식 (6)을 이용하여 얻어진다.

$$P_{xk}(t) = \begin{cases} P_{yk}(t) - \alpha\widehat{P_{dk}}(t), & \text{if } P_{yk}(t) > \dfrac{\alpha}{1-\beta}\widehat{P_{dk}}(t) \\ \beta P_{yk}(t), & otherwise \end{cases} \tag{6}$$

여기서 $P_{xk}(t)$는 Spectral Subtraction (SS) 기법을 적용하여 얻은 개선된 음성 신호의 파워 스펙트럼이며, α는 over-estimation factor, 그리고 β는 flooring factor를 나타낸다. 본 논문에서는 α 및 β는 일반적으로 많이 사용되고 있는 값 0.5, 0.1을 이용한다 [1].

2.1.2. MMSE-STSA

잡음 신호를 $d(t)$, 원 음성 신호를 $x(t)$, 그리고 왜곡된 입력 음성 신호를 $y(t)$로 가정하면 이들의 관계는 시간 영역에서는 식 (1), 그리고 주파수 영역에서는 식 (3)과 같이 나타낼 수 있으며 각 신호들의 주파수 성분을 크기와 위상으로 나타내면 식 (2)와 같이 나타낼 수 있다.

인간의 청각 특성은 음성 신호의 위상보다 스펙트럼 크기에 더욱 민감하므로 원 음성 신호의 추정은 식 (7)과 같이 원 음성 신호의 스펙트럼 크기 $A_k(t)$에 대한 MMSE 추정치 $\widehat{A_k}(t)$를 구하는 문제로 간략화 할 수 있다. $\Gamma[\bullet]$는 Gamma 함수를 의미 한다($\Gamma(1.5) = \sqrt{\pi}/2$). 여기서 posterior SNR $\gamma_k(t)$, priori SNR $\xi_k(t)$는 식 (9), 식 (10)과 같이 정의할 수 있다.

$$\begin{aligned} \widehat{A_k}(t) &= E\{A_k(t)|Y_k(t)\} \\ &= \Gamma(1.5)\frac{\sqrt{\upsilon_k(t)}}{\gamma_k(t)}M(\upsilon_k(t))R_k(t) \\ &= G_{MMSE}(\xi_k(t), \gamma_k(t))\, R_k(t) \end{aligned} \tag{7}$$

$$\upsilon_k(t) = \frac{\xi_k(t)}{1+\xi_k(t)}\gamma_k(t) \tag{8}$$

$$\gamma_k(t) = \frac{R_k^2(t)}{P_{dk}(t)} \tag{9}$$

$$\xi_k(t) = \frac{P_{xk}(t)}{P_{dk}(t)} = \alpha\frac{\widehat{A_k^2}(t-1)}{P_{dk}(t-1)} + (1-\alpha)O[\gamma_k(t)-1] \tag{10}$$

$$\begin{aligned} M(\upsilon_k(t)) &= \exp(-\frac{\upsilon_k(t)}{2})[(1+\upsilon_k(t)) \\ &I_0(\frac{\upsilon_k(t)}{2}) + \upsilon_k(t)I_1(\frac{\upsilon_k(t)}{2})] \end{aligned} \tag{11}$$

그리고 $P_{xk}(t)$와 $P_{dk}(t)$는 각각 원 음성 신호와 잡음 신호의 k번째 스펙트럼 빈 (bin)의 파워 스펙트럼을 의미하고, α는 이전 프레임까지의 추정된 음성 신호의 파워 스펙트럼을 반영하는 forgetting factor, $O[\bullet]$는 양의 값을 가지기 위한 연산자이다. 여기서 $M[\bullet]$은 식 (11)과 같이 정의되며 I_0, I_1은 각각 0차와 1차 modified Bessel function을 나타낸다.

왜곡된 입력 음성 신호로부터 음성 신호가 존재할 확률 (SPP: Speech Present Probability)을 식 (7)에 도입하면 식 (12)와 같이 나타낼 수 있다. 여기서 $p(H_k^1|Y_k(t))$는 왜곡된 입력 음성 신호의 스펙트럼 $Y_k(t)$에 대한 음성 존재 확률로서 likelihood ratio $\Lambda_k(t)$와 사전 음성 부재 확률 $p(H_k^0)$을 이용하여 식 (13), 식 (14)와 같이 계산된다.

$$\begin{aligned} \widehat{A_k}(t) &= p(H_k^1|Y_k(t))\; G_{MMSE}(\xi_k(t), \gamma_k(t))R_k(t) \\ &= \frac{\widetilde{\Lambda}_k(t)}{1+\widetilde{\Lambda}_k(t)}G_{MMSE}(\xi_k(t), \gamma_k(t))R_k(t) \end{aligned} \tag{12}$$

$$\begin{aligned} \widetilde{\Lambda}_k(t) &= \frac{p(H_k^1)p(Y_k(t)|H_k^1)}{p(H_k^0)p(Y_k(t)|H_k^0)} = \frac{1-p(H_k^0)}{p(H_k^0)}\frac{p(Y_k(t)|H_k^1)}{p(Y_k(t)|H_k^0)} \\ &= \frac{1-p(H_k^0)}{p(H_k^0)}\Lambda_k(t) \end{aligned} \tag{13}$$

$$\Lambda_k(t) = \frac{1}{1+\xi_k(t)}exp(\upsilon_k(t)) \tag{14}$$

H_k^1과 H_k^0는 각각 음성 신호의 존재와 부재에 대한 가설이며, $p(H_k^1)$는 사전 음성 신호의 존재 확률을 나타낸다. 본 논문에서는 실험을 위해 사전 음성 부재 확률 $p(H_k^0)$를 0.2로, 그리고 forgetting factor α를 0.99로 설정한다 [5].

2.1.3. Wiener Filtering (WF) [3][6]

잡음 신호를 $d(t)$, 원 음성 신호를 $x(t)$, 그리고 왜곡된 입력 음성 신호를 $y(t)$로 가정하면 이들의 관계는 시간 영역에서 식 (1), 그리고 주파수 파워 스펙트럼에서는 식 (4)와 같이 나타낼 수 있었다.

Wiener Filter의 전달 함수는 식 (15)를 이용하여 얻을 수 있다. 여기서 $\sqrt{P_{dk}(t)}$, $\sqrt{P_{xk}(t)}$ 및 $\sqrt{P_{yk}(t)}$는 각각 잡음 신호, 원 음성 신호 및 왜곡된 입력 음성 신호의 t번째 프레임에서 k번째 스펙트럼 빈의 스펙트럼 값을 나타낸다.

$$H_k(t)_{WF} = \frac{\sqrt{P_{xk}(t)}}{\sqrt{P_{xk}(t)}+\sqrt{P_{dk}(t)}} = \frac{\sqrt{P_{yk}(t)}-\sqrt{P_{dk}(t)}}{\sqrt{P_{yk}(t)}} \quad (15)$$

2.1.4. Mel-warped Wiener Filtering (MWF) [7][8]

일반적인 Wiener Filter의 전달 함수 추정은 선형적인 주파수 영역에서 식 (15)를 구성하는 성분들의 추정 오차를 최소값으로 설정하여 얻을 수 있다. 그러나 실제 음성 인식에서 많이 사용하고 있는 MFCC 또는 PLP 특징 파라미터는 청각 인지 주파수 영역에서 추출하고 있다 [8][18]. 그러므로 A. Agarwal [8]는 일반적인 Wiener Filter를 청각 인지 특성과 유사한 멜(mel) 주파수 영역으로 변환을 시도하여 그 유효성을 확인하였다. 그리고 현재 유럽 통신 표준 기구 (ETSI)에서 진행 중인 Aurora 프로젝터에서는 Mel-warped Wiener Filtering (MWF)에 기반한 2단계 MWF 기법 [8]을 표준 잡음 제거 기법으로 채택하고 있다. 본 논문에서는 구현의 복잡도와 인식 성능을 동시에 고려하여 1단계 MWF 기법을 선정한다. 1단계 MWF의 전달 함수의 유도 과정은 다음과 같다.

먼저 식 (16)을 이용하여 선형 주파수 영역에서 mel-warped 주파수 영역으로 변환한다.

$$k_{mel}(i) = i \times \frac{MEL\{k_s/2\}}{M+1} \quad (16)$$

$$MEL\{k\} = 2595 \times \log_{10}(1+k/700) \quad (17)$$

$$k_{cnt}(i) = 700 \times [10^{k_{mel}(i)/2595} - 1], \quad 1 \le i \le M \quad (18)$$

여기서 i는 mel-warped 필터 뱅크의 인덱스, M은 전체 필터 뱅크의 수, k_s는 샘플링 주파수, 그리고 $k_{cnt}(i)$은 i번째 필터 뱅크의 중심 주파수를 나타낸다.

앞에서 설명한 mel-warped 변환 과정을 Wiener Filter의 전달 함수에 적용한다. 이 후 IDCT (Inverse Discrete Cosine Transform)를 이용하여 얻은 Wiener Filter의 임펄스 응답은 식 (19)와 같다.

$$\begin{aligned} h(t) &= \frac{1}{2}\int_0^{\pi} H(k_{mel}(i))\cos(k_{mel}(i)\times t)\, dk_{mel}(i) \\ &\approx \sum_{i=0}^{M} H(k_{mel}(i))\cos\left(\frac{2\pi \times t \times k_{cnt}(i)}{k_s}\right) \\ &\quad \times \left(\frac{k_{cnt}(i+1)-k_{cnt}(i-1)}{k_s}\right), \\ &\quad 0 \le t \le M+1, \;\; 0 \le i \le M+1 \end{aligned} \quad (19)$$

최종적으로 식 (19)에서 얻은 Mel-warped Wiener Filter (MWF)의 임펄스 응답을 잡음 제거에 적용하기 전에 먼저 음의 시간 영역에 해당하는 계수들을 복원하는 미러링 (mirroring) 과정을 수행한다. 또한 필터 적용의 계산량을 줄이기 위하여 인과적 임펄스 응답 계수들의 양쪽 끝부분에서 중요하지 않는 성분을 걸러내는 작업도 동시에 수행한다. 마지막으로 해닝 원도우 (hanning window)를 이용하여 스무딩 (smoothing) 작업을 수행한다. 따라서 필터 길이 $L=17$인 Mel-warped Wiener Filter (MWF)의 임펄스 응답을 얻을 수 있다. 자세한 과정은 식 (20)과 같다.

$$h_{MWF}(t) = Hanning(t) \times h(t) \quad (20)$$

$$Hanning(t) = 0.5 - 0.5\cos(2\pi(t+0.5)/L), \quad 0 \le t \le L-1 \quad (21)$$

2.2. 모델 보상 방식 (model compensation method)

일반적으로 HMM (Hidden Markov Models)을 이용한 음성 인식은 음성 신호의 정확한 특성을 반영하기 위해서 깨끗한 음성 신호를 사용하여 음향 모델을 작성한다. 그러나 실제 환경에서 인식 과정은 음성 신호에 잡음이 부가되어 깨끗한 환경에서 학습된 모델을 이용하여 원 음성 신호의 특징을 정확하게 나타낼 수는 없다. 이러한 문제는 다양한 배경 잡음이 존재하는 음성 신호의 모든 어휘들을 학습함으로써 해결할 수 있으나 계산량이 많고 잡음의 스펙트럼 특성이 바뀔 때 마다 새로운 학습 과정을 반복적으로 수행해야 하는 단점이 있다. 이를 해결하기 위하여 모델 보상 방식의 기법들이 등장하였다. 대표적

인 기법으로 PMC (Parallel Model Combination), VTS (Vector Taylor Series) 및 JA (Jacobian Adaptation) 기법 등이 있다. 본 논문에서는 분석하고자 하는 인식 성능과 계산량을 고려하여 이들 기법 중에서 PMC 기법을 간략히 살펴본다.

2.2.1. PMC (Parallel Model Combination) [9-10]

모델 보상 방식에서 가장 대표적인 기법으로 사용되는 PMC 기법의 처리 과정을 그림 1과 같이 나타낼 수 있다.

그림 1에서 보는 바와 같이 PMC는 켑스트럼 (cepstrum) 영역의 HMM 파라미터들을 선형 주파수 영역으로 변환한 후 깨끗한 음성 HMM 파라미터 값과 잡음 HMM 파라미터 값을 결합하여 왜곡된 음성 HMM 모델을 도출하는 방법이다. 각 단계 별 수행과정은 다음과 같다.

1) 켑스트럼 영역의 HMM 파라미터들을 식 (22)을 이용하여 IDCT (Inverse Discrete Cosine Transformation) 변환을 수행한다.

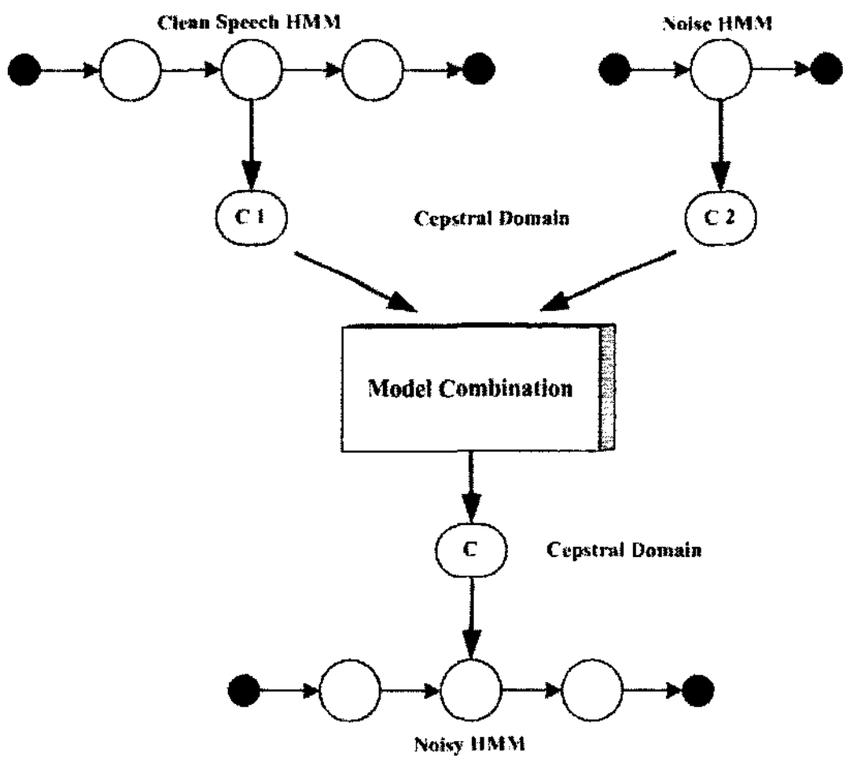


그림 1. PMC 기법의 처리 과정
Fig. 1. Process of Parallel Model Combination.

$$\mu^l = C^{-1}u^c,\ \ \Sigma^l = C^{-1}\Sigma^c(C^{-1})^T \tag{22}$$

여기서, μ^l과 Σ^l는 각각 로그 스펙트럼 영역에서의 평균 벡터와 공분산 행렬이며 μ^c와 Σ^c는 켑스트럼 영역에서의 평균 벡터와 공분산 행렬이다.

2) 지수 함수 변환을 수행하여 로그 스펙트럼 영역의 HMM 파라미터 값을 선형 주파수 영역으로 변환한다. 이 과정을 식 (23)과 식 (24)에 나타낸다.

$$\mu_i = \exp(u_i^l + \frac{\Sigma_{ii}^l}{2}) \tag{23}$$

$$\Sigma_{ij} = \mu_i\mu_j[\exp(\Sigma_{ij}^l) - 1] \tag{24}$$

여기서 μ_i와 Σ_{ij}는 각각 선형 주파수 영역의 파라미터 값인 평균 벡터 μ와 공분산 행렬 Σ의 구성 원소이다.

3) 1)과 2)과정에서 얻어진 깨끗한 음성과 잡음의 HMM 파라미터 값을 선형 주파수 영역에서 식 (25)와 같이 결합한다.

$$\hat{\mu} = \mu + \overline{\mu},\ \ \hat{\Sigma} = \Sigma + \overline{\Sigma} \tag{25}$$

여기서 $\hat{\mu}$와 $\hat{\Sigma}$는 각각 선형 주파수 영역에서의 결합된 평균 벡터와 공분산 행렬이고, 그리고 $\overline{\mu}$와 $\overline{\Sigma}$는 각각 잡음에 관한 평균 벡터와 공분산 행렬이다.

4) 마지막으로 켑스트럼 영역의 잡음이 보상된 음성 HMM의 평균 벡터 $\hat{\mu^c}$와 공분산 행렬 $\hat{\Sigma^c}$을 얻기 위해 먼저 식 (26), 식 (27)과 같이 로그 변환을 수행한 후 식 (28)을 이용하여 DCT 변환을 수행한다.

$$\hat{\mu_i^l} = \log(\hat{\mu_i}) - \frac{1}{2}log(\frac{\hat{\Sigma_{ii}}}{\hat{\mu_i^2}} + 1) \tag{26}$$

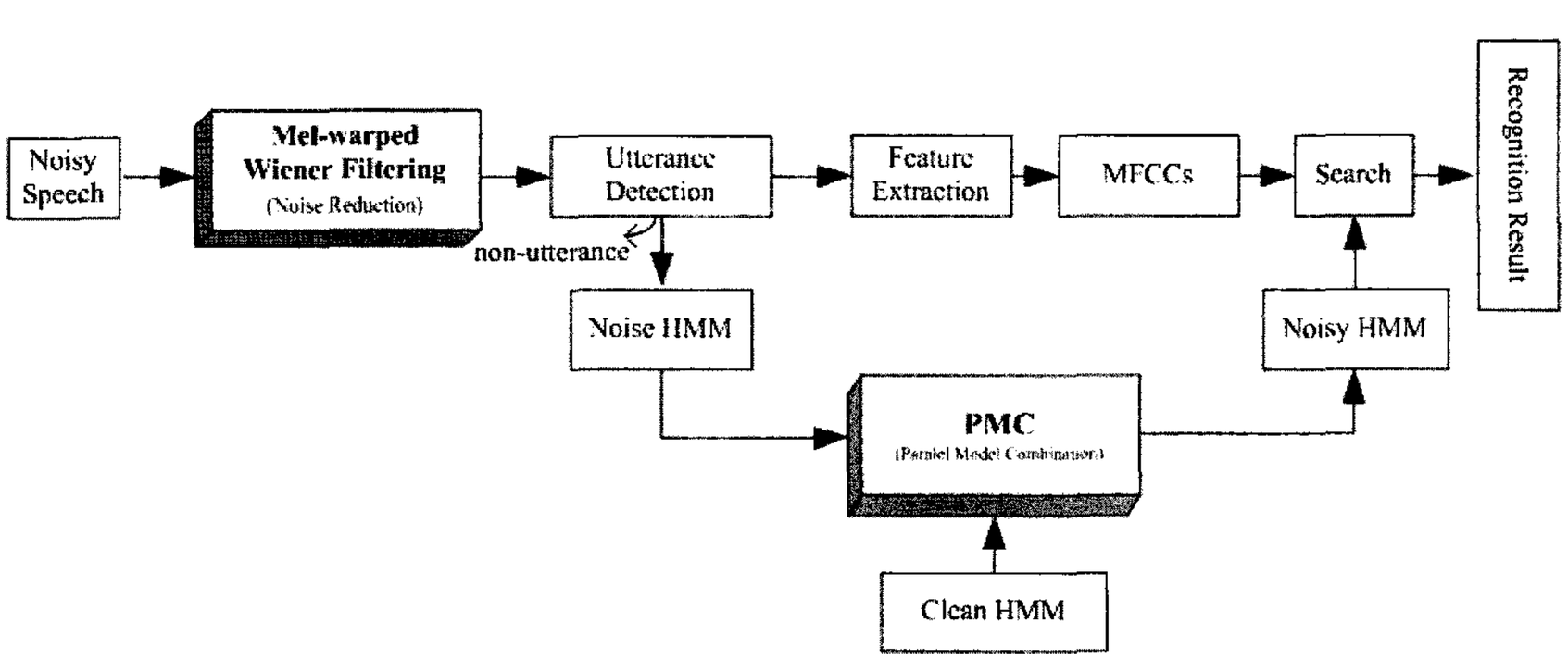


그림 2. MWF-PMC 기법의 처리 과정
Fig. 2. Process of MWF-PMC method.

$$\widehat{\Sigma_{ij}^{l}} = \log\left(\frac{\widehat{\Sigma_{ij}}}{\hat{\mu}_i \hat{\mu}_j} + 1\right) \tag{27}$$

$$\widehat{\mu^{c}} = C\widehat{\mu^{l}} \ , \ \widehat{\Sigma^{c}} = C\widehat{\Sigma^{l}}C^{T} \tag{28}$$

III. 제안한 음성 개선과 모델 보상 결합 방식

3.1. MWF-PMC 기법

앞서 소개한 전처리 단계에서의 음성 개선 방식인 Spectral Subtraction (SS), MMSE-STSA, Wiener Filtering (WF) 및 Mel-warped Wiener Filtering (MWF)기법과 후처리 단계에서의 모델 보상 방식인 PMC 기법은 각각의 장점과 단점을 가지고 있다. 전자의 경우 왜곡된 입력 음성 신호로부터 잡음을 제거하여 음성 신호 자체를 개선하므로 우수한 인식 성능을 기대할 수 있다. 그러나 이 방식은 SNR이 낮은 잡음 음성의 경우, 잡음 제거 과정에서의 음성 신호도 함께 제거되어 왜곡이 많이 발생하기 때문에 이와 같은 신호를 이용한 음성 인식의 성능 향상에는 한계가 있다.

한편 모델 보상 방식은 음성 신호에 부가된 잡음 신호를 채취하여 깨끗한 음성의 HMM을 적응시킨다. 따라서 인식과 훈련의 환경적 불일치 (mismatch)를 감소시킴으로써 인식 성능 향상을 얻을 수 있다. 그러나 이 방식도 낮은 SNR에서는 음성 신호의 특징이 잡음에 의해 왜곡이 많이 발생하기 때문에 그 효력이 급격히 저하되는 문제점을 나타내고 있다. 이를 음성 개선 방식과 병행할 경우 음성 개선 방식에서 발생하는 음성 왜곡에 의한 인식 성능 저하를 감소시킬 수 있을 것으로 기대할 수 있다.

따라서 본 논문에서는 잡음 환경에서 강인한 음성 인식을 위하여 두 가지 방식을 결합하는 새로운 알고리즘을 제안한다. 이는 각 방식의 장점은 취하고 단점은 보완할 수 있는 좋은 모델로 판단된다. 기존의 SS-PMC [13], CSS-PMC [14], MMSESTSA-PMC/JA [4] 및 WF-NOVO [15] 기법들도 이러한 이유 때문에 등장하였다. 본 논문에서는 현재까지 발표된 위 방법들보다 보다 우수한 음성 개선 성능이 기대되는 MWF 기법과 PMC 기법을 결합하는 MWF-PMC 음성 인식 기법을 제안한다.

그림 2는 제안된 MWF-PMC 기법의 처리 과정을 나타내고 있다. 먼저, 음성 HMM은 깨끗한 음성 데이터를 사용하여 작성하며 인식 대상의 잡음 음성 데이터에 대해서는 MWF 기법을 적용하여 잡음을 제거한 후 특징 파라미터를 추출하게 된다. 다음으로 MWF 기법으로부터 개선된 음성 데이터의 묵음 구간에서 잡음을 채취하여 잡음 HMM을 작성하여 PMC 기법으로 잡음이 보상된 음성 HMM을 얻게 된다. 마지막으로 보상된 음성 HMM을 이용하여 음성 인식을 수행하게 된다.

IV. 실험 및 결과

4.1. 음성 데이터 및 특징 파라미터

제안된 MWF-PMC 기법의 유효성을 확인하기 위하여 인식 실험을 실시하고 이 결과를 기존의 결합 방식들과 비교하였다. 인식 실험을 위해서는 국어공학연구소 (KLE)에서 작성한 PBW (Phoneme Balanced Words) 452단어의 1회 발성한 단어 음성 데이터를 8 kHz로 다운 샘플링한 후, 남성 35명을 모델 훈련에, 나머지 3명을 화자 독립 음성 인식 실험 평가에 사용하였다. 잡음 음성 데이터는 다양한 잡음 환경을 고려하여 3종류의 잡음 (Subway, Car, Exhibition)을 5단계의 신호 대 잡음비 (SNR)를 0, 5, 10, 15, 20 dB로 음성 데이터에 부가하여 작성하였다.

인식기는 HM-Net [19]을 기반 인식기를 이용하였으며, 이 인식기의 기본 인식단위는 39개의 PLU (Phoneme Like Unit)로 하였으며 단어 음향 모델은 4 혼합수 (mixture)의 1000 상태 (state)로 구성하였다. 모든 음성 데이터는 8 kHz로 샘플링 된 후 16 bits로 A/D변환 하였으며, 프리엠퍼시스 (Pre-emphasis)필터를 통과한 후 25 ms의 해밍 윈도우 (hamming window)를 이용하여 10 ms씩 이동하면서 분석하였다. 이를 통해 음성 특징 파라미터는 12차의 MFCCs 와 로그에너지 그리고 이들에 대한 delta를 포함하여 총 26차원의 특징 파라미터를 작성하여 인식에 사용하였다.

4.2. 인식 실험 결과

표 1은 3가지 잡음 환경에서 5단계의 SNR 레벨에 따른 기본 인식기의 실험 결과를 나태내고 있다. 표 1로 부터 입력 음성 신호의 SNR이 낮아짐에 따라 인식률도 현저히 저하되는 것을 알 수 있다.

표 2는 본 논문에서 시험한 기존의 음성 개선 기법들에 대한 인식 실험 결과를 나타낸다. 전반적으로 표 1에 비해 인식률이 많이 향상됨을 알 수 있으며, 특히 MWF 기법은 WF 기법과 MMSE-STSA 기법에 비해 Subway 잡음 환경에서 2.6%와 0.7%, Car 잡음 환경에서 5.6%와 2.9%, Exhibition 잡음 환경에서 4.4%와 4.6% 각각 향상됨을

표 1. 3 종류(Subway, Car 및 Exhibition)의 잡음 환경 하에서 SNR 값에 따른 기본 인식기의 인식률

Table 1. Recognition rates of base line recognizer according to 5 SNR levels in 3 kinds of noise (Subway, Car and Exhibition) environments.

SNR \ Noise	Subway	Car	Exhibition	CLEAN
20 dB	90.78	91.89	89.23	
15 dB	72.94	71.39	68.36	
10 dB	43.95	33.63	33.92	98.23
5 dB	19.03	6.12	6.93	
0 dB	1.70	0.59	0.00	
Avg.	45.68	40.72	39.69	98.23

표 2. 3 종류 (Subway, Car 및 Exhibition)의 잡음 환경에서 음성 개선 방식의 각 기법에 대한 인식률 비교

Table 2. Recognition rates of each speech enhancement method in 3 kinds of noise (Subway, Car and Exhibition) environments.

SNR \ Noise	(a) Subway noise			
	SS	MMSE STSA	WF	MWF
20 dB	92.04	95.72	95.80	94.99
15 dB	82.52	89.82	87.17	89.68
10 dB	62.32	72.86	69.03	73.75
5 dB	32.45	43.07	39.53	44.91
0 dB	9.07	11.87	12.76	13.72
Avg.	55.68	62.67	60.86	63.41
SNR \ Noise	(b) Car noise			
	SS	MMSE STSA	WF	MWF
20 dB	95.21	97.27	97.35	96.24
15 dB	89.90	93.44	93.29	93.14
10 dB	67.92	77.95	74.12	81.34
5 dB	29.28	45.80	38.42	52.88
0 dB	4.35	7.23	4.87	12.61
Avg.	57.33	64.34	61.61	67.24
SNR \ Noise	(c) Exhibition noise			
	SS	MMSE STSA	WF	MWF
20 dB	91.89	95.21	94.76	93.22
15 dB	84.07	88.35	88.20	87.98
10 dB	62.39	67.26	68.22	71.83
5 dB	28.02	32.23	32.45	42.55
0 dB	2.73	2.51	2.58	12.83
Avg.	53.82	57.11	57.24	61.68

알 수 있다. 또, 표 2로부터 MWF, MMSESTSA, WF 및 SS 기법 순으로 우수한 인식 결과를 나타내고 있음을 알 수 있으며, 이 중에서 MMSESTSA 기법은 Subway, Car 잡음환경에서 WF 기법에 비해 약간 우수한 성능을 나타내고 있다. SS 기법이 다른 기법들에 비해 가장 저조한

표 3. 3 종류 (Subway, Car 및 Exhibition)의 잡음 환경에서 PMC 기법에 대한 인식률

Table 3. Recognition rates of PMC method in 3 kinds of noise (Subway, Car and Exhibition) environments.

SNR \ Noise	PMC		
	Subway	Car	Exhibition
20 dB	93.29	93.36	91.96
15 dB	91.67	90.27	87.76
10 dB	84.44	78.02	67.77
5 dB	58.33	40.78	30.83
0 dB	22.12	8.85	8.19
Avg.	69.97	62.26	57.30

인식 결과를 나타내고 있는 원인은 잡음 추정의 정확도의 한계 때문으로 생각된다. 한편, MWF 기법은 4가지 기법 중 가장 우수한 평균 인식률을 보이고 있으며, 특히 Car, Exhibition 잡음 환경에서는 매우 우수한 인식 결과를 나타내고 있는 데 이것은 인간의 청각 인지 주파수 기반의 MWF 기법이 잡음 음성 인식에 적합한 기법이라 볼 수 있다. 그러나 5단계의 SNR 레벨에서의 평균 인식률이 70%이하로 잡음 환경에서 효과적인 음성 인식을 위해서는 더욱 향상된 음성 개선 기법의 개발이 요구된다고 할 수 있다.

표 3은 PMC 기법에 의한 인식 결과를 나타낸다. 표 3과 표 2의 결과를 비교해 보면, Subway 잡음 환경에서는 PMC 기법이 모든 음성 개선 방식 기법들에 비해 우수한 성능을 보이고 있으나 Car, Exhibition 잡음 환경에서는 그렇지 못하다. 또, PMC 기법의 평균 인식률은 WF 기법과 거의 동등하게 나타나고 있다. 표 2와 표 3의 결과를 비교해 보면 음성 개선 기법은 SNR 레벨이 15 dB~20 dB의 환경에서, PMC기법은 0 dB~10 dB의 환경에서 비교적 우수한 성능을 나타내고 있음을 알 수 있다. 따라서 이 두 방법을 결합할 경우, 모든 SNR 환경에서 전반적인 인식 성능 향상을 기대할 수 있을 것으로 생각된다.

표 4에 기존의 결합 방식과 본 논문에서 제안하는 MWF-PMC 기법에 대한 인식 결과를 나타낸다. 전반적으로 3종류의 잡음 환경에서 평균 인식률이 MWF-PMC, WF-PMC, MMSESTSA-PMC, SS-PMC 기법 순으로 우수한 인식 결과를 나타내고 있다. 이 중 본 논문에서 제안한 MWF-PMC 기법에 의한 평균 인식률은 73.2%로 음성 개선 방식의 MWF 기법에 의한 평균 인식률 64.1%, PMC 기법에 의한 평균 인식률 63.2%에 비해 약 9%~10%의 인식률 향상을 나타냄을 알 수 있어 그 유효성을 확인할 수 있다. 특히 낮은 SNR 레벨에서의 인식률 개선 효과가 현저함을 알 수 있는 데 이 결과는 낮은 SNR 레벨에서

표 4. 3 종류 (Subway, Car 및 Exhibition)의 잡음 환경에서 각 결합 방식의 인식률

Table 4. Recognition rates of each combined method in 3 kinds of noise (Subway, Car and Exhibition) environments.

Noise / SNR	SS -PMC	MMSESTSA -PMC	WF -PMC	MWF -PMC
	(a) Subway noise			
20 dB	86.95	91.37	91.52	93.29
15 dB	87.17	90.56	90.63	92.70
10 dB	82.74	86.14	86.80	86.50
5 dB	69.40	69.84	75.29	70.50
0 dB	40.86	40.04	49.31	41.37
Avg.	73.42	75.59	78.71	76.87
	(b) Car noise			
20 dB	85.55	90.41	91.30	92.55
15 dB	87.24	87.91	90.41	91.96
10 dB	84.88	81.27	86.28	86.28
5 dB	68.73	60.62	70.94	70.06
0 dB	34.66	22.20	31.49	32.96
Avg.	72.21	68.48	74.09	74.76
	(c) Exhibition noise			
20 dB	83.85	87.32	88.86	91.45
15 dB	80.60	81.12	85.91	88.86
10 dB	70.80	66.08	76.70	79.79
5 dB	51.18	41.45	56.12	56.12
0 dB	24.93	11.36	20.13	24.19
Avg.	62.27	57.46	65.55	68.08

음성 개선 방식을 이용하여 음성 신호의 음질을 개선한 후 묵음 구간으로부터 취한 잡음을 PMC 기법으로 인식 모델에 보상할 경우 잡음에 강인한 인식기를 구성하는 방법이 유효함을 증명하고 있다. 그러나 20 dB의 SNR 레벨에서 결합 방식 기법들의 인식 결과는 표 2의 음성 개선 방식만 사용하였을 때 보다 약간 저하되어 나타남을 볼 수 있다. 이는 음성 개선 과정에서 원 음성 신호에 왜곡이 발생한 결과라 추정된다. 향 후 이 문제점을 해결하기 위해 보다 원 음성 신호의 정보를 왜곡하지 않으면서 개선 효과가 우수한 음성 개선 방식에 대한 연구를 진행할 예정이다.

V. 결 론

본 논문에서는 잡음 환경하의 음성 인식을 위해 음성 개선 방식의 Mel-warped Wiener Filter (MWF) 기법과 모델 보상 방식의 PMC 기법을 결합한 MWF-PMC 기법을 제안하였다. 제안한 MWF-PMC 기법의 유효성을 확인하기 위하여 Subway, Car 및 Exhibition 3종류의 잡음을 KLE 452 단어 음성 데이터에 부가하여 다양한 SNR 레벨 환경에서 인식 실험을 수행하여 기존의 결합 방식들과 비교 평가를 하였다. 인식 실험 결과, 본 논문에서 제안한 MWF-PMC 기법이 기존 결합 방식의 기법들에 비해 전반적으로 향상된 인식 성능을 얻을 수 있었다. 특히, 각 결합 방식은 낮은 SNR 레벨의 잡음 환경에서 인식률 개선이 현저함을 확인할 수 있었다.

참 고 문 헌

1. J. Chen, K. K. Paliwal, S. Nakamura, "Sub-Band Based *Additive Noise Removal for Robust Speech Recognition*," Proc. Eurospeech, 70-73, 2001.
2. H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis of Speech," Proc. JASA, 1738-1752, 1990.
3. S. V. Vaseghi, Advanced Signal Processing and Digital Noise Reduction (Wiley & Teubner Publishers, 1996), Chap. 5, 140-162.
4. 김희근, 정용주, 배건성, "음질향상 기법과 모델보상 방식을 결합한 강인한 음성인식 방식," 음성과학, **14**(2), 115-126, 2007.
5. Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," Proc. ICASSP, **ASSP-32**(6), 1109-1121, 1984.
6. R. J. McAulay, M. L. Malpass, "Speech Enhancement Using A Soft-Decision Noise Suppression Filter," Proc. IEEE Trans. on Acoustic Speech Signal Processing, **28**(2), 1995.
7. ETSI final draft standard doc., "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms," ETSI ES 202 050, v1.1.1, 2002.
8. A. Agarwal, Y. M. Cheng, "Two-Stage Mel-warped Wiener Filter for Robust Speech Recognition," Proc. ASRU, 67-70, 1999.
9. M. J. Gales, S. Young, "An Improved Approach to The Hidden Markov Model Decomposition of Speech and Noise," Proc. ICASSP, I-233-236, 1992.
10. M. J. Gales, S. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination," Proc. Computer Speech and Language, 289-307, 1995.
11. 정용주, 이승욱, "자동차 잡음환경 고립단어 음성인식에서의 VTS와 PMC의 성능비교," 음성과학, **10**(3), 251-261, 2003.
12. 김남수, "잡음 환경에서의 음성인식," Telecommunications Review, **13**(5), 650-661, 2003.
13. J. A. Nolazco Flores, S. Young, "Adapting A HMM-based Recogniser for Noisy Speech Enhanced by Spectral Subtraction," CUED/F-INFENG/TR.123, Cambridge University, England, 1993.
14. J. A. Nolazco Flores, S. Young, "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptatioin," Proc. ICASSP, **1**, 409-412, 1994.

15. K. Satoshi, S. Sumitaka, Y. Yoshikazu, T. Satoshi, "Robust Speech Recognition Based on HMM Composition and Modified Wiener Filter," Proc. ICSLP, 2053-2056, 2004.
16. F. Martin, K. Shikano, Y. Minami, "Recognition of Noisy Speech by Using The Composition of Hidden Markov Models," Proc. ASJ, 1-7-10, 1992.
17. S. Sagayama, Y. Yamaguchi, S, Takahashi, "Jacobian Adaptation of Noisy Speech Models," Proc. ASU, 396-403, 1997.
18. S. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," Proc. IEEE Trans. on Acoustics, Speech, and Signal Processing, ASSP-28(4), 357-366, 1980.
19. S. J. Oh, H. Y. Chung, C. J. Hwang, B. K. Kim, A. Ito, "New State Clustering of Hidden Markov Network with Korean Phonological Rules for Speech Recognition," Proc. IEEE 4th Workshop on Multimedia Signal Processing, 39-44, 2001.

저자 약력

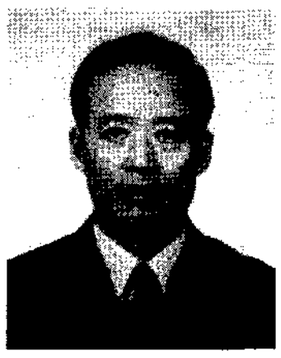
•신 광 호 (Guanghu Shen)

2002년 8월: (중국) 연변대학교 사범대학 수학교육학과 (이학사)
2005년 8월: 영남대학교 정보통신공학과 (공학석사)
2005년 9월~현재: 영남대학교 대학원 정보통신공학과 박사과정
※주관심분야: 잡음처리, 잡음음성인식, 디지털 신호처리

•정 호 열 (Ho-Youl Jung)

1988년 2월: 아주대학교 전자공학과 (공학사)
1990년 2월: 아주대학교 전자공학과 (공학석사)
1993년 2월: 아주대학교 전자공학과 (박사수료)
1998년: (프)리옹국립응용과학원(INSA de Lyon) 전자공학전공(공학박사)
1998년 4월~1998년 12월: (프)CREATIS 박사후 과정
1999년 3월~현재: 영남대학교 전자정보공학부 교수
※주관심분야: 음성, 영상 신호처리, 디지털 워터마킹

•정 현 열 (Hyun-Yeol Chung)



1975년: 영남대학교 전자공학과 (공학사)
1989년: 일본 동북대학교 정보공학과 (공학박사)
1989년 3월~현재: 영남대학교 전자정보공학부 교수
1992년 7월~1993년 7월: 미국 CMU Robotics 연구소 객원연구원
1994년 12월~1995년 2월: 일본 토요하시기술과학대학 외국인 연구자
2000년 6월~2000년 8월: 미국 Qualcomm Inc. 수석 엔지니어
※주관심분야: 음성인식, 화자인식, 음성합성 및 DSP 응용분야