한국음향학회지 제25권 제8호 pp. 370~374 (2006)

포만트 기반의 가우시안 분포를 가지는 필터뱅크를 이용한 멜-주파수 켑스트럴 계수

Mel-Frequency Cepstral Coefficients Using Formants-Based Gaussian Distribution Filterbank

손 영 우*, 홍 재 근*
(Young-Woo Son*, Jae-Keun Hong*)

*경북대학교 전자공학과

(접수일자: 2006년 9월 7일; 수정일자: 2006년 11월 1일; 채택일자: 2006년 11월 17일)

음성인식의 특징벡터로서 멜-주파수 켑스트럴 계수 (MFCC, mel-frequency cepstral coefficients)가 가장 널리 사용되고 있다. MFCC 추출과정은 입력되는 음성신호를 푸리에 변환한 후, 주파수 대역별로 필터를 취하여 에너지 값을 구하고 이산 코사인 변환을 하여 그 계수값을 구한다. 본 논문에서는 멜-스케일된 주파수 대역필터를 취할 때 가중함수에 의해서 구해진 각 대역필터별 가중치를 적용하여 필터의 출력 에너지를 계산한다. 여기서 가중치를 구하기 위해 사용된 가중함수는 포만트가 존재하는 대역을 중심으로 인접한 대역들이 가우시안 분포를 가지는 함수이다. 제안한 방법으로 실험한 결과, 잡음이 거의 없는 음성신호에 대해서는 기존의 MFCC를 사용했을 때와 비슷한 인식률을 보이고 잡음 성분이 많을수록 가중치가 적용된 방법이 인식률에서 보다 높은 성능 향상을 가져온다.

핵심용어: 멜-주파수 켑스트럴 계수, 포만트, 가우시안 분포, 음성인식

투고분야: 음성처리 분야 (2.5)

Mel-frequency cepstral coefficients are widely used as the feature for speech recognition. In MFCC extraction process, the spectrum, obtained by Fourier transform of input speech signal is divided by mel-frequency bands, and each band energy is extracted for the each frequency band. The coefficients are extracted by the discrete cosine transform of the obtained band energy. In this paper, we calculate the output energy for each bandpass filter by taking the weighting function when applying mel-frequency scaled bandpass filter. The weighting function is Gaussian distributed function whose center is at the formant frequency. In the experiments, we can see the comparative performance with the standard MFCC in clean condition, and the better performance in worse condition by the method proposed here.

Key words: MFCC (mel-frequency cepstral coefficients), Formant, Gaussian distribution, Speech recognition

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

음성인식을 이용해 인간과 기계 사이의 의사소통을 위한 연구가 수 십년 동안 계속되고 있다. 음성인식의 방법으로 HMM (hidden Markov model) [1]을 이용한 방법이 여러가지 이유로 인하여 최근까지 가장 널리 사용된다. 패턴인식의 한 분야로 알려진 음성인식은 입력되는 음성신호의 유사한 특징을 보이는 부분들로 나누고 그 각 부분들에 대한 특징을 확률적으로 처리한 후 그 부분에 대한 하나의 패턴을 형성한다. 이러한 패턴 형성을 위해 음성으로부터 추출된 값을 음성의 특징벡터라고 일컫는다. 음성의 특징벡터로는 선형예측계수 (linear prediction coefficients), 켑스트럼 (cepstrum), 필터뱅크 (filterbank), 멜-주파수 켑스트럴 계수 등이 있는데, 이 중 인간의 청각적인 특성을 고려한 필터뱅크를 사용하는 MFCC가 그 우수한 인식 성능으로 인해 현재까지 가장 널리 사용되고 있다. 그러나 음성인식 시스템을 현실적으로 적용하여 인간과 기계사이의 의사소통을 위해서는 무엇보다도 잡음에 강인한 음성인식 시스템이 필요

책임저자: 손 영 우 (syw@ee.knu.ac.kr)
702-701 대구광역시 북구 산격동 1370 경북대학교 전자공학과
(전화: 053-940-8634; 팩스: 053-950-5505)

하다. 잡음에 강인한 인식 시스템의 필요성으로 인하여 여러가지 기술 범주로 나뉘어 많은 연구가 진행되고 있다. 잡음의 문제를 해결하기 위해서 연구되는 기술의 범주는 일반적으로 크게 세가지 정도로 구별된다. 각각의 범주는 잡음에 강인한 특징벡터를 추출하는 방법 [2-5], 잡음이 섞인 음성신호에서 잡음을 제거하여 음질을 향상시키는 방법 [6], 그리고 잡음의 영향으로 손상된 모델을 보상하는 방법 [7] 이 그것이다. 이 중 음질향상을 위해 잡음을 제거하는 방법은 특징벡터를 추출하기 전의 전처리 단계에 해당하며 음성인식 시스템과는 독립적인 하나의 시스템으로 간주할 수 있고 연산량이 많은 기법이다. 또한 모델을 보상하는 방법은 인식모델을 변환하여 적응된 모델이 현재의 잡음이 섞인 음성으로부터 훈련된 것처럼 하는 방법으로 이 방법을 이용할려면 음성인식 시스템의 수정이 필요하고 많은 연산량이 필요한 단점이 있다. 그리고 잡음에 강인한 특징벡터를 추출하는 방법에는 특징벡터를 추출하는 과정에서 잡음의 영향을 제거하는 방법이 일반적이다. 이는 음성인식 시스템의 전처리 과정에서 잡음을 제거하고 특징벡터를 추출하는 방법과 유사하다. 따라서 잡음을 제거하는 과정에서 잡음의 영향을 제거하는 방법이 일반적이다. 이는 음성인식 시스템의 전처리 과정에서 잡음을 제거하고 특징벡터를 추출하는 방법과 유사하다. 따라서 잡음을 제거하는 과정 없이 적은 연산량으로 특징벡터를 추출하는 방법이 필요하다.

본 논문에서는 잡음을 제거하지 않고 잡음에 강인한 특징벡터 추출을 목적으로 한다. 이를 위해 음성이 가지는 고유한 특성을 살펴볼 필요성이 있다. 음성은 일반적으로 독립적 모음이나 자음과 모음의 조합으로 구성된다. 모음신호의 에너지는 자음신호의 에너지와 비교하여 훨씬 크고 신호 자체도 어느정도의 규칙성을 가지지만 자음신호는 규칙성이 거의 없는 잡음과 같은 특성을 가진다. 특히 이러한 음성에 잡음이 섞일 경우에는 자음신호와 잡음과의 구별이 상당히 어렵지만 모음신호는 어느 정도의 잡음에 큰 영향을 받지 않고 구별이 가능하다. 따라서 음성에 잡음이 섞일 경우 자음은 잡음화 되고 모음만 음성으로 간주 될 수 있다. 이러한 이유로 인하여 잡음환경에서는 음성인식 시스템의 특징벡터 추출시 자음의 영향은 거의 없고 모음의 영향에 의해 특징벡터가 추출되기 때문에 음성인식 시스템이 모음에 의존될 수 밖에 없다. 특히 모음의 여러가지 특성 중에서 발음되는 음성에 따라 성도의 변화에 의해 형성되는 포만트는 큰 파워스펙트럼을 가질 뿐만 아니라 잡음이 많이 섞인 음성에서도 뚜렷한 특성을 보인다. 이러한 포만트의 특징을 부각시키기 위해 기존의 MFCC 추출과정에서 사용하는 멜-스케일 된 동일한 진폭특성을 가지는 대역통과 필터를 대신하여 가우시안 (Gaussian) 분포의 가중함수가 부가된 대역통과 필터가 사용된다. 이러한 가우시안 분포의 평균값의 위치는 포만트가 존재하는 대역통과 필터가 되고 그 대역통과 필터를 중심으로 가우시안 분포의 가중치를 가지는 대역통과 필터가 사용된다. 그리고 몇 개의 포만트가 존재하기 때문에 각 포만트 별 가중치의 합으로 전체 대역통과 필터가 형성된다. 이는 포만트가 존재하지 않는 고주파수의 스펙트럼의 영향을 줄이고 포만트가 존재하는 저주파수 영역의 스펙트럼을 강조하는 역할을 한다.

제안한 방법으로 특징벡터인 MFCC를 추출하여 실험을 한 결과 잡음이 존재하지 않은 음성에 대해서는 기존의 MFCC를 사용한 결과와 비슷한 성능을 보이지만 신호대잡음비가 낮은 음성에 대해서는 상당한 성능향상을 보인다.

II. 포만트 기반의 가우시안 분포를 가지는 필터뱅크

일반적인 MFCC 추출과정에서 파워 스펙트럼에 대역통과 필터를 취하여 아래과 같이 에너지를 계산한다.

$$E(k)=\sum_{i} W_k(i)P(i) \tag{1}$$

여기서 $E(k)$는 k 번째 필터의 출력 에너지이고 $W_k(i)$는 k 번째 필터에 대한 i 번째 가중치 값이되며 $P(i)$는 음성의 파워 스펙트럼이다. 식 (1)에서 가중치로 사용되는 $W_k(i)$는 멜주파수 특성을 가지는 삼각형의 주파수 크기 특성을 가지는 필터이다. 식 (1)에서 사용된 대역통과 필터에 다음 식과 같이 가중함수를 사용함으로써 음성의 특성을 부각시킬 수 있다.

$$E(k)=\sum_{i} \alpha(k)W_k(i)P(i) \tag{2}$$

여기서 $\alpha(k)$는 다음 식에서 나타내는 가중함수이다.

$$\alpha(k) = \sum_{n=1}^{m} \frac{\exp[-\frac{(x_{n,k} - \mu_n)^2}{2\sigma^2}]}{\sqrt{2\pi}\,\sigma} + \beta \qquad (3)$$

위의 식을 보면 전체적으로 가우시안 분포의 형태를 가지는 함수가 되고 k 는 필터밴드 위치를 결정하는 값이고 μ_n 은 n 번째 포만트를 가지는 필터밴드의 위치를 나타낸다. $x_{n,k}$ 는 n 번째 포만트를 가지는 필터밴드를 중심으로 k 번째 필터의 위치를 나타낸다. σ^2 는 가우시안 분포의 분산을 의미하고 β 는 동적범위를 제한하기 위한 기준값을 나타낸다. 그리고 m 은 전체 포만트의 개수 이다. 음성구간에서 가중치를 계산하기 위해 우선적으로 포만트가 존재하는 기준 필터밴드를 구하는 것이 필요하다. 포만트를 구하기 위한 여러가지 방법[7]이 있으나 본 논문에서는 정확한 포만트 주파수가 필요한 것이 아니라 포만트가 존재하는 대역의 필터밴드의 위치가 중요하기 때문에 선형예측계수 (LPC)를 이용한 정확한 포만트 주파수 검출방법을 이용하지 않고 입력 음성의 스펙트럼으로부터 최고점 검출방법을 이용하여 포만트의 위치를 검색하였다. 포만트가 존재하는 필터밴드의 위치를 기준으로 하여 인접한 대역통과 필터의 가중치를 식 (3)을 이용하여 구할 수 있다. 가우시안 분포를 가지는 가중치를 계산할 때 가우시안 분포의 분산이 중요한 요소가 됨을 알 수 있다. 분산이 너무 작을 경우는 포만트가 존재하는 대역에만 영향을 미치게 되어 포만트가 존재하지 않는 대역의 스펙트럼 특성이 지나치게 왜곡되고, 분산이 너무 클 경우에는 전대역에 걸쳐 가중치가 부가되게 됨에 따라 가중함수를 사용하지 않는 것과 큰차이가 없기 때문이다. 이러한 가우시안 분포의 가중함수를 적용하면 포만트가 존재하는 비교적 주파수가 낮은 대역에서는 높은 가중치를 가지게 되지만, 포만트가 존재하지 않는 높은 주파수 대역에서는 가중치가 낮아지게 됨에 따라 높은 주파수 성분이 많이 존재하는 잡음의 영향이 줄어드는 역할이 된다. 그림 1에서 가우시안 분포의 가중함수가 적용된 대역통과 필터의 주파수 크기특성의 예를 나타낸다. 여기서 임의의 포만트 두 개만을 이용하고 σ 를 4로 설정한 대역통과 필터의 특성이 다음 그림과 같이 나타난다. 각 포만트에 의해 결정된 가중치를 각 대역통과 필터들에 적용하고 그 합을 그 대역의 가중치로 결정한다.

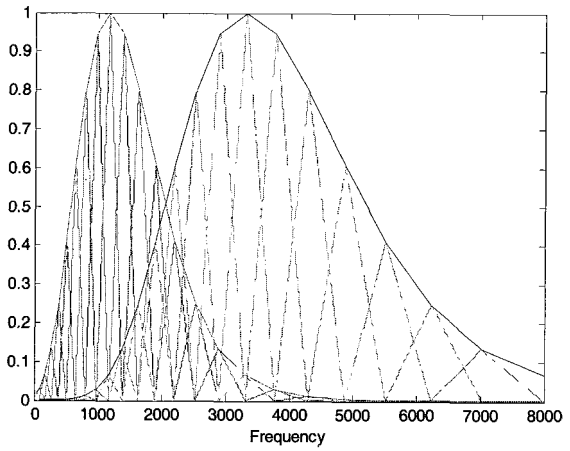

그림 1. 가우시안 분포의 가중함수를 가지는 대역통과 필터의 예
Figure 1. Example of bandpass filters with weighting funcion of Gaussian distribution.

III. 실험 및 결과

포만트가 존재하는 대역통과 필터를 기준으로 가우시안 분포의 가중함수를 적용하도록 제안한 MFCC 추출 과정의 전체 블록 다이어그램은 그림 2와 같다. MFCC

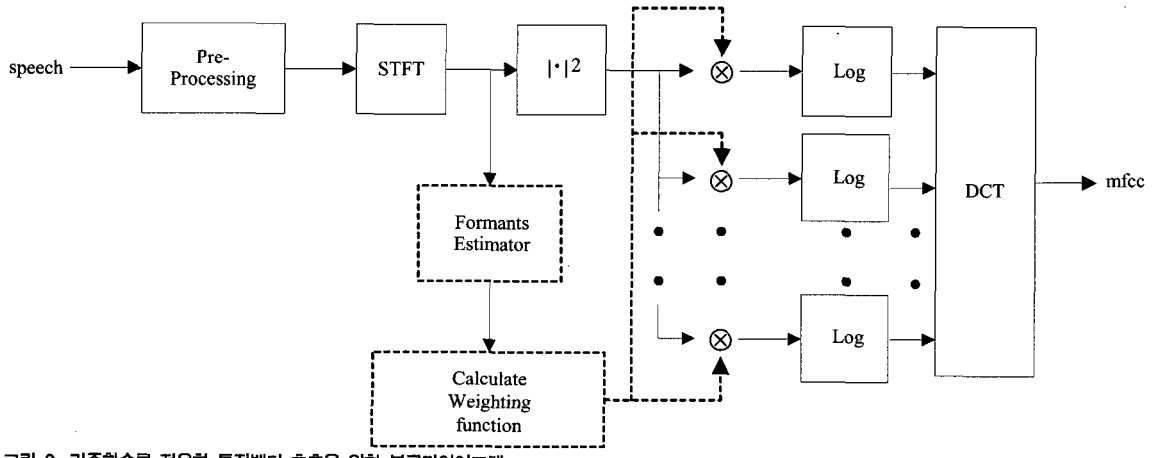

그림 2. 가중함수를 적용한 특징벡터 추출을 위한 블록다이어그램
Figure 2. The blockdiagram for feature extraction using weighting function.

추출 과정에서 제안한 부분은 점선으로 표시되었다. 그림 2에서 보듯이 전처리 과정을 거친 입력 음성이 푸리에변환 (FFT) 된 후 최고점 검출법으로 포만트의 위치를 찾고 검출된 포만트를 기준으로 식 (3)에 표현된 각 대역별 가중함수인 $\alpha(k)$ 를 결정한다. 식 (3)으로 구해진 $\alpha(k)$ 를 식 (2)의 가중치로 사용하여 각 대역별 에너지를 구한다.

그림 3은 깨끗한 음성과 약 10dB 정도의 백색잡음이 섞인 음성의 크기 스펙트럼을 나타낸다. 그리고 수직선은 스펙트럼을 이용하여 구한 포만트 주파수의 위치를 표시한다.

그림 3에서 볼 수 있듯이 잡음이 섞인 음성의 스펙트럼은 잡음으로 인해 높은 주파수 성분의 에너지가 증가하게 되어 신호대잡음비가 낮아질수록 포만트의 검출은 보다 어려워진다. 포만트 검출이 잘못 이루어 질 경우 스펙트럼의 심한 왜곡이 발생하여 정보의 손실이 발생할 수 있다. 따라서 본 논문에서는 3번째 포만트가 존재할 가능성이 있는 3kHz 주파수 이상에서 검출되는 포만트는 무시한다. 즉, 최대 포만트 수를 3개로 제한하되 그 포만트 주파수가 제한 주파수인 3kHz 이하일 경우에만 포만트가 검출된 것으로 판단한다. 기준이 되는 포만트 대역을 중심으로 식 (3)에 나오는 가우시안 분포의 가중함수를 계산할 때 가중치의 범위를 제한하기 위하여 가우시안 분포의 분모 부분은 계산에서 제외하였다. 그리고 β 값은 가중함수의 최저값으로서 그 값이 크질수록 가우시안 분포함수의 영향을 거의 받지 않는 대역에서도 필터특성이 기존의 필터 특성과 거의 같아지게 되어 정보의 손실은 없어지지만 잡음성분이 많이 포함되어 있는 고주파수 성분의 감소가 없기 때문에 인식률의 향상을 기대하기 어렵다. 반면 그 값이 작아질 수록 포만트가 존재하는 대역과 그 안접한 대역의 정보만 남고 고주파수 영역에서의 정보는 거의 없어지기 때문에 β 의 설정이 중요하다. 실험에서는 분산과 β 값으로 5와 0.1을 사용하였다.

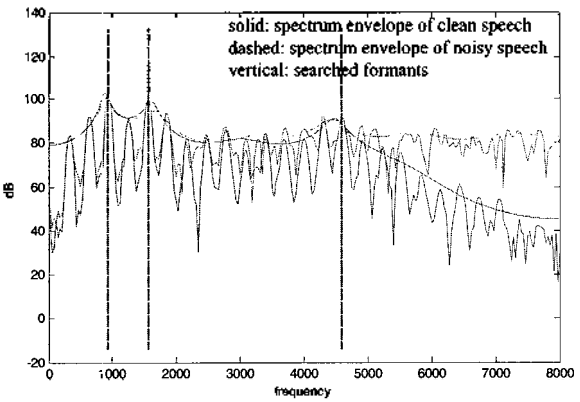

그림 3. 신호대잡음비에 따른 음성의 포만트와 주파수 특성
Figure 3. Formants and frequency responses of speech according to SNR.

그림 2의 블록다이어그램을 기반으로 특징벡터를 추출 하고 다음과 같은 실험 환경에서 인식실험을 실시하였다. 실험에 사용된 음성의 데이터베이스는 한국 전자통신 연구원 (ETRI, Electronics and Telecommunications Research Institute)에서 제공하는 445DB를 사용하였다. 실험에 사용된 445DB는 40명의 남녀 화자가 445개의 단어를 각 2번씩 발음한 고립단어 한국어이다. 샘플링률이 16kHz인 음성신호에 대해 한 프레임을 20ms로 정하고 한 프레임 길이의 해밍 (Hamming) 윈도우를 취하였다. 그리고 프레임 이동률을 10ms로 하였다. 실험에 사용된 특징벡터는 제안한 방법의 필터뱅크를 사용하여 각 밴드별 에너지를 구하고 DCT (discrete consine transform)를 통해 13차의 특징벡터가 얻어지면 0차의 켑스트럼 성분을 제외한 12차의 MFCC에 1차의 로그에너지를 추가하여 13차의 MFCC 특징벡터를 구성한다. 그리고 이것의 delta 성분 13차와 acceleration 성분 13차를 구하여 최종적으로 분석 프레임당 총 39차의 MFCC 특징벡터를 구하게 된다. 깨끗한 음성 데이터베이스로부터 추출 MFCC를 바탕으로 HTK ver 3.2.1 [9]의 훈련절차를 이용하여 tri-phone모델을 생성하였다. 각 모델은 7개의 상 (state)와 상태당 6 mixtures의 HMM으로 이루어져 있다. 인식실험은 위에서 언급한 데이터베이스를 사용하여 HTK 인식시스템을 이용하였고 백색잡음이 부가된 신호대 잡음비에 따라 남녀 각 3명씩의 화자에 대해 추출된 특징벡터를 이용하여 라운드 로빈 (round robin) 방식으로 실시하였다. 위의 조건으로 실험을 실시한 결과는 표 1과 그림 4에 나타나 있다.

표 1과 그림 4에서 볼 수 있듯이 기존의 방법으로 추출한 MFCC와 제안한 방법으로 추출한 MFCC를 비교해

표1. 기존의 방법과 제안한 방법의 인식률 비교
Table 1. Recognition rates of convential method and the proposed method.

Recognition Rate(%)		
SNR(dB)	Conventional MFCC	Proposed MFCC
Clean	96.06	95.26
30	92.54	94.40
25	84.46	92.65
20	62.13	88.57
15	28.73	75.83
10	8.63	48.02
Average	62.09	82.46

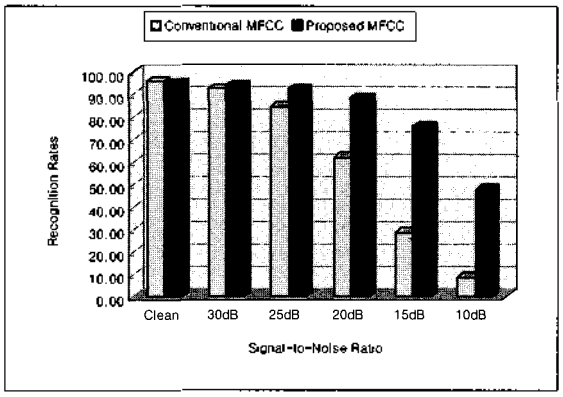

그림 4. 기존의 방법과 제안한 방법의 인식률
Figure 4. Recognition rates of conventional method and the proposed method.

볼 때 깨끗한 음성일 경우는 비슷한 인식률을 나타내고 있지만 신호대잡음비가 작아질수록 제안한 방법의 성능이 우수함을 알 수 있다.

IV. 결 론

본 논문에서는 모음의 여러가지 특성 중에 잡음에 의해 크게 영향을 받지 않는 포만트 성분을 강조하여 잡음 환경에서의 강인한 음성인식을 위한 특징벡터 추출을 목적으로, 포만트가 존재하는 멜-스케일된 대역을 중심으로 대역별 가중치를 부가하는 방법을 제안하였다. 대역별 가중치는 포만트가 존재하는 대역을 중심으로 가우시안 분포의 형태를 가지는 가중함수를 각 대역별로 계산하여 구하였으며, 이렇게 구해진 가중치를 적용한 대역통과 필터를 이용하여 제안한 방법의 MFCC를 구하였다. 제안한 방법은 전체적으로 인식률에 있어서 성능향상을 가졌으며, 특히 신호대잡음비가 작아질수록 높은 성능향상을 보였다. 제안한 방법의 가장 중요한 요소인 포만트를 신호대잡음비가 아주 낮은 잡음 환경에서도 명확하게 검색할 수 있다면 보다 높은 인식률 향상을 가져오게 될 것이다. 그리고 다양한 음성신호의 데이터베이스를 바탕으로 여러가지 특성을 가지는 잡음에 대한 연구가 필요하다.

참 고 문 헌

1. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, 77(2) 257-286, Feb. 1989.
2. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am 87 1738-1752, April 1990.
3. K. K. Chu, S. H. Leung and C. S. Yip, "Perceptually non-uniform spectral compression for noisy speech recognition", Proc. ICASSP 2003, 404-407, 2003.
4. K. K. Chu, S. H. Leung, "Feature extraction based on perceptually non-uniform spectral compression for speech recognition", Proc. ISCAP 2003, 726-729, 2003.
5. K. K. Chu and S. H. Leung, "SNR-dependent non-uniform spectral compression for noisy speech recognition", Proc. ICASSP 2004, 973-976, 2004.
6. P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection for robust speech recognition in cars", Speech Communication, 11 215-228, June 1992.
7. M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise", Speech Communication, 12 231-239, 1993.
8. L. Welling and H. Ney, "Formant estimation for speech recognition", IEEE Trans. On Speech and Audio Processing, 6(1) Jan. 1998.
9. S. Young, D. Kershaw, J. Odell, D. Ollason, and P. Woodland, *The HTK Book version 3.2.1*, 2002.

저자 약력

• 손 영 우 (Young-Woo Son)

1995년 2월 : 경북대학교 전자공학과 졸업(공학사)
1997년 2월 : 경북대학교 대학원 전자공학과 공학석사
1997년 ~ 2001년 : Siemens VDO Halla 연구원
2001년 ~ 2002년 : Voiceware 연구원
2003년 ~ 현재 : 경북대학교 대학원 전자공학과 박사과정
※ 주관심 분야 : 음성인식, 음성신호처리

• 홍 재 근 (Jae-Keun Hong)

1975년 2월 : 경북대학교 전자공학과 졸업(공학사)
1979 년 2월 : 경북대학교 대학원 전자공학과 공학석사
1985년 2월 : 경북대학교 대학원 전자공학 공학박사
1979년 3월 ~ 1983년 2월 : 경북전문대학 교수
1983년 4월 ~ 현재 : 경북대학교 전자전기공학부 교수
※ 주관심 분야 : 음성신호처리, 음향신호처리