잡음 환경에서의 음성 인식을 위한 청각 표현

Auditory Representations for Robust Speech Recognition in Noisy Environments

김 도 석*, 이 수 영*, 길 이 만**
(Doh-Suk Kim*, Soo-Young Lee*, Rhee M. Kil**)

요 약

본 논문에서는 잡음 환경에서의 음성 인식을 위한 전처리기로서 청각 모델을 제안하였다. 제안된 청각 모델은 와우각 대역 통과 필터와 비선형단으로 구성되어 있으며, 잡음 환경에서도 신호의 주파수 정보와 강도 정보를 효과적으로 표현할 수 있다. 주파수 정보는 신호의 영교차 간격에 의해서, 또 강도 정보는 피크 검출기와 포화 비선형 함수에 의해서 구해진다. 영교차 간격이 교란되는 양의 분산을 교차 레벨 값의 함수로 표현함으로써 영교차 간격을 사용하는 것이 레벨 교차 간격에 비해 잡음에 둔감한 특성이 있음을 보였다. 제안된 청각 모델은 다른 청각 모델에 비해 계산량이 적고, 미리 많은 파라미터를 정해줄 필요가 없다. 화자 독립 격리단어 인식 실험 결과 제안된 방법은 잡음 환경에서 우수한 성능을 보였다.

ABSTRACT

An auditory model is proposed for robust speech recognition in noisy environments. The model consists of cochlear bandpass filters and nonlinear stages, and represents frequency and intensity information efficiently even in noisy environments. Frequency information of the signal is obtained by zero-crossing intervals, and intensity information is also incorporated by peak detectors and saturating nonlinearities. Also, the robustness of the zero-crossings in estimating frequency is verified by the developed analytic relationship of the variance of the level-crossing interval perturbations as a function of the crossing level values. The proposed auditory model is computationally efficient and free from many unknown parameters compared with other auditory models. Speaker-independent speech recognition experiments demonstrate the robustness of the proposed method.

I. 서 론

음성 인식 시스템은 학습된 환경과 실제 사용되는 환경이 달라지면 인식 성능이 저하되게 되며, 이러한 환경의 불일치의 한 요인으로는 주변의 잡음을 들 수 있다. 음성 인식 시스템이 사용될 수 있는 환경에는 여러 가지 잡음이 존재하게 되며, 실용화를 위해서는 이러한 여러 잡음 환경에서도 인식 성능이 크게 저하되지 않아야 한다. 음성 인식 시스템에서 잡음의 영향을 제일 먼저 받는 부분은 음성 신호로부터 인식에 필요한 특징 파라미터를 뽑아내는 전처리 단계이다. 잡음에 둔감한 음성 인식을 위한 전처리기는 잡음 환경에서도 인식에 유용한 부분만을 취하고 인식에 불필요한 정보와 음성 신호에 내재되어 있는 변이는 줄일 수 있어야 한다. 현재 가장 널리 사용되고 있는 특징 파라미터는 LPC-켑스트럼이며, 이는 음성의 발성 기관을 모델링하는 것에 그 기반을 두고 있다. 그러나 음성 인식을 위해서는 음성 인지 과정을 모델링하는 것이 음성 발성 과정을 모델링하는 것보다 더 자연스럽고, 또한 바람직할 것이며, 실제로 인간은 인공적인 시스템에 비하여 잡음이 존재하는 환경에서 매우 뛰어난 인식 능력을 지니고 있다. 이러한 생물학적 청각 기관의 기능을 모델링함으로써 잡음에 강한 특징 추출기를 설계하고자 하는 연구는 꾸준히 계속되고 있다[1, 2, 3, 4, 5, 6]. 청각 모델링은 생물학, 심리 음향학, 물리학, 전자공학 등의 많은 분야의 협력이 요구되는 분야이다. 그러나 대부분의 연구는 실험 자체에 심하게 의존하고 있으며, 청각 시스템의 다단계 비선형성 때문에 분석적인 접근 방법이 매우 어려운 문제점이 있다.

본 논문에서는 잡음 환경에서도 잡음이 없는 경우와 마찬가지로 유용한 특징 파라미터를 추출하기 위하여 생물학적 청각 기관에 근거를 둔 청각 모델을 제안하였다. 그리고 제안된 방법으로부터 얻어진 특징 파라미터가 배경 잡음에 둔감한 특성이 있다는 것을 해석적인 방법과 실험적인 방법으로 입증하였다. 논문의 구성은 다음과 같

*한국과학기술원 전기 및 전자공학과
**한국과학기술원 기초과학부

접수일자 : 1996년 7월 5일

다. Ⅱ절에서는 제안된 청각 모델인 Zero-Crossings with Peak Amplitudes(ZCPA)에 대하여 설명하고, ZCPA와 Ensemble Interval Histogram(EIH)[4, 5]를 비교 고찰하였다. Ⅲ절에서는 레벨값의 크기가 레벨 교차 간격에 미치는 영향에 대해 확률적인 분석을 제시하였고, 이 분석에 의해 잡음이 섞여 있는 신호의 주파수을 추정함에 있어 영교차를 사용하는 것이 레벨 교차를 사용하는 것보다 우수함을 보였다. Ⅳ절에서는 화자 독립 음성 인식 실험을 통해 제안된 방법이 다른 특징 추출 기법보다 잡음에 둔감함을 보였고, Ⅴ절에서 결론을 맺었다.

Ⅱ. Zero-Crossings with Peak Amplitudes(ZCPA) 모델

2.1 모델의 구조와 특징 추출

음향 신호는 청각 시스템에 의해 인지에 필요한 형태로 변환된다. 음향 신호에 의한 공기의 진동은 외이를 거쳐 고막을 진동시키게 되며, 고막과 외이에서의 기계적인 진동은 중이와 oval window를 거쳐 와우각 내부의 액체 흐름으로 변환된다. 와우각은 가느다란 관이 달팽이 모양으로 말려 있는 기관이며, 이 관은 basilar membrane에 의해 길이 방향으로 분할되어 있다. 이 basilar membrane의 길이 방향으로는 섬모세포(hair cell)들이 나열되어 있고, 이 섬모세포들은 auditory nerve fiber에 연결되어 대뇌로 신호가 전달된다. 외부의 자극에 의한 와우각 내부의 파동은 basilar membrane 상에서 진행파를 형성하게 되며, 이 진행파가 최대가 되는 위치는 주파수에 따라 달라지게 된다. 중이에 가까운 부분에서는 고주파가, 그리고 끝으로 갈수록 저주파 성분이 잘 검출된다. 즉 basilar membrane에는 일종의 대역 통과 필터군이 존재하여 음성 신호를 주파수 분할하는 것으로 알려져 있다. 진행파와 basilar membrane에서의 기계적인 진동은 내섬모세포(inner hair cell)에 달려있는 섬모의 운동을 유발하게 되며, 이는 auditory nerve fiber의 신경 발화(neural firing)로 변환된다. 이 신경 발화는 cochlear nucleus, inferior colliculus, medial geniculate body등의 여러 세포군을 거쳐 대뇌의 auditory cortex로 전달된다[7, 8]. 이 세포군들에서는 인지에 필요한 복잡한 특징 검출기가 존재하는 것으로 보이지만, 아직 많은 부분에 대한 정보는 거의 알려져 있지 않다.

그림 1은 제안된 Zero-Crossings with Peak Amplitudes (ZCPA) 모델의 블럭도이다. 이 시스템은 청각시스템의 auditory nerve fiber까지를 모델링한 것이며, 대역통과 와우각 필터뱅크와 각 와우각 필터의 출력단에 연결되어 있는 비선형 변환단으로 구성되어 있다[9].

와우각 필터뱅크는 basilar membrane의 기계적인 운동과 주파수 선택성을 모델링한 것이며, 본 논문에서는 Kates가 제안한 진행파 필터[10]가 사용되었다. 그림 2의 블럭도와 같이 이 필터는 와우각 내에서의 진행파와, basilar membrane과 tectorial membrane간의 복잡한 상호 작용에 의한 필터링을 선형 필터들의 직렬연결로 모델링한 것이다. 이 진행파 필터는 Lyon과 Mead[11]에 의해 제안된 아날로그 와우각 모델에 근거를 두고 있으며, 실제 생물학적 시스템의 관측 결과와 비슷하도록 설계되었다. 각 와우각 필터의 전달함수는

$$Coch_i(z) = H_{hp,i}(z)\, F_i(z) \prod_{k=1}^{i} H_k(z) \qquad (1)$$

와 같이 표현할 수 있다. 여기서 $H_k(z)$는 진행파 필터의 한 구간을 나타내는 저역 통과 필터이다. $H_k(z)$의 공진 주파수는 Greenwood에 의해 제안된 basilar memberane 상에서의 위치와 주파수 관계식[12]

$$F = A(10^{ax} - 1) \qquad (2)$$

에 의해 200 Hz에서 5000 Hz까지 분포되어 있다. 여기서 F는 Hz 단위로 나타낸 주파수이고, x는 basilar membrane

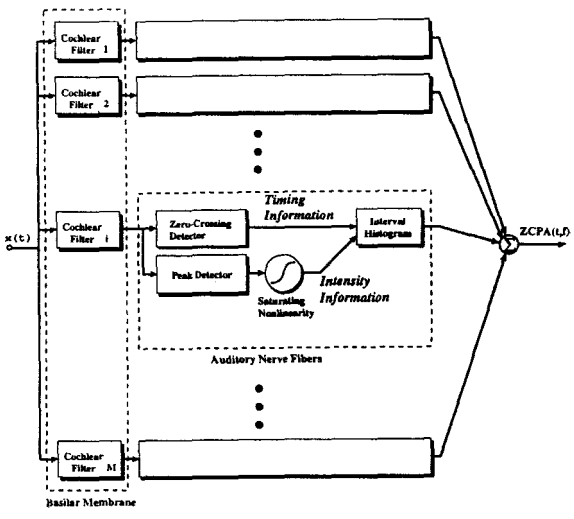


그림 1. ZCPA 모델의 블럭도
Fig 1. Block diagram of the ZCPA model.

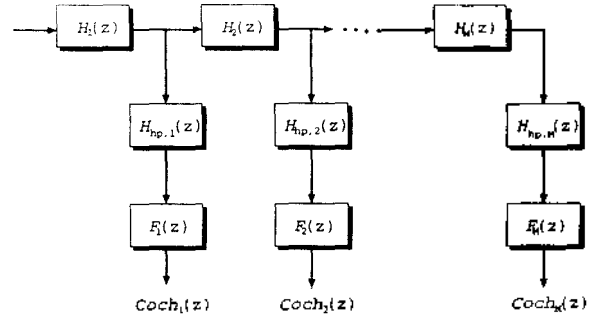


그림 2. 진행파 필터로 구성된 와우각 필터뱅크의 블럭도
Fig 2. Block diagram of the cochlear filters.

에서의 규준화된 거리로 0에서 1사이의 값을 갖는다. 인간의 특성에 적절한 상수값으로 $A=165.4$와 $a=2.1$이 사용되었다.

$H_{hp,i}(z)$는 압력을 속도로 변환시켜 주는 극점이 하나인 고역 통과 필터이고, $F_i(z)$는 최종 응답이 생물학적 시스템의 응답과 같이 이중 공진 특성을 보이도록 해주는 너치 필터이다. 이 필터들의 자세한 설계 방법은 [10]에 기술되어 있으며, 본 논문에서 사용된 20개의 필터의 주파수 응답 특성을 그림 3에 나타내었다. 각 응답 특성은 저주파쪽에 비해서 고주파쪽이 매우 급격한 경사를 이루고 있는 비대칭 특성을 보인다. 또한 저주파에 비해 고주파쪽의 필터가 날카로운 공진 특성을 가지고 있다.

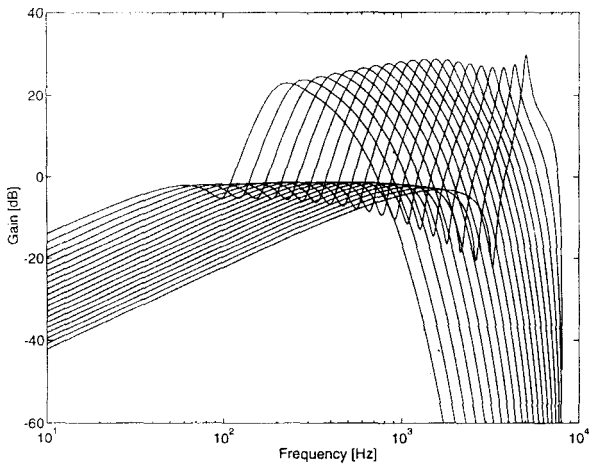


그림 3. 20개 와우각 필터의 주파수 응답 특성
Fig 3. Frequency responses of 20 cochlear filters.

Auditory nerve fiber에서 음성 신호가 어떻게 표현되는가 하는데에는 두가지 이론이 존재하며, 그중 temporal representation 이론에 의하면 auditory nerve fiber는 자극 신호에 동기되어 반응을 하게 되며, 이 발화 패턴은 신호의 주파수 특성을 반영하게 된다[13, 14, 15, 16, 17, 18, 19]. 또한 발화율은 자극의 강도가 증가함에 따라 커지게 되며, 어느 정도 이상이 되면 포화하는 특성을 보인다[20]. 이는 신호의 주파수와 강도 정보가 auditory nerve fiber에 의해 신경 신호로 표현된다는 것을 의미한다. 제안된 청각 모델에서는 각 대역통과 필터를 거친 신호가 영점을 교차하는 점에서 신경 발화가 일어나는 것으로 청각 신경 섬유가 신호에 동기되어 반응하는 특성을 모델링 하였다. 영교차점은 신호가 증가하는 방향, 즉 영점에서의 미분값이 양수일 경우만 고려되며, 신호의 영교차가 발생할 때마다 인접한 두 교차점간의 시간 간격을 측정하고, 그 역수를 주파수 히스토그램에 누적시킨다. 그리고 자극 신호의 강도와 청각 신경 섬유의 발화율과의 관계 특성을 반영하기 위하여, 인접한 영교차점간의 신호 피크값을 검출하여 비선형 함수를 통과한 값을 주파수 히스토그램에 가중치로 사용한다. 최종 ZCPA의 출력은 모든 채널의 히스토그램을 더함으로써 얻어지며, n번째 프레임의 ZCPA의 출력은

$$y(n,i)=\sum_{channel}\sum_{k=1}^{K-1}\delta_{ij_k}\,g(A_k),\qquad 1\le i\le M \tag{3}$$

와 같이 표현할 수 있다. 여기서 K는 각 채널에서 상향 영교차점의 갯수를, M은 주파수 히스토그램에서의 주파수 분할구간의 갯수를, j_k는 k번째와 $(k+1)$번째 영교차점으로부터 계산된 주파수 구간의 index이다. 또 A_k는 k번째와 $(k+1)$번째 영교차점 사이 신호의 피크값이고, δ_{ij}는 Kronecker 델타, $g()$는 센 자극에 대해서 발화율이 포화되는 것을 반영하기 위한 단조 함수이다. ZCPA 동작의 이해를 돕기 위해 정현파 신호가 입력되었을 때의 한 예를 그림 4에 나타내었다.

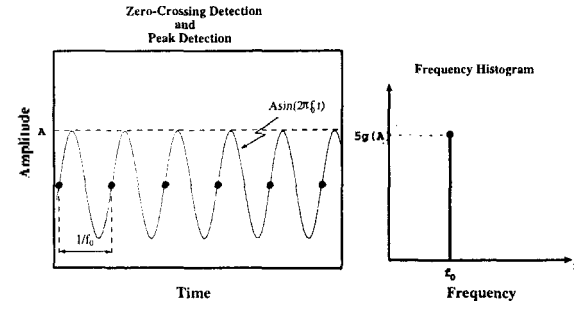


그림 4. ZCPA 동작의 한 예
Fig 4. An example of the operation of the ZCPA.

한 채널의 출력신호가 $A\sin(2\pi f_0 t)$와 같은 정현파라고 가정하면, 그림 4에 나타난 것과 같이 여섯번의 영교차점이 검출되고, 인접한 영교차점 사이의 시간 간격은 모두 f_0의 역수가 된다. 영교차점이 검출될 때마다 현재 영점과 인접한 과거 영점간의 간격을 측정하고, 주파수 히스토그램에서 그 역수에 해당하는 주파수 구간의 값을 증가시킨다. 증가시킬때 영점간의 피크값(이 예에서는 A)을 비선형 함수 $g()$를 통과시킨 $g(A)$ 만큼의 양을 증가시키게 된다. 여기서는 다섯번의 영교차 간격이 측정되고 이 간격의 역수는 모두 f_0이므로 최종 히스토그램의 값은 f_0 구간에서만 $5g(A)$가 된다.

ZCPA 출력의 계산을 위해서는 각 채널마다 유한한 길이의 신호가 고려되어야 한다. 각 채널 와우각 필터의 중심 주파수를 F_c라고 할때, 시간 t에서의 ZCPA 출력을 얻기 위해서는 구간 $[t-10/F_c,\ t]$ 만큼의 신호가 고려된다. 이렇게 되면 저주파 채널에서는 긴 신호가, 고주파 채널에서는 상대적으로 짧은 신호가 사용되므로 저주파 부분에서는 주파수 분해도가 좋아지는 반면 시간 분해도가 나빠지게 되고, 고주파 부분에서는 이와 반대의 특성이

나타난다. 이는 실제 사람에게서 나타나는 특성과 일치한다.

2.2 EIH 모델과의 비교

Ensemble Interval Histogram(EIH)는 Ghitza에 의해 제안된 청각 모델의 일종이다[4, 5]. 이 모델은 비교적 계산량이 적고 잡음 환경에서의 음성인식 결과 우수한 성능을 보여주었다. EIH는 와우각 필터뱅크와 여러개의 레벨교차 검출기들로 구성되어 있다. 기본적으로 EIH와 ZCPA 모두 신호의 교차점을 이용함으로써 주파수 정보를 추출한다. 그러나 EIH는 ZCPA와 달리 여러 레벨값을 갖는 레벨교차 검출기를 이용함으로써 신호의 강도 정보를 얻게 된다. 여러 레벨교차 검출기는 하나의 내섬모세포에 연결되어 있는 auditory nerve fiber를 모델링 한 것이다. 각 레벨은 그 고유의 문턱값을 지닌 auditory nerve fiber를 의미하고, 레벨값들은 모두 양수이며 대수적으로 균등하게 분포되어 있다. 각 레벨교차 검출기는 신호의 레벨교차점을 검출하고, 인접한 교차점간의 시간 간격을 히스토그램에 누적하게 된다. 따라서 같은 신호에 대해서 한 채널 내부에서도 독립적인 여러 타이밍 정보가 존재한다. 만약 신호의 진폭이 증가하게 되면 여러 레벨검출기가 동작하게 되고, 그만큼 히스토그램에 누적되는 값이 커지게 되어 신호의 강도 정보를 표현할 수 있게 된다.

그러나 EIH는 레벨값과 레벨 갯수 등 여러 파라미터를 결정해주어야 하는 단점이 있다. 레벨과 관련된 파라미터가 EIH에 어떤 영향을 미치는지 정성적으로 간략히 살펴보면 다음과 같다. 만약 레벨값들이 모두 0에 가깝다면 레벨교차 검출기로부터 얻어지는 타이밍 정보는 많은 부분이 중복될것이며, 신호의 세기 정보는 제대로 표현되지 못하게 된다. 반면에 레벨값이 신호의 크기에 비해 너무 높다면 높은 레벨값을 갖는 레벨교차 검출기는 쓸모없이 될 것이다. 그러므로 이 적당한 레벨 갯수와 레벨값을 결정해주는 것이 매우 중요한 문제가 된다. 그러나 아직까지 이를 위한 이론이나 분석이 되어있지 못한 상태이며, 따라서 여러번의 시행착오를 거쳐 결정해주어야만 한다. 더우기, 보다 근본적인 문제는 높은 레벨교차 검출기로부터 얻어지는 타이밍 정보는 잡음이 존재하는 환경에서는 부정확하게 된다는 점이다.(이 특성은 Ⅲ절에 나타내었다.) 제안된 청각 모델은 영교차만을 사용해 주파수 정보를 얻게 되므로 잡음에 둔감한 특성이 있으며, 신호의 강도 정보 또한 이용하게 된다.

Ⅲ. 레벨값과 레벨교차 간격과의 관계

EIH와 ZCPA 모두 신호의 교차점을 측정함으로써 스펙트럼을 추정하게 된다. 이러한 접근 방법은 특히 신호에 부가 잡음이 존재하면 레벨값에 따라 그 성능이 크게 영향을 받게 된다. 레벨값이 증가할수록 추정되는 스펙트럼은 부가 잡음에 민감하게 된다. 이를 입증하기 위해서 식 (4)와 같은 입력 신호를 고려해 보자.

$$x(t)=\sum_{i=0}^{M-1} A_i\cos(\omega_i t+\theta_i)+gv(t) \tag{4}$$

여기서 $v(t)$는 대역폭이 W [rad/sec]으로 제한된 백색 가우시안 잡음이며 평균과 분산이 모두 1이라고 하자. 고정된 A_i에 대해서 신호대잡음비는 g에 의해 결정된다. 이 신호가 필터뱅크를 통과한 후에 각 정현파 성분이 분리되었다고 가정하면 각 필터의 출력은 대역통과된 잡음 성분만 존재하거나, 또한 하나의 정현파와 대역통과된 잡음 성분이 더해진 형태 두가지 중의 하나일 것이다. k 번째 필터의 출력이 후자의 경우라면

$$x_k(t)=A_i\cos(\omega_i t+\phi_i)+gv_k(t) \tag{5}$$

와 같이 표현될 수 있다. 그림 5에 나타난 것 같이 상향 레벨교차점, 즉 $x_k(t_n)=l$이 되는 시간을 t_n이라고 하고, 잡음에 의해 교차점이 이동한 시간, 즉 교란된 양을 r_n이라고 하자. 만일 $A_i \gg g$라면 dominant frequency principle[21]에 의해 레벨교차 간격의 평균은 $2\pi/\omega_i$로 근사될 수 있다.

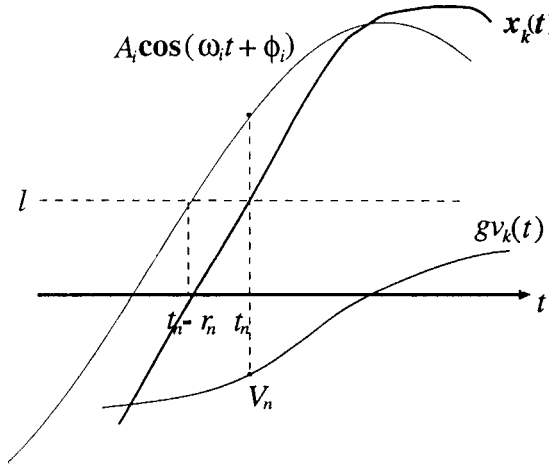


그림 5. 대역통과필터 출력에서의 신호와 잡음 성분
Fig 5. Signal and noise component at the output of a bandpass filter.

V_n을 t_n에서의 대역통과 잡음의 amplitude라고 하면 그림 5로부터 식 (6)과 (7)의 관계를 얻을 수 있다.

$$A_i\cos(\omega_i t_n+\phi_i)=l-V_n \tag{6}$$

$$A_i\cos(\omega_i(t_n-r_n)+\phi_i)=l \tag{7}$$

다시 $\alpha=\omega_i t_n+\phi_i$, $\beta=\cos^{-1}(l/A_i)$로 놓으면 식 (8)과 같은 관계가 성립한다.

$$\omega_i \tau_n = \alpha - \beta \tag{8}$$

교차점의 미분이 양수인 경우만 고려하므로 β의 범위는 $3\pi/2 \le \beta \le 2\pi$로 제한되고, 식 (8)로부터 식 (9)를 얻게 된다.

$$\cos(\omega_i \tau_n) = \frac{l - V_n}{A_i} \cdot \frac{l}{A_i} + \left[\left(1 - \left(\frac{l - V_n}{A_i} \right)^2 \right) \left(1 - \left(\frac{l}{A_i} \right)^2 \right) \right]^{1/2} \tag{9}$$

만약 $\omega_i \tau_n$가 매우 작고 $R = l/A_i$라고 한다면, 식 (9)는 식 (10)과 같이 근사된다.

$$\tau_n^2 \approx \frac{2}{\omega_i^2} \left[1 - R \left(R - \frac{V_n}{A_i} \right) \right] - \frac{2}{\omega_i^2} \left[1 - \left(R - \frac{V_n}{A_i} \right)^2 \right) (1 - R^2) \Bigg]^{1/2} \tag{10}$$

이제 잡음에 의해 두 인접한 레벨교차점이 각각 τ_n과 τ_{n+1}만큼 교란되었다면 레벨교차점간의 시간 간격은 $|\tau_n - \tau_{n+1}|$만큼 교란될 것이다. 랜덤변수 τ_n과 τ_{n+1}의 평균이 0이고 서로 상관관계가 없다면 이 교란의 분산은

$$\sigma_i^2 = E\{|\tau_n - \tau_{n+1}|^2\} = E\{\tau_n^2\} + E\{\tau_{n+1}^2\} \tag{11}$$

와 같다. $E\{V_n\} = 0$이므로, 식 (10)으로로부터 식 (12)를 얻을 수 있다.

$$E\{\tau_n^2\} \approx \frac{2}{\omega_i^2} (1 - R^2) - \frac{2}{\omega_i^2} E\left\{ \left[(1 - R^2) \left(1 - \left(R - \frac{V_n}{A_i} \right)^2 \right) \right]^{1/2} \right\} \tag{12}$$

한편 랜덤변수 X의 평균을 η_X, 표준편차를 σ_X^2, 확률 밀도 함수를 $f_X(x)$라고 할때 Taylor 급수 전개에 의해

$$E\{h(x)\} = \int_{-\infty}^{\infty} h(x) f_X(x) d(x) \approx h(\eta_X) + h''(\eta_X) \frac{\sigma_X^2}{2} \tag{13}$$

가 성립하므로 식 (12)의 우변의 두번째 항은 $V_n/A_i \ll 1$일 때 $E\{V_n^2\} = Bg^2/W$를 이용하여 다음과 같이 근사될 수 있다.

$$E\left\{ \left[(1 - R^2) \left(1 - \left(R - \frac{V_n}{A_i} \right)^2 \right) \right]^{1/2} \right\} \approx [1 - R^2] - \left[\frac{1}{A_i^2} \left(1 + \frac{R^2}{1 - R^2} \right) \cdot \frac{1}{2} \left(\frac{B}{W} g^2 \right) \right] \tag{14}$$

식 (11), (12), (14)로부터 인접한 두 레벨교차점간 시간간격 교란의 분산은 식 (15)와 같이 된다.

$$\sigma_i^2 = \frac{2g^2 B/W}{(\omega_i A_i)^2} \cdot \frac{1}{1 - (l/A_i)^2} = \sigma_{i_0}^2 \frac{1}{1 - (l/A_i)^2} \tag{15}$$

여기서 $\sigma_{i_0}^2$는 영교차일 경우의 분산이다. 이제 시간간격의 신호대잡음비는 다음과 같이 정의할 수 있다.

$$SNR_i \equiv \frac{2\pi/\omega_i}{\sigma_i} = \left(\frac{A_i \pi}{g} \right) \left[\frac{2W}{B} (1 - (l/A_i)^2) \right]^{1/2}. \tag{16}$$

시간간격 교란의 분산은 $l = 0$일때 최소값 σ_{i_0}가 된다. 주어진 A_i와 g에 대해서 레벨값 l이 증가함에 따라 분산은 증가하고 신호대잡음비는 감소하게 된다. 그러므로 높은 레벨교차 검출기로부터 추정되는 주파수는 부가잡음에 더 민감하게 되는 특성이 있다. 이 관계를 그림 6에 보였다.

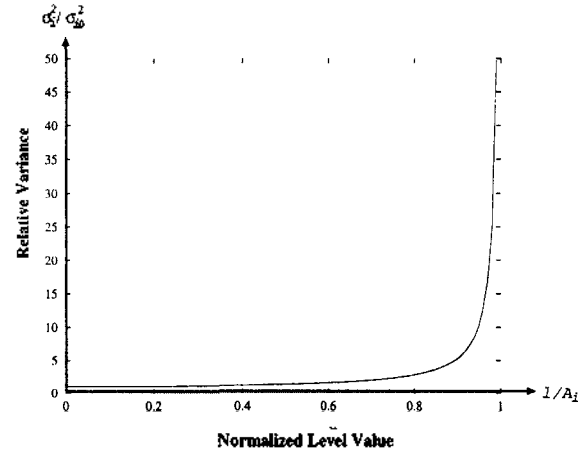


그림 6. 레벨값과 레벨교차 간격 교란의 분산과의 관계
Fig 6. Relation between level value and the variance of level-crossing interval perturbation.

Ⅳ. 실험 결과

4.1 실험 조건

제안된 특징추출 방법의 잡음 환경에서의 성능을 평가하기 위하여 화자 독립 단어 인식 실험을 수행하였다. 실험에 사용된 음성 데이타는 음운학적으로 균형을 이룬 75단어로 이루어져 있으며, 조용한 사무실 환경에서 Sennheiser HMD224X headset을 통해 16 kHz, 16 bit로 A/D변환되었다[22]. 20명의 화자가 발음한 분량을 다섯명이 한조가 되도록 하여 4조로 나누었으며, 이중 3조를 인식 실험의

기준 패턴으로 하고 나머지 1조를 테스트 패턴으로 사용하였다. 데이타의 분량이 비교적 적으므로, 보다 신뢰도 있는 결과를 얻기 위하여 기준 패턴과 테스트 패턴의 조합을 각각 다르게 하여 4가지 서로 다른 인식 실험을 하였다. 잡음은 백색 가우시안 잡음을 원하는 신호대 잡음비에 맞도록 크기를 조절하여 음성에 더해주었는데, 여기서 신호대 잡음비는 단어 전체를 다 고려한 global 신호대 잡음비를 사용하였다[4].

EIH와 ZCPA에 사용된 필터뱅크는 20개의 와우각 필터로 구성되었다. 히스토그램은 0에서 5000 Hz까지의 주파수 범위를 critical-band rate[23]에 따라 18개의 구간으로 나누었으며, 각 주파수 구간의 경계주파수는 f를 주파수(kHz)라고 할때 바크 z의 범위 $1.5 \le z \le 18.5$에서 식 (17)과 같이 결정된다.

$$f = \left(\frac{\exp(0.219z)}{354} + 0.1 \right) z - 0.032 \exp(-0.15(z-5)^2) \quad (17)$$

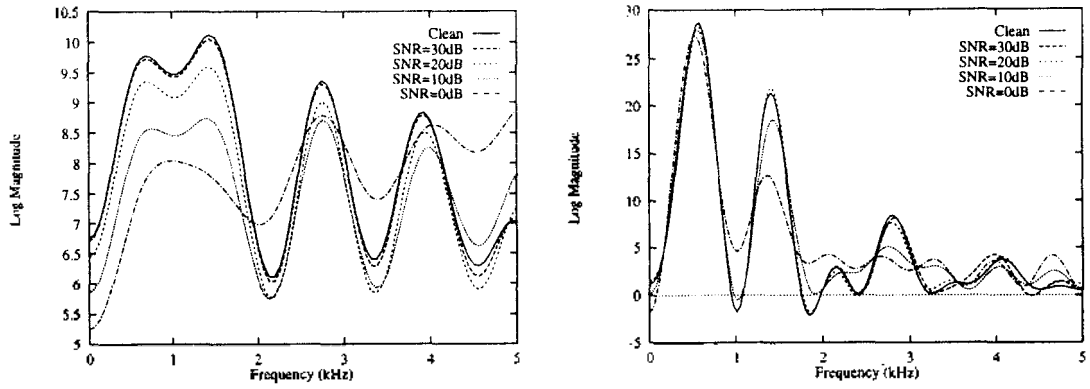
인식기로는 nearest neighbor 인식기를 사용하였으며, 음성의 시간 변이를 흡수하기 위하여 trace-segmentation algorithm[24, 25]을 적용하였다.

4.2 ZCPA와 LPC-켑스트럼의 비교

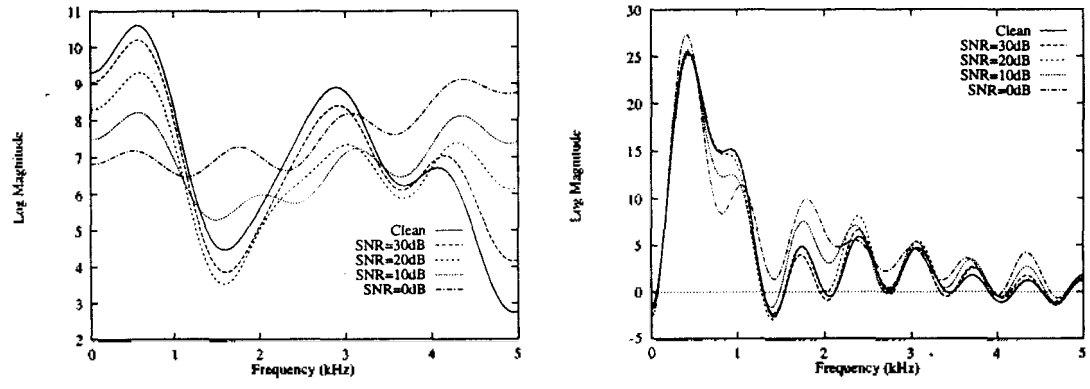
켑스트럼을 이용하여 모음 부분의 스펙트럼을 smoothing 시킨 결과를 그림 7에 나타내었다. ZCPA는 일종의 log 스펙트럼이라고 간주할 수 있으므로 그 출력을 역푸리에 변환시킴으로써 켑스트럼을 구하였다. 각 스펙트럼은 20차 켑스트럼으로부터 smoothing되어 glottal pulse에 의한 급격한 변화 성분이 제거되고 발성 기관에 의한 스펙트럼 부분만 반영되었다. ZCPA의 경우가 LPC-켑스트럼에 비해 잡음 환경에서도 스펙트럼의 변이가 비교적 적은 특성이 있음을 알 수 있다.

4.3 하나의 레벨을 사용한 EIH의 결과

그림 8은 각 채널마다 하나만의 레벨교차 검출기가 있는 EIH의 인식 실험 결과를 레벨값과 신호대 잡음비를



(a) 모음 /a/



(b) 모음 /o/

그림 7. 여러 신호대 잡음비에서 구해진 모음 (a)/a/, (b)/o/의 smoothing된 스펙트럼. 왼쪽의 그림들은 LPC-켑스트럼으로부터, 오른쪽은 ZCPA-켑스트럼으로부터 구한 것이다.

Fig 7. Smoothed Spectrum of the vowels (a)/a/, (b)/o/ at various SNRs. Each left plot is the cepstral curve fit of the LPC-derived cepstrum, and each right plot is the cepstral curve fit of the ZCPA.

변화시키면서 측정한 결과이다. 레벨값은 수평축에 나타내었으며, 레벨값이 증가함에 따라 인식율이 저하됨을 알 수 있다. 레벨값이 (1/2)0.02·2^{15}까지는 - 이 값은 음성 신호가 취할 수 있는 최대값 2^{15}의 1%이다. - 비교적 조용한 환경에서 인식율의 차이는 크게 나지 않지만 잡음 정도가 심해지면 인식율이 저하된다. 이는 레벨값이 커질수록 레벨교차가 잡음에 민감해지므로 스펙트럼이 제대로 추정되지 못했다는 것을 의미한다. 따라서 레벨교차를 이용함에 있어 낮은 레벨을 사용하는 것이 바람직하다고 할 수 있다.

한편, 낮은 신호대 잡음비에서 레벨값이 0.02·2^{15}이상으로 증가함에 따라 인식율이 오히려 증가하는 경우가 있다. 이 경우는 음성 신호에서 모음과 같이 국소적으로 신호대 잡음비가 높은 부분만이 특징벡터로 표현되고, 레벨값보다 작은 부분은 무시되었기 때문이다. 이 때문에 오히려 높은 레벨값에서도 인식율이 증가하는 부분이 발생하지만 이때에는 작은 신호 성분이 무시됨으로 인해 정보의 손실이 생기며 이로 인해 조용한 환경에서는 인식율이 저하되게 된다.

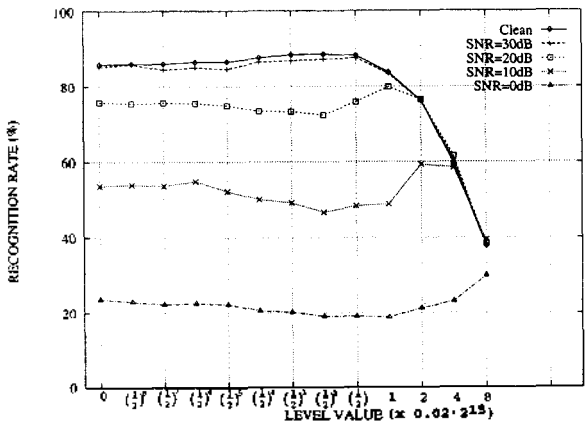


그림 8. 여러 신호대 잡음비에서 하나의 레벨교차 검출기만 갖는 EIH의 인식 결과(%)

Fig 8. Recognition rates (%) of the EIHs with single levels at various SNRs.

4.4 여러 레벨을 사용한 EIH의 결과

그림 9는 EIH에서 레벨값과 레벨갯수를 변화시켜 가며 인식율을 측정한 결과이다. EIH의 레벨값은 log2 scale상에서 균등하게 분포되어 있으며, "L"뒤에 오는 첫번째 숫자는 레벨 갯수를, 두번째 숫자는 레벨값 분포의 범위를 나타낸다. 두번째 숫자가 클수록 레벨값들이 낮게 설정되어 있다는 것을 의미한다. 예를 들면 L3.1의 가장 높은 레벨값은 L3.3의 가장 높은 레벨값보다 4배가 높다. 그리고 서로 다른 레벨 갯수에 대해 두번째 숫자가 같다는 것은 가장 높은 레벨값들이 같다는 것을 의미한다. 예로 L5.7의 최상위 레벨값은 L3.7과 L7.7의 최상위 레벨값과 같으며, 따라서 L5.7은 L3.7보다는 2개의 낮은 레벨이 더 있는 것이 된다.

레벨값들이 낮게 설정될수록 또 레벨 갯수가 증가할수록 EIH의 인식율이 증가하는 경향이 있다. 그러나 레벨값들이 지나치게 작은 경우는 낮은 레벨들로부터 얻어지는 정보는 중복될 것이며, 이때에는 레벨 갯수가 감소된 것과 같아져서 인식율이 오히려 약간 감소하게 된다. 또한 그림 8과 그림 9를 비교해 보면 L7.1의 인식율이 영교차만 사용했을 경우보다 더 나쁘지만 L7.7의 인식율은 영교차에 비해 clean, 30dB, 20dB의 경우 각각 2%, 2.2%, 3.4% 높다. 이 결과들로부터 EIH의 성능은 레벨 갯수와 레벨값에 크게 영향을 받으며, 레벨값들이 적절히 결정되었을 때에는 여러 레벨을 사용하는 경우가 영교차만 사용하는 경우보다 우수한 인식율을 얻게 되지만, 그렇지 못할 때에는 오히려 성능 저하를 가져올 수 있음을 알 수 있다.

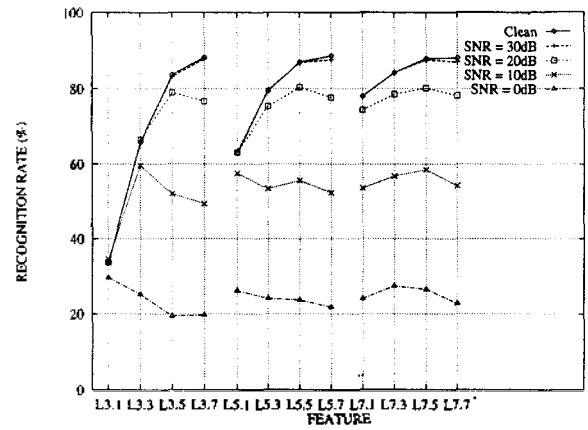


그림 9. 여러 신호대 잡음비에서 EIH의 인식율(%). "L"뒤에 오는 첫번째 숫자는 사용된 레벨교차 검출기의 갯수를, 두번째 숫자는 레벨값들이 얼마나 낮게 분포되어 있는지를 나타낸다.

Fig 9. Recognition rates (%) of the EIHs with multiple levels at various SNRs. The first digit following the "L" denotes the number of levels used in the EIH, and the second digit denotes how low the level values are distributed.

4.5 ZCPA의 결과와 다른 특징 파라미터와의 비교

ZCPA의 성능과 다른 특징 추출 방법과의 비교를 표 1에 나타내었다. LPC-켑스트럼의 차수는 12에서 18까지 변화시키며 실험하여 여기에서는 가장 좋은 18차일때의 결과를 나타내었다. ZC(Zero-Crossing)는 EIH에서 레벨교차 검출기를 하나만 사용하고 레벨값을 0으로 한 경우이다. EIH는 7개의 레벨을 사용한 경우중에서 가장 나쁜 예(L7.1)와 가장 좋은 예(L7.5)를 함께 나타내었다. EIH (L7.1)는 최상위 레벨값이 신호가 취할수 있는 최대값의

6.4%, EIH(L7.5)는 0.4%로 되어 있다.

LPC-켑스트럼의 인식 성능은 잡음이 심해질수록 크게 저하되며 EIH나 ZCPA에 비해 잡음에 매우 민감하다. ZCPA와 EIH(L7.1)을 비교하면 ZCPA가 각 잡음 조건에서 7.2%에서 11.1% 높은 인식율을 보였다. 또 EIH(L7.5)보다는 20dB에서 1.5%, 10dB에서 6.2% 높은 인식율을 보여 ZCPA가 부가 잡음이 있는 환경에서 우수한 성능을 보임을 알 수 있다.

표 1. 여러 신호대 잡음비에서 ZCPA와 다른 특징추출 방법과의 인식율(%) 비교

Table 1. Comparison of recognition rates (%) of the ZCPA with other features at various SNRs.

Feature	Clean	30dB	20dB	10dB	0dB
LPC cepstrum	86.8	73.9	37.1	12.5	3.3
ZC	85.9	85.3	75.7	53.7	23.6
EIH (L7.1)	78.0	77.9	74.4	53.6	24.1
EIH (L7.5)	87.9	87.5	80.1	58.5	26.6
ZCPA	88.3	86.8	81.6	64.7	34.1

V. 결 론

본 논문에서는 생물학적 청각 기관에 근거를 둔 특징 추출 방법을 제안하였다. 제안된 ZCPA 모델은 와우각 필터뱅크와 비선형단으로 구성되어 있다. 와우각 필터뱅크는 basilar membrane에서의 주파수 선택성을, 비선형단은 자극 신호에 동기되어 반응하는 auditory nerve fiber를 모델링 하였다. 비선형단은 영교차 검출기와 피크 검출기, 그리고 포화 비선형 함수로 이루어져 있다. EIH에서 사용되는 레벨교차 검출보다 영교차 검출이 잡음에 둔감한 특성이 있음을 해석적인 방법과 실험적인 방법으로 입증하였다. ZCPA는 다른 청각 모델에 비해 비교적 계산량이 간단하고 많은 파라미터들을 결정해 주지 않아도 된다는 잇점이 있다. 화자 독립 단독음 인식 실험 결과 제안된 특징 추출 방법이 기존의 방법보다 우수한 성능을 나타내었다.

ZCPA는 매우 단순화된 청각 모델이며, 적응 기능 등 생물학적으로 관측되는 결과들을 고려함으로써 성능을 향상시킬 수 있으리라 기대된다. 또한 청각 시스템의 깊은 이해와 모델의 향상을 위해서는 수학적인 분석과 입증이 매우 중요하다. ZCPA는 신호 처리 관점에서 보면 대역 통과된 신호를 영교차 표현한 것이라고 간주할 수 있다. 대역 통과된 신호의 영교차는 의미있는 표현이며, 만약 대역이 제한된 신호가 차수 n인 다항식으로 근사될 수 있으면 다시 이 다항식은 영점들이 각 항에 나타나는 곱셈식으로 표현될 수 있다. 다항식의 영점들이 모두 실수인 경우에 이 신호는 영점과 가장 높은 차수의 계수 a_n만으로 표현이 가능하게 된다. 주기적인 신호의 경우에 그 신호의 실수인 영점만으로 신호를 복원할 수 있으며 임의의 두 DFT 계수를 계산해 낼 수 있다. 비주기적인 신호의 경우에는 근사적으로 신호를 복원할 수 있으며 [26], Sreenivas와 Niederjohn[27]은 영교차점의 통계적인 특성을 이용하여 스펙트럼을 분석하는 방법을 제안하였다. 이 관점은 앞으로 ZCPA의 특성을 설명하고 분석하는 하나의 제시가 될수 있으며, 이를 통해 더욱 성능이 향상된 청각 모델의 개발을 기대할 수 있을 것이다.

참 고 문 헌

1. J. B. Allen, "Cochlear modeling", *IEEE-ASSP Magazine*, pp. 3-29, 1985.
2. J. R. Cohen, "Application of an auditory model to speech recognition", *J. Acoust. Soc. America*, vol. 85, pp. 2623-2629, 1989.
3. S. Seneff, "A joint synchrony/mean-rate model of auditory processing", *J. Phonetics*, vol. 16, no. 1, pp. 55-76, 1988.
4. O. Ghitza, "Auditory nerve representation as a basis for speech processing", in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds., pp. 453-485, Marcel Dekker, New York, 1992.
5. O. Ghitza, "Auditory models and human performances in tasks related to speech coding and speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, part II, pp. 115-132, 1994.
6. K. Wang and S. A. Shamma, "Self-normalization and noise-robustness in early auditory representations", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 421-435, 1994.
7. E. B. Goldstein, *Sensation and perception*, Wadsworth Publishing Company, 1989.
8. Stephen Handel, *Listening: An introduction to the perception of auditory events*, The MIT Press, 1993.
9. D. S. Kim, J. H. Jeong, J. W. Kim, and S. Y. Lee, "Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments", in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, May 1996, pp. 61-64.
10. J. M. Kates, "A time-domain digital cochlear model", *IEEE Trans. Signal Processing*, vol. 39, no. 12, pp. 2573-2592, 1991.
11. R. F. Lyon and C. Mead, "An analog electronic cochlea", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 7, pp. 1119-1134, 1988.
12. D. Greenwood, "A cochlear frequency-position function for several species-29 years later", *J. Acoust. Soc. America*, vol. 87, no. 6, pp. 2592-2650, 1990.
13. M. B. Sachs and E. D. Young, "Encoding of steady state vowels in the auditory-nerve: representation in terms of discharge rate", *J. Acoust. Soc. America*, vol. 66, pp. 470-479, 1979.
14. E. D. Young and M. B. Sachs, "Representation of steady-

state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers", *J. Acoust. Soc. Am.*, vol. 66, no. 5, pp. 1381-1403, 1979.

15. B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: I.", *J. Acoust. Soc. America*, vol. 75, no. 3, pp. 866-878, 1984.
16. B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: III.", *J. Acoust. Soc. America*, vol. 75, no. 3, pp. 887-896, 1984.
17. B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: IV.", *J. Acoust. Soc. America*, vol. 75, no. 3, pp. 897-907, 1984.
18. B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: V.", *J. Acoust. Soc. America*, vol. 75, no. 3, pp. 908-918, 1984.
19. M. B. Sachs, C. C. Blackburn, and E. D. Young, "Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus", *Journal of Phonetics*, vol. 16, pp. 37-53, 1988.
20. M. B. Sachs and P. J. Abbas, "Rate versus level functions for auditory nerve fibers in cats: Tone burst stimuli", *J. Acoust. Soc. America*, vol. 56, no. 6, pp. 1835-1847, 1974.
21. B. Kedem, "Spectral analysis and discrimination by zero-crossings", *Proc. IEEE*, vol. 74, no. 11, pp. 1477-1493, November 1986.
22. 최인정, 권오욱, 박종렬, 김도영, 정호영, 은종관, "자동통역용 한국어 음성 데이타베이스", 음성 통신 및 신호 처리 워크샵 논문집, 1994, pp. 287-290.
23. E. Zwicker and E. Terhart, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", *J. Acoust. Soc. America*, vol. 68, pp. 1523-1525, 1980.
24. H. F. Silverman and N. R. Dixon, "State constrained dynamic programming (SCDP) for discrete utterance recognition", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, 1980, pp. 169-172.
25. D. S. Kim and S. Y. Lee, "Intelligent judge neural network for speech recognition", *Neural Processing Letters*, vol. 1, no. 1, pp. 17-20, 1994.
26. S. M. Kay and R. Sudhaker, "A zero crossing-based spectrum analyzer", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 1, pp. 96-104, Feb. 1986.
27. T. V. Sreenivas and R. J. Niederjohn, "Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise", *IEEE Trans. Signal Processing*, vol. 40, no. 2, pp. 282-293, 1992.

▲김 도 석(Doh-Suk Kim) 1968년 2월 23일생
1987년 3월~1991년 2월 : 한양대학교 전자공학과 (공학사)
1991년 3월~1993년 2월 : 한국과학기술원 전기 및 전자공학과 (공학석사)
1993년 3월~현재 : 한국과학기술원 전기 및 전자공학과 박사과정
1993년 1월~1996년 7월 : 시스템공학연구소 별정직 연구원
※주관심분야 : 청각모델링, 음성인식, 신호처리, 신경회로망 등

▲이 수 영(Soo-Young Lee) 1952년 10월 15일생
1975년 2월 : 서울대학교 공과대학 전자공학과(공학사)
1977년 2월 : 한국과학원 전기공학과 (공학석사)
1984년 5월 : Polytechnic Institute of New York, Electrophysics (공학박사)
1977년~1980년 : 대한엔지니어링 (주) 과장대리
1983년~1985년 : General Physics Corp. Staff Scientist/Senior Scientist
1986년~현재 : 한국과학기술원 전기 및 전자공학과 조교수/부교수/교수

▲길 이 만(Rhee M. Kil)
1985년 9월~1991년 12월 : University of Southern California, 전기공학과 컴퓨터공학 박사과정
1983년 9월~1985년 5월 : University of Southern California, 전기공학과 컴퓨터공학 석사과정
1975년 3월~1979년 2월 : 서울대학교 공과대학 전기공학과 학사과정
1994년 8월~현재 : 한국과학기술원 기초과학과정 조교수
1991년 11월~1994년 7월 : 한국전자통신연구소 기초기술연구부 선임연구원으로서 비선형 시계열예측 및 자료부호화 관련 과제책임자로 재직
1988년 1월~1991년 5월 : University of Southern California 전기공학과 실험, 연구, 강좌 조교
1979년 2월~1982년 10월 : 국방과학연구소 전자단 연구원
※주관심분야 : 시계열예측, 자료부호화, 비선형제어, 패턴인식, 학습이론, 병렬처리컴퓨터