技術解說

잡음 환경에서 음성 인식을 위한 신호 처리
Signal Processing for Speech Recognition in Noisy Environment

金 元 九 · 林 龍 勳 · 車 日 煥 · 尹 大 熙
(Weon Goo Kim, Yong Hoon Lim, Il Whan Cha, Dae Hee Youn)
(연세대학교 전자공학과)

要 約

본 논문에서는 잡음 환경에서 음성 인식 시스템의 성능을 개선할 수 있는 잡음제거 방식과 거리 측정 방법을 연구하고 백색 및 유색 잡음 환경에서 거리 측정 방법에 따른 음성 인식 시스템의 성능을 평가하였다. 잡음 제거 방법으로는 음성 인식 시스템의 전처리 과정으로서 사용될 수 있는 스펙트럼 차감법, 자기 상관 차감법, 적응 잡음 제거, 적응 빔 형성기가 있으며 거리 측정 방법으로는 Log Likelihood Ratio(d_{LLR}), 켑스트럼에 의한 거리 측정(d_{CEP}), 가중 켑스트럼 거리 측정(d_{WCEP}), 스펙트럼 기울기에 의한 거리 측정(d_{RPS}), 켑스트럼 투영 거리 측정 방법(d_{CP}, d_{BCP}, d_{WCP}, d_{BWCP})들이 있다. 백색 및 자동차 잡음 환경에서의 화자 종속 단독음 인식 실험 결과, 켑스트럼 계수와 높은 차수에 큰 가중을 두는 거리 측정 방법인 d_{RPS}, d_{WCEP}가 잡음에 강한 특성을 나타내었으며, 잡음이 존재할 때는 pre-emphasis를 하지 않은 경우가 높은 인식율을 얻을 수 있었다.

ABSTRACT

This paper studies noise subtraction methods and distance measures for speech recognition in a noisy environment, and investigates noise robustness of the distance measures applied to the problem of isolated word recognition in white Gaussian and colored noise (vehicle noise) environments. Noise subtraction methods which can be used as a pre-processor for the speech recognition system, such as the spectral subtraction method, autocorrelation subtraction method, adaptive noise cancellation and acoustic beamforming are studied, and distance measures such as Log Likelihood Ratio(d_{LLR}), cepstral distance measure(d_{CEP}), weighted cepstral distance measure(d_{WCEP}), spectral slope distance measure(d_{RPS}) and cepstral projection distance measure (d_{CP}, d_{BCP}, d_{WCP}, d_{BWCP}) are also investigated.

Testing of the distance measures for speaker-dependent isolated word recognition in a noisy environment indicate that d_{RPS} and d_{WCEP} which weigh higher order cepstral coefficients more heavily give considerable performance improvement over d_{CEP} and d_{LLR}. In addition, when no pre-emphasis is performed, the recognizer can maintain higher performance under high noise conditions.

1. 서 론

음성 인식 시스템의 실용화가 눈어남에 따라 주변 잡음에 대한 인식 시스템의 성능 저하가 문제시되고

있다. 그러한 이유는 잡음이 없거나 비교적 조용한 실험실 환경에서는 잘 작동하는 음성 인식 시스템의 성능이 입력에 잡음이 존재할 때는 급격히 떨어지기 때문이다. 실제로 잡음이 존재할때의 문제점은 잡음이 존재할 때 화자의 발음이 조용한 환경에서 발음한 것과 다르고 음성 인식 시스템의 입력으로 잡음과 음성이 동시에 들어가기 때문에 기준 패턴과 다른 형태를 갖는 다는 점이다. 또한 잡음은 음성 인식 시스템이 사용되는 상황에 따라 다양한 형태로 존재하기 때문에 특정한 잡음 환경만을 가정하여 음성 인식 시스템을 학습시킬 수도 없다. 따라서 주변 환경의 변화, 즉 잡음에 적응할 수 있는 음성 인식 시스템의 구현은 실용적인 음성 인식 시스템을 개발하기 위하여 고려해야 할 중요한 문제중의 하나로 연구되고 있다.

현재, 잡음 환경에서 음성 인식을 하기 위하여 두 가지 방향으로 연구가 진행되고 있다. 첫번째는 실험실 환경에서 성공을 거둔 음성 인식 시스템을 잡음 환경에 대하여 강하게 만드는 방법으로서 인식 시스템의 입력 전단에서 음성에 섞인 잡음을 제거하는 잡음 제거 시스템을 부착하는 것이다[23]. 이러한 방법은 기존의 음성 인식 시스템의 구조를 변화시키지 않는 장점이 있다. 두번째 방법은 음성 인식 시스템을 설계할 때부터 입력 음성에 첨가되는 잡음에 강하게 제작하는 것이다[2]. 이 방법은 일반적으로 사용되는 방법과는 다르게 잡음의 영향을 적게받는 특징 벡터나 거리 측정 방법을 사용하는 인식 시스템을 사용하는 것이다. 이러한 방법에 사용되는 특징 벡터나 거리 측정 방법은 조용한 환경에서도 좋은 성능을 나타내면서 잡음의 영향을 적게 받아야 하는 두가지 조건을 만족해야 한다.

본 논문에서는 첫번째 방법인 잡음 제거 방식중에서 스펙트럼 차감법(spectral subtraction method)[23], 자기 상관 차감법(autocorrelation subtraction method)[11], 적응 잡음 제거법(adaptive noise cancellation)[19]과 음향 빔 형성법(acoustic beamforming)[20]에 대하여 기술하고 두번째 방법으로서 거리 측정(distance measure)의 계산이 간단한 LPC 계열의 여러가지 거리 측정 방법들과 그것들의 잡음에 대한 영향을 비교한다. 거리 측정 방법으로는 음성 인식 시스템에서 많이 사용되고 있는 Log Likelihood Ratio(LLR)[3], 켑스트럼 거리 측정(cepstral distance measure켑[4]과 잡음의 영향을 적게 받는 거리 측정 방법으로 제안된 가중 켑스트럼 거리 측정(weighted cepstral distance measure)[5], 스펙트럼 기울기에 의한 거리 측정(spectral slope distance measure)[6], 켑스트럼 투영 거리 측정(cepstral projection distance measure)[7] 방법들을 사용한다. 또한 잡음 환경에서 화자 종속 단독음 인식 실험에서는 컴퓨터에서 발생한 백색 잡음과 실제 자동차 운전중에 발생한 잡음을 녹음해서 사용하였으며, 여러 거리 측정 함수들의 잡음에 대한 강인성을 평가하기 위하여 잡음이 섞이지 않은 기준패턴을 가지고 인식 실험을 수행하였다.

Ⅱ. 잡음 제거 방식

음성인식 시스템이 잡음환경에서 사용될 때, 전처리과정의 목적은 음성에 포함된 잡음을 제거하여 잡음으로 인한 음성인식 시스템의 성능저하를 최소로 하는 것이다. 이러한 목적의 잡음제거 방법은 음성인식 시스템에 추가되는 계산량이 적어야 하며, 시스템이 사용되는 잡음환경의 특성에 준하여 설계되어야 한다. 예를들어 잡음의 형태를 미리 알고있거나 소음의 통계적 특성이 서서히 변하는 환경에서는 스펙트럼 차감법(spectral subtraction) 부류의 잡음제거 방법을 사용할 수 있으며, 잡음의 특성이 시변인 환경에서는 적응잡음제거(adaptive noise cancelling) 방법이 적절하다. 화자 방향으로 지향성빔을 형성하여 주변소음을 제거하는 마이크로폰 어레이(microphone array) 기법도 전처리 과정으로서 응용되고 있다.

2.1 스펙트럼 차감법(spectral sutraction method)

스펙트럼 차감법[13]은 주변 잡음에 의해 손상된 음성 스펙트럼에서 잡음 스펙트럼의 크기 성분만을 제거하는 방법이다. 이는 주변잡음이 음성에 산술적으로 더해진다는 가정과, 음성을 인지하는 청각의 특성은 음성의 주파수 성분별 위상정보 보다는 크기 정보에 더 많이 영향을 받는다는 연구결과[13]에 기초한다. 대부분의 음성 인식 시스템에서 사용하는 음성의 특징 파라메타(feature parameter)로서 스펙트럼의 크

기정보를 사용하므로 스펙트럼 차감법은 음성인식 시스템에 효과적인 잡음제거 방법이라 할 수 있다.

스펙트럼 차감법은 다음과 같은 조건을 만족하는 잡음 환경에서 사용될 수 있다.

- 배경소음의 스펙트럼 형태를 미리 알고 있거나, 잡음의 스펙트럼을 추정하기에 충분한 묵음구간(약 300ms)이 주어져야 한다.
- 배경소음은 최소한 부분적으로 stationary한 특성을 갖아야 하며, 통계적 특성이 서서히 변화하는 환경에서는 음성이 존재하는 구간과 잡음만이 존재하는 구간을 검출할 수 있는 방법이 필요하다.

잡음신호 n(k)가 음성신호 s(k)에 더해졌을 때, 손상된 음성신호 X(k)는 다음과 같이 나타낼 수 있다고 가정한다.

$$x(k)=s(k)+n(k) \tag{1}$$

이 신호에 윈도우(window)를 취하여 단시간 푸리어 변환(short time fourier transform)을 하면,

$$X(\omega)=S(\omega)+N(\omega) \tag{2}$$

음성신호의 크기 추정치는 잡음 스펙트럼의 평균을 사용하여 다음과 같이 구해진다.

$$|\hat{S}(\omega)|=|X(\omega)|-\mu(\omega) \tag{3}$$

여기에서 $\mu(\omega)=E[|N(\omega)|]$을 나타내지만, 실제로는 음성이 없는 잡음구간에서 수 프레임(frame)의 데이타로 부터 구한 샘플 평균(sample mean) $\hat{\mu}(\omega)$로 대치하여 사용한다.

$$\hat{\mu}(\omega)=\frac{1}{M}\sum_{i=1}^{M}|N_i(\omega)| \tag{4}$$

여기에서, $N_i(\omega)$는 i번째 윈도우 데이터의 푸리어 변환을 나타내며, 평균을 취하는 윈도우의 수가 많을수록 $\hat{\mu}(\omega)$는 $\mu(\omega)$에 접근한다.

잡음이 제거된 음성신호의 시간축 파형을 얻기 위해서는 식(3)으로 표현된 음성의 크기정보 이외에 음성의 위상정보가 필요하다. 그러나 주어진 상황에서 위상정보를 얻는 것은 어려우며, 청각적인 측면에서 이 음질향상이나 음성인식 과정에서 위상정보가 중요한 변수가 아니라는 사실을 고려하여 음성의 위상정보 $\theta_s(\omega)$를 손상된 음성의 위상정보 $\theta_x(\omega)$로 대치하여 사용한다. 결과적으로 개선된 음성의 주파수 영역에서의 표현은 다음과 같다.

$$\hat{S}(\omega)=[|X(\omega)|-\hat{\mu}(\omega)]\exp(-j\,\theta_x(\omega)) \tag{4}$$

이상에서 살펴본 스펙트럼 차감법을 일반화하면 그림 1과 같이 나타낼 수 있다. 여기에서 a는 잡음제거 정도에 가변성을 주는 파라메터이며, a=1인 경우는 Boll[23]에 의해 제안되어 연구되었고, a=2인 경우는 파워스펙트럼 차감법(power spectral subtraction) 또는 자기상관 차감법[14]이라 한다.

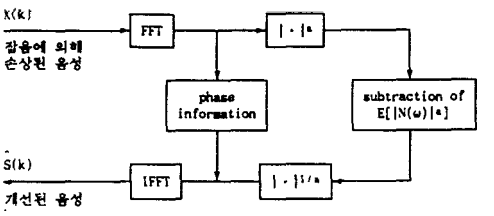

그림 1. 스펙트럼 차감법의 블럭도
Fig. 1. Block diagram of spectral subtraction

2.2 자기상관 차감법(autocorrelation subtraction method)

음성인식을 위한 특징 파라메터를 추출하는 과정에서 많은 경우에 음성신호의 자기상관함수를 구하는 과정이 선행된다. 자기상관 차감법은 잡음이 섞인 자기상관 함수로 부터 잡음성분을 제거하는 방법으로, 특징 파라메터 추출하는 중간과정에서 적은 계산량으로 간단히 구성하여 잡음에 강인한 파라메터 추출을 가능하게 한다.

식(1)로 가정된 신호의 단구간 자기상관함수(short time autocorrelation)는 다음과 같이 나타낼 수 있다.

$$r_{xx}(\tau)=r_{ss}(\tau)+r_{sn}(\tau)+r_{ns}(\tau)+r_{nn}(\tau) \tag{6}$$

여기에서,

$r_{xx}(\tau)=\sum_{k=1}^{N} x(k)x(k+\tau)$: 잡음에 손상된 음성의 단구간 자기상관함수

$r_{sn}(\tau)=\sum_{k=1}^{N} s(k)n(k+\tau)$: 음성과 잡음의 단구간 상호상관함수

$r_{nn}(\tau)=\sum_{k=1}^{N} n(k)n(k+\tau)$: 잡음의 단구간 상호상관함수

음성신호와 잡음 사이에 상관관계가 없다고 가정하면 $r_{sn}(\tau)\approx 0$, $r_{ns}(\tau)\approx 0$ 이므로, 식(6)은 다음과 같이 간단히 된다.

$$r_{xx}(\tau)=r_{ss}(\tau)+r_{nn}(\tau) \tag{7}$$

잡음이 섞인 신호 x(k)의 자기상관함수는 음성신호의 자기상관함수와 잡음의 자기상관함수의 합으로 생각할 수 있다. 식(7)로부터 음성신호의 자기상관함수 추정치는 다음과 같이 구해진다.

$$\hat{r}_{ss}(\tau)=r_{xx}(\tau)-E[r_{nn}(\tau)] \tag{8}$$

여기에서, $E[r_{nn}(\tau)]$는 잡음신호에 대한 단구간 자기상관함수의 ensemble 평균을 의미하지만, 실제로는 스펙트럼 차감법에서와 같이 음성이 없는 구간에서 측정한 샘플 평균을 사용한다.

$$\hat{r}_{ss}(\tau)=r_{xx}(\tau)-\hat{E}[r_{nn}(\tau)] \tag{9}$$

여기에서,

$$\hat{E}[r_{nn}(\tau)]=\frac{1}{M}\sum_{i=1}^{M} r^{i}_{nn}(\tau)$$

을 나타내고, $r^{i}_{nn}(\tau)$는 i번째 윈도우에서 구한 잡음의 단구간 자기상관함수이다.

실제상황에서 $\hat{E}[r_{nn}(\tau)]$ 값이 너무 크거나 또는 $r_{xx}(\tau)$ 값이 너무 작아서, $\hat{r}_{ss}(\tau)$로 구성되는 자기상관함수 행렬의 positive definite가 성립하지 않은 경우가 발생한다[11]. 이러한 경우 음성신호의 LPC 계수를 구할 수 없으므로 식(9)는 다음과 같이 변형되어 사용된다.

$$\hat{r}_{ss}(\tau)=r_{xx}(\tau)-\alpha\cdot\hat{E}[r_{nn}(\tau)],\quad 0\le\alpha\le 1 \tag{10}$$

WSS(Wide Sense Stationary) 환경에서 자기상관함수와 파워스펙트럼이 푸리어 변환쌍인 관계에 있으므로, 자기상관차감법은 그림 1에 나타낸 파워스펙트럼차감법(a=2인 경우)과 등가로 볼 수 있다.[13].

2.3 적응잡음제거(adaptive noise cancellation)[19]

그림 2는 적응디지탈필터(adaptive digital filter)를 사용하여 음성 신호 s(k)에 첨가된 잡음 n(k)를 제거하는 시스템의 블럭도이다. 주입력(primary input)신호 x(k)에 포함되어 있는 첨가잡음신호 n(k)는 잡음원신호 $\tilde{n}(k)$가 선형 시스템 W(z)를 통과한 출력신호이며, 원하는 신호 s(k)와 상관관계가 없다고 가정한다.

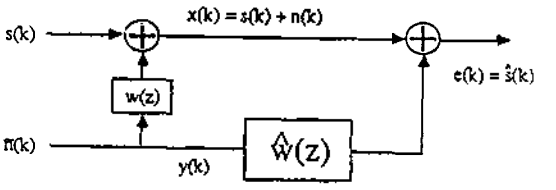

그림 2. 적응잡음제거기의 블럭도
Fig. 2. Block diagram of adaptive noise canceller

잡음 제거방법은 잡음경로시스템 W(z)를 추정하고, 추정된 필터 $\hat{W}(z)$에 참고입력 $y(k)\sim\tilde{n}(k)$을 통과시켜 얻은 잡음추정신호 $\hat{n}(k)$를 주입력 신호 x(k)로부터 빼내므로써 얻을 수 있다. 이때 오차신호 e(k)는 신호대잡음비(SNR : Signal-to-Noise Ratio)가 향상된 음성추정신호 $\hat{s}(k)$가 된다.

$$e(k)=s(k)=x(k)-\sum_{m=1}^{M} w_m\, y(k-m) \tag{10}$$

$$=x(k)-W^{T}Y(k) \tag{11}$$

여기서,

$$W=[w_0\ w_1\ w_2 \cdots w_M]^T$$
$$Y(k)=[y(k)\ y(k-1)\ y(k-2) \cdots y(k-M)]^T$$

이며, "T"는 전치행열을 나타낸다.

입력신호들이 stationary인 경우 최적필터 계수는 평균자승오차 $E[e^2(k)]$를 최소화하는 Wiener 필터 계수[18]로 주어진다.

$$W_{opt}=R_{yy}^{-1}P_{xy} \quad (12)$$

여기에서,

$R_{yy}=E[Y(k)Y^T(k)]$: 참고입력신호의 자기상관행열
$P_{xy}=E[x(k)Y(k)]$: 주입력신호와 참고입력신호벡터 간의 상호상관행열

실제로 잡음제거 시스템이 신호의 통계적 특성을 모르거나 시변(time varying) 환경에서 운용되는 경우 식(12)로 주어지는 최적필터의 계산과 실제 사용에 따르는 어려움을 극복하기 위해 적응디지탈필터가 연구되었다. 대표적인 적응 알고리듬으로는 steepest descent 방법[15]을 사용하는 LMS(Least Mean Square) 알고리듬[19]과 LS(Least Squares) 최적화조건을 만족하는 RLS(Recursive Least Squares) 알고리듬[16] 등이 있다.

그림 3에 자동차안에서 적응디지탈필터를 사용한 잡음제거 시스템의 예를 나타내었다. 적응필터의 주입력신호는 운전자의 음성과 자동차내의 엔진소음이며, 참고입력신호는 잡음발생원인 엔진소음을 본네트안에 있는 마이크로폰으로 받은 신호이다. 잡음이 제거된 개선된 신호는 적응필터의 오차신호 e(k)로서 음성인식시스템에 연결된다.

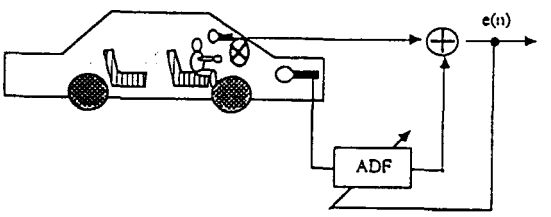

그림 3. 적응잡음제거 시스템의 예
Fig. 3. An example of adaptive noise cancellation system

3.4 음향 빔형성법

원하는 신호와 잡음의 방향이 다른 환경 또는 화자의 위치가 고정되어 배경 잡음에 대해 특정방향의 신호대잡음비가 높은 잡음환경에서는 공간상에서 신호의 방향정보를 이용하는 잡음제거방법을 사용하여 전처리과정에 응용할 수 있다. 그 대표적인 예로서 지향성 마이크로폰을 사용하는 방법이 있으나, 만족할 만큼의 예리한 지향특성을 얻기가 어렵다.

신호의 방향정보를 이용하는 다른 방법은 그림 4와 같이 마이크로폰 어레이에 수신된 신호를 시간지연시킨 후 더하여(delay-and-sum) 출력신호를 얻는 방법으로, 시간지연은 어레이를 특정방향으로 회전시킨 효과를 냄으로써 수신신호에 대해 지향성을 형성한다. 이를 고전적인 빔형성기(conventional beamformer)라 하며, 지향성 패턴에서 주엽(main lobe)의 각 ψ는 다음과 같이 주어진다[15].

$$\psi=\sin^{-1}\frac{c\delta}{d}\text{[radian]} \quad (13)$$

여기에서, δ는 시간지연[sec]을 나타내고, c는 음속[m/sec], d는 마이크로폰 사이의 간격[m]을 나타낸다.

고전적 빔형성기를 사용하여 음성신호방향으로 주엽이 형성되도록 설계함으로써 음성신호와 신호대잡음비를 높일 수 있지만, 광대역 음성신호에 대해 주파수 분해능이 떨어지며, 빔 형태가 어레이의 기학적 구조에 의해 결정되므로 외부환경이 변하면 성능이 저하되는 단점이 있다. 이러한 단점을 해결하기 위해

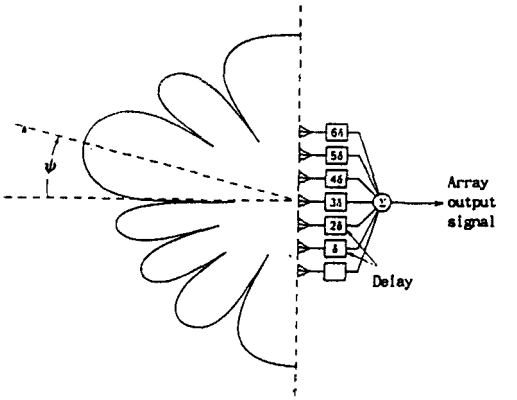

그림 4. 고전적 빔 형성기
Fig. 4. Conventional Beamformer

원하는 신호와 잡음원에 대한 선행지식이 없어도 제거하는 동시에 원하는 신호를 개선시키는 적응빔형성기(adaptive beamformer)에 대한 연구가 진행되고 있다[20][21][23].

빔형성법을 실제 환경에 적응하기 위해서는 3차원 공간상에서 음성신호 방향으로 지향성빔이 형성되어야 하므로 마이크로폰 어레이는 평면상에 배치되어야 한다.

Ⅲ. 거리 측정 방법

3.1 Log Likelihood Ratio

Itakura satio는 Itakura-Satio distortion measure로 불리우는 maximum likelihood distortion measure를 단구간 스펙트럼(short time spectrum)에 최초로 사용하였고[3] 다음과 같이 정의된다.

$$d_{IS}(S,f)=\int_{-\pi}^{\pi}\left\{\frac{S(w)}{f(w)}+\ln\frac{f(w)}{S(w)}-1\right\}\frac{dw}{2\pi} \tag{14}$$

여기서 S(w)는 음성 신호의 단구간 스펙트럼 밀도함수(short time spectrum density function or periodogram)이고 f(w)는 p차 전극 모델의 스펙트럼 밀도함수로서 다음과 같다.

$$f(w)=\frac{\sigma^2}{\left|1+\sum_{k=1}^{p}a_k e^{jw}\right|^2}=\frac{\sigma^2}{|A(e^{jw})|^2} \tag{15}$$

또한 Itakura는 Itakura-Satio distortion measure를 음성 인식에 사용하기 위하여 음성의 크기에 따른 영향을 최소화하는 gain optimized distortion measure를 제안하였고[3] 두 전극 모델 f(w)와 f′(w) 사이의 거리는 다음과 같이 정의 된다.

$$d_{LLR}(f,f')=\min_{\beta\geq 0} d_{IS}(f(w),\beta f'(w))=\ln\frac{a'^{T}Ra'}{a^{T}Ra} \tag{16}$$

위 식을 Itakura distortion measure[3]또는 Log Likelihood Ratio(LLR)라고 한다.

3.2 켑스트럼에 의한 거리 측정

두 스펙트럼 모델 f(w)와 f′(w)에 대해서 rms 대수 스펙트럼 거리(rms log spectral distance) d_{CEP}를 정의하면[4]

$$d_{CEP}=\int_{-\pi}^{\pi}\{\ln(f(w)/f'(w))\}^2(dw/2\pi) \tag{17}$$

이고 Parseval 정리를 d_{CEP}에 적용하여 p차 켑스트럼 거리 측정(cepstral distance measure)로 근사화 시키면 다음과 같다.

$$d_{CEP}=\sum_{k=1}^{p}(c_k-c_k')^2 \tag{18}$$

3.3 가중 켑스트럼에 의한 거리 측정

음성 분석 과정에서 발생되는 결점과 켑스트럼 계수의 가변성을 감소시키기 위하여 식(19a)와 같은 가중 함수가 제안 되었다[10]. 이러한 함수는 가변성이 큰 낮은 차수와 높은 차수의 켑스트럼 계수에 작은 가중을 두는 특징을 갖는다. 또한 위와 같이 켑스트럼 계수의 각 차수에 가중을 두는 거리 측정 방법으로서 켑스트럼 계수의 통계적 분포에 따라서 가중 함수를 결정하는 방법이 제안되었다[5]. 이 방법은 기준 패턴의 켑스트럼 계수의 분산의 역으로 가중을 두는 것으로서 분산이 정규화 된 유클리드안(Euclidian) 거리 측정 방법으로 식(19b)와 같다.

$$w(k)=1+h\sin(\pi k/L) \tag{19a}$$

$$w(k)=\frac{\sigma_1^2}{\sigma_k^2} \tag{19b}$$

여기서 h와 L은 실수이고 k=1,2,…L 이고 σ_k^2은 k번째 켑스트럼 계수의 분산이다.

이와 같은 가중 함수를 사용한 거리 측정 함수는 식 (20)과 같다.

$$d_{CEP}=\sum_{k=1}^{p}[w(k)(c_k-c_k')]^2 \tag{20}$$

3.4 스펙트럼 기울기에 의한 거리 측정

스펙트럼 기울기에 의한 거리 측정은 지각에 의하여 음성간의 거리(perceived phonetic distance)를

측정하는 연구에서 Klatt에 의하여 제안되었고[6] 이러한 개념은 선형 예측 분석에 따른 계산상의 이점이 있는 전극 모델 스펙트럼에 직접 적용되었다. 그 거리 측정 함수는 다음과 같다[6].

$$d_{RPS}=\int_{-\pi}^{\pi}\{-\frac{\partial}{\partial w}\log(f(w))-\frac{\partial}{\partial w}\log(f'(w))\}^2\frac{dw}{2\pi} \quad (21)$$

식 (21)은 켑스트럼 계수를 이용해서 다음과 같이 근사화 시킬 수 있다.

$$d_{RPS}=\sum_{k=1}^{p}[k(c_k-c_k')]^2 \quad (22)$$

식 (22)에서 index k와 k번째 차수의 켑스트럼 계수 c_k와 곱 $k\cdot c_k$를 index weighted 켑스트럼 계수 또는 root power sum(RPS)라 한다.

3.5 켑스트럼 투영거리의 측정

D.Mansour와 B.H.Juang등은 LPC 켑스트럼 계수를 특징 벡터로 사용하여 주위 환경의 변화에 영향을 덜 받는 거리 측정 방법을 제안하였다[7]. 이들은 실험을 통하여 백색 잡음이 섞였을 때의 켑스트럼 계수는 다음과 같은 특징이 있음을 파악하였다.

1) 켑스트럼 계수가 norm이 줄어든다.
2) norm이 작은 켑스트럼 계수가 norm이 큰 것보다 더 많은 영향을 받는다.
3) 잡음에 의한 켑스트럼 계수의 변화는 계수의 norm보다는 방향(orientation or direction)이 영향을 적게 받는다.
4) 백색 잡음이 존재할 때 낮은 차수보다 높은 차수의 켑스트럼 계수의 변화가 적다.

이러한 실험을 바탕으로 제안한 거리 측정 방법은 다음과 같다.

$$d_{CP}=|(c_t/|c_t|)-(c_r/|c_r|)|^2 \quad (23)$$
$$=2(1-\cos\beta)$$

$$d_{WCP}=|c_t|(1-\cos\beta)=|c_t|-c_tc_r/|c_r| \quad (24)$$

여기서 c_t와 c_r은 각각 시험 패턴과 기준 패턴의 켑스트럼 계수로 구성된 행벡터이고 $\cos\beta=c_tc_r/|c_t||c_r|$으로서 비교되는 두 벡터들 간의 방향 코사인(directional cosine)이고 $|c_t|$는 시험 벡터의 norm을 의미한다. d_{CP}는 정규화된 두 벡터들간의 차이이고 d_{WCP}는 시험 패턴의 norm으로 가중으로 둔 거리 측정방법이다.

식 (24)에서 $|c_t|$는 상수이므로 인식 결과에 영향을 미치지 못하고 $\hat{c}_r$을 정규화된 (normalized) 기준 켑스트럼 벡터라 하면 다음과 같이 쓸 수 있다.

$$d_{WCP}=-c_t\cdot\hat{c} \quad (25)$$

Ⅳ. 실험 및 결과

위에서 기술한 거리 측정 방법에 대하여 조용한 연구실에서 녹음한 음성에 컴퓨터에서 발생시킨 백색 잡음과 주행중인 자동차에서 녹음한 잡음을 여러가지 레벨로 섞어 화자 종속 단독음 인식 실험을 수행하였다[12].

1. 음성 인식 시스템

녹음기에 녹음된 음성은 4.5kHz의 차단 주파수를 갖는 저역 통과 필터를 통과한 후 16비트, 10kHz로 A/D 변환하였다. 이렇게 샘플링된 신호는 끝점 검출(endpoint detection)을 거친 후 전달 함수 $H(z)=1-0.95z^{-1}$인 디지탈 필터를 통과한 후에 20ms의 크기를 갖는 해밍(hamming) 윈도우를 사용하여 10ms씩 이동하면서 차수 p=14의 선형 예측 계수와 LPC 켑스트럼 계수를 구한다. 이때 잡음이 섞인 경우에는 pre-emphasis 과정을 생략한 특징 벡터(feature vector)도 구한다. 시험 패턴과 기준 패턴의 비교는 DTW(Dynamic Time Warping)을 사용하고 결정 법칙은 NN(Nearest Neighbor) 법칙을 사용하였다. 사용된 음성 인식 시스템은 그림 5과 같다.

2. 데이타 베이스

음성 인식에 사용된 단어는 10개의 단독 숫자음(0-9)과 '공', '예', '걸어', '시작', '다음'을 포함하는 총 15개이다. 기준 패턴은 조용한 연구실에서 녹음된 20-30대 남성 화자 7명이 단독음을 10회씩 발음한 것을 MKM(Modified K-Means)[11] 방식으로 집단화하여 사용하였다. 시험 패턴은 동일한 화자들

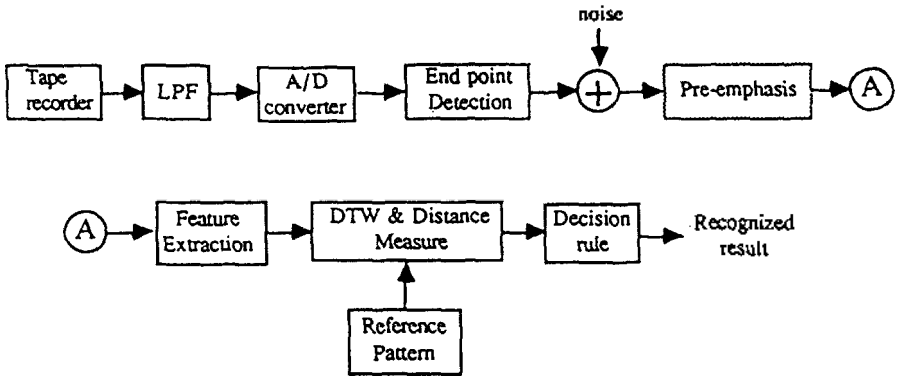

그림 5. 음성 인식 시스템
Fig. 5. Speech recognition system

이 한덩 경과 후 10번씩 발음한 총 1050개의 단독음에 백색 잡음과 자동차 잡음을 SNR 20,15,10,5,0 dB가 되도록 섞어 음성 패턴으로 구성하였다.

3. 거리 측정 방법에 따른 인식율

가중 켑스트럼에 의한 거리 측정 방법은 사용되는 가중 함수가 $w(k)=1+20\sin(\pi k/35)$일때 d_{WCEP}라고 정의하고 dv_{WCEP}의 가중 함수 $w(k)$는 k번째 켑스트럼 계수의 분산의 역수를 취한 것으로 기준 패턴을 사용하여 구한 가중 함수는 그림 6과 같다.

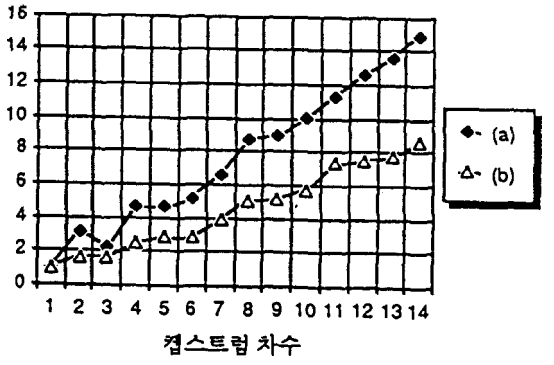

그림 6. dv_{WCEP}의 가중 함수
(a) pre-emphasis를 하지 않은 경우
(b) pre emphasis를 한 경우
Fig. 6. Weighting function of dv_{WCEP}
(a) no pre-emphasis
(b) pre-emphasis

백색 잡음이 SNR 20,15,10,5,0 dB로 첨가된 음성에 대하여 여러가지 거리 측정 방법에 대한 단독음 인식 실험을 수행하였다. Ⅲ장에서 기술한 거리 측정 방법들을 사용하고 pre-emphasis를 한 경우와 하지 않은 경우에 대하여 그림 7,8,9에 나타내었다.

그림 7는 5가지 거리 측정 방법인 d_{LLR}, d_{CEP}, d_{RPS}, d_{WCEP}, dv_{WCEP}에 대하여 실험한 결과이다. 인식 결과에서 잡음이 첨가되지 않은 음성의 인식율은 비슷하지만 잡음이 첨가되면 각각의 방법에 따른 차이가 매우 커지는 것을 알 수 있다. 특히 일반적인 인식 시스템에서 많이 사용되는 d_{LLR}과 d_{CEP}의 경우에는 잡음이 없을 때 높은 인식율을 나타내지만 잡음이 첨가되면서 가장 급격히 감소되는 것을 알수 있다. 따라서 이러한 방법을 사용하는 시스템의 성능은 잡음에 따라 크게 영향을 받게 되는 것이다. 이에 비하여 d_{RPS}, d_{WCEP}, dv_{WCEP}는 잡음이 첨가되어도 인식율의 감소가 상당히 둔화되어 d_{LLR}에 비하여 40%이상, d_{CEP}에 비하여 20~30%이상 인식율이 향상되는 것을 알 수 있다. 이러한 것은 SNR 측면에서 약 15~30dB의 향상을 나타내는 것이다. 이러한 결과는 d_{RPS}, d_{WCEP}, dv_{WCEP} 모두가 켑스트럼 계수의 높은 차수에 큰 가중을 두는 것을 고려할 때, 높은 차수의 켑스트럼 계수가 잡음의 영향을 적게 받는다는 것을 의미한다.

그림 8와 그림 9는 pre-emphasis를 하지 않은 입력 패턴을 대상으로 위와 같은 인식 시험을 한 결과이다. 실험은 d_{LLR}, d_{CEP}, d_{RPS}, d_{WCEP}, dv_{WCEP}와 켑스트럼 투영 거리 측정 방법인 d_{CP}, d_{WCP}, d_{BCP}, d_{BWCP}에 대하여 수행하였다. 여기서 d_{BCP}, d_{BWCP}는 d_{CP}와

d_{WCEP}에 $w(k)=1+20\sin(\pi k/35)$로 가중을 둔 켑스트럼 계수를 사용한 켑스트럼 투영 거리 측정 방법이다. 그림 7와 그림 8를 비교하면, 잡음이 존재할 때 pre emphasis를 하지 않은 경우가 5~10%정도 높은 인식율 나타내었다. 이러한 이유는 음성 신호의 에너지가 대부분 저주파 영역에 존재하기 때문에 백색잡음이 존재할 때 고주파 영역을 강조시키는 pre-emphasis는 잡음의 영향을 더욱 크게 만들기 때문이다. 그림 9에서, 켑스트럼 투영 방법은 잡음이 존재하지 않을 때는 다른 방법보다 5%가량 인식율이 떨어지나 잡음이 증가함에 따라서 감소되는 인식율의 변화가 완만하다. 특히 잡음이 매우 큰 0 dB에서는 d_{VWCEP}보다 8%정도 우수한 성능을 나타내었다.

그림 10는 시속 60km/h로 주행중인 자동차에서 녹음한 잡음을 SNR 20,15,10,5,0 dB로 첨가한 음성에 대하여 d_{LLR}, d_{CEP}, d_{RPS}, d_{WCEP}, d_{VWCEP}, d_{BCP}와 6가지 거리 측정 방법에 따른 단독음 인식 실험을 수행하였다. 자동차 잡음의 경우에도 높은 차수에 많은 가중을 두는 d_{RPS}, d_{WCEP}, d_{VWCEP}가 높은 인식율을 보였다. 전체적인 인식율은 자동차 잡음의 에너지가 저주파 영역에 집중되어 있기 때문에 백색 잡음 환경에 비하여 높다.

[dB] Method	Clean	20	15	10	5	0
LLR	99.33	48.57	34.86	24.95	15.52	10.57
CEP	99.52	67.44	51.43	39.14	28.19	18.10
RPS	96.86	92.29	88.00	78.76	60.95	37.71
WCEP	99.24	92.19	88.95	76.19	55.52	37.43
VWCEP	98.95	92.19	97.17	78.67	62.48	34.29

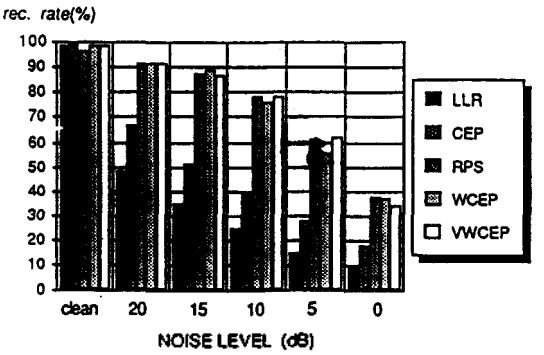

그림 7. 백색 잡음이 섞인 음성에 대하여 pre-emphasis를 한 경우, 거리 측정 방법과 SNR에 따른 인식율

Fig. 7. Recognition rate as a function of input SNR for the distance measures in white noise (with pre-emphasis)

[dB] Method	Clean	20	15	10	5	0
LLR	99.81	61.33	34.67	24.38	15.43	11.52
CEP	99.81	71.14	53.44	38.67	28.95	19.81
RPS	99.52	95.81	91.05	82.95	65.81	46.10
WCEP	99.43	95.52	92.38	82.76	64.95	46.00
VWCEP	99.05	95.33	92.48	87.14	78.29	59.81

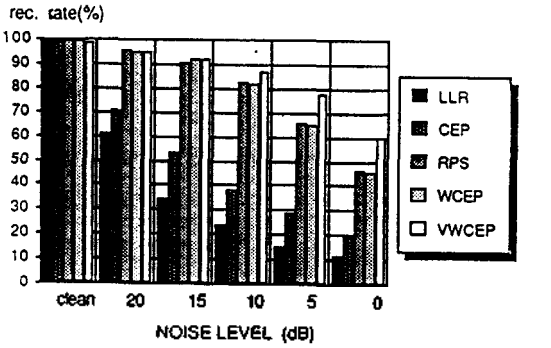

그림 8. 백색 잡음이 섞인 음성에 대하여 pre-emphasis를 하지 않은 경우, 거리 측정 방법과 SNR에 따른 인식율

Fig. 8. Recognition rate as a function of input SNR for the distance measures in white noise (with no pre-emphasis)

[dB] Method	Clean	20	15	10	5	0
CP	98.48	89.24	82.48	73.62	64.57	54.38
WCP	94.38	76.48	70.38	62.48	56.86	51.43
BCP	95.14	89.43	86.67	80.48	75.24	66.57
BWCP	94.57	85.90	82.19	78.29	73.52	67.90

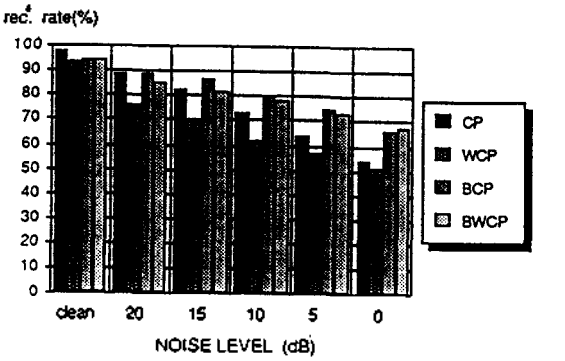

그림 9. 백색 잡음이 섞인 음성에 대하여 pre-emphasis를 하지 않은 경우, 켑스트럼 투영 방법과 SNR에 따른 인식율

Fig. 9. Recognition rate as a function of input SNR for the cepstral projection measures in white noise (with no pre-emphasis)

[dB] Method	Clean	20	15	10	5	0
LLR	99.81	98.29	94.19	85.90	73.81	55.62
CEP	99.81	99.05	97.62	94.86	88.38	80.10
RPS	99.52	99.24	98.76	98.19	96.38	93.52
WCEP	99.43	99.71	99.33	98.48	97.05	94.00
VWCEP	99.05	98.95	98.86	98.19	96.29	93.52
BCP	95.14	93.71	93.71	92.19	90.19	87.33

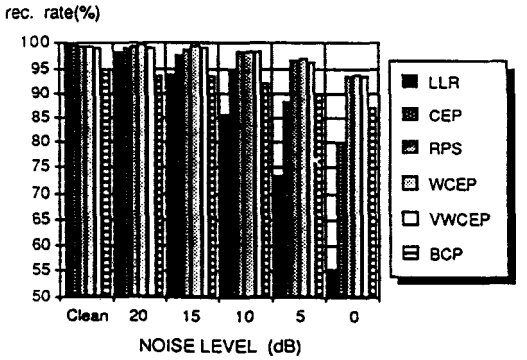

그림 10. 자동차 잡음이 섞인 음성에 대하여 pre-emphasis를 하지 않은 경우, 거리 측정 방법과 SNR에 따른 인식율

Fig. 10. Recognition rate as a function of input SNR for the distance measures in white noise (with no pre-emphasis)

V. 결 론

음성 인식 시스템의 실용화를 위해서는 잡음에 대한 대책이 반드시 필요하다. 이러한 이유는 잡음이 없거나 비교적 조용한 실험실 환경에서는 잘 동작하는 음성인식 시스템의 성능이 입력에 잡음이 존재할 때는 급격히 떨어지기 때문이다. 이러한 문제점을 해결하는 방법으로서 스펙트럼 차감법, 자기 상관 차감법, 적응 잡음제거, 음향 빔 형성법등과 같은 잡음 제거 시스템을 음성 인식 시스템의 전단에 사용하거나, 잡음에 강한 거리 측정 방법이나 특징 벡터를 사용하는 음성 인식 시스템을 설계하는 것이다. 또한 잡음에 의한 영향뿐만 아니라 잡음 환경에서 발음한 화자의 발음이 조용한 환경에서 발음한 것과 다를 때 발생하는 문제점도 고려해야 할 중요한 문제이다.

잡음 환경에서 거리 측정 방법에 따른 성능 분석 결과에서, 기존에 많이 사용되는 Log Likelihood Ratio나 켑스트럼 거리 측정 방법은 잡음 환경에서는 급격한 성능의 저하를 나타냈지만 켑스트럼 계수의 높은 차수에 큰 가중을 두는 가중 켑스트럼 거리 측정 방법이나 켑스트럼 투영 방법은 잡음에 강한 특징을 나타내었다.

현재 외국에서는 음성 인식 기술이 부분적으로 실용화 단계까지 발전하여 잡음에 대한 문제도 중요한 연구로서 진행되고 있으나, 아직 국내에서는 음성 인식에 대한 연구가 충분히 이루어지지 못하였기 때문에 잡음에 대한 문제도 중요시 되고 있지 못하다. 그러나 음성 인식 시스템의 잡음에 대한 대책은 음성 인식 시스템의 설계와 병행하여 이루어져야만 보다 잡음에 강한 음성 인식 시스템을 제작할 수 있다. 또한 음성 인식 시스템의 실시간 처리 문제와 더불어 잡음 제거 시스템의 실시간 구현 문제도 중요하다.

참 고 문 헌

1. S.Kay, "Noise Conpensation for Autoregressive Spectral Estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No.3, pp.292-303, June, 1980.
2. Y.Ephraim, J.G.Wilpon, and L.R.Rabiner, "A Linear Predictive Front-end Processor for Speech Recognition in Noisy Environment," *Proc. ICASSP-87*, pp.1324-1327, Apr. 1987.
3. F.Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-23, pp. 67-72, Feb. 1975.
4. A.H.Gray, Jr, J.D. Markel, "Distance Measures for Speech Processing," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.ASSP-24, pp.380-391, Oct. 1976.

5. Y.Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *Proc. ICASSP 86* pp.765-768, Apr. 1986.
6. B.A.Hanson, H Wakita "Spectral Slope Distance Measure with Linear Prediction Analysis for Word Recognition in Noise," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, No.7, pp.968-973, July. 1989.
7. D.Mansour, B.H.Juand, "A Family of Distance Measure Based upon Projection Operation for Robust Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-37, No.11, pp. 1659-1671, Nov. 1989.
8. D.H.Klatt, "Prediction of Perceived Phonetic Distance from Critical Band Spectra," *Proc. ICASSP-82*, pp.1278-1281, May. 1982.
9. C.Myers, L.R.Rabiner, A.E.Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithm for Isolated Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 6, pp.623-635, Dec. 1980.
10. B.H.Juang, L.R.Rabiner, J.G.Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans Acoust., Speech, Signal Processing*, Vol. ASSP-35, No.7, pp.947-954, July. 1987.
11. J.G.Wilpon, L.R.Rabiner, "A Modified K-Means Clustering Algorithm for Use in Isolated Word Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-33, No.3, June. 1985.
12. 김락용, 김원구, 임용훈, 차일환, 윤대희, "백색 및 잡음 환경하에서의 단독음 인식," 대한전자공학회논문지 제28권 B편 제 6호, pp.24-31, 1991년 6월.
13. J.S.Lim and A.V.Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. of The IEEE*, Vol.67, No.12, pp.1586-1604, Dec. 1979.
14. J.S.Lim, "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise," *IEEE Trans on Acoust., Speech and Signal Processing*, Vol. ASSP-26, No.5, pp. 471-472, Oct 1985
15. B.Widrow and S.D.Stern, *Adaptive Signal Processing*, Prentice-Hall, Inc., 1985.
16. S.Hakin, *Adaptive Filter Theory*, Prentice-Hall, Inc., 1986.
17. J.S.Lim, *Speech Enhancement*, Prentice-Hall, Inc., 1982.
18. N.Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, New York : Technology Press and Wiley 1949.
19. B.Widrow, et al., "Adaptive Noise Cancelling Principles and applications," *Proc. IEEE*, Vol.63, No.12, pp.1692 1716, Dec. 1975.
20. Y.Kaneda and J.Ohga, "Adaptive Microphone Array for Noise Peduction," *IEEE Trans. on Acoust., Speech, and Signal Processing*, Vol. ASSP-34, No.6, Dec. 1986.
21. M.M.Sondhi and G.W.Elko, "Adaptive Optimization of Microphone Arrays Under a Nonlinear Constraint," *Proc. ICASSP-86*, Tokyo, pp.981-984, 1986.
22. O.L.Frost Ⅲ, "An Algorithm for Linearly Constrainted Adaptive Array Processing," *Proc. IEEE*, Vol.60, No.8, pp.926-935, Aug. 1972.
23. S.F.Boll, "Supression of Acoustic Noise in Speech Using Spectral subtraction," *IEEE Trans. on Acoust., Speech, and Signal Processing*, Vol. ASSP-27, No.2, April. 1979.

본 연구는 한국과학재단의 연구비 지원에 의하여 이루어진 것임.

▲金 元 九 (Won Goo [illegible])

196[illegible]년 11월 3일 생.
1987년 2월 : 연세대학교 전자공학과 졸업
1989년 8월 : 연세대학교 전자공학과 석사학위 취득
1989년 9월~현재 : 연세대학교 전자공학과 박사과정 재학중
*주 관심분야는 디지탈 신호처리 및 음성 신호처리 등임.

▲林 龍 勳 (Yong Hoon LIM)

1963년 6월 26일생
1989년 2월 : 연세대학교 전자공학과 졸업
1991년 2월 : 연세대학교 전자공학과 석사학위취득
1991년 3월~현재 : 연세대학교 전자공학과 박사과정 재학중
*주 관심 분야는 디지탈 신호 처리 및 음성 신호 처리 등임.

▲尹 大 熙 (Dae Hee YOUN)
1991년 제10권 3호 참조

▲車 日 煥 (Il Whan CHA)
1991년 제10권 3호 참조