한국음향학회지 제26권 제1호 pp. 16~24 (2007)

UMP 테스트에 근거한 새로운 통계적 음성검출기

A New Statistical Voice Activity Detector Based on UMP Test

장 근 원*, 김 동 국*, 장 준 혁**
(Keun Won Jang*, Dong Kook Kim*, Joon-Hyuk Chang**)

*전남대학교 전자컴퓨터 공학부, **인하대학교 전자전기 공학부
(접수일자: 2006년 10월 14일; 수정일자: 2007년 1월 2일 채택일자: 2007년 1월 5일)

음성검출기는 이동 통신이나 음성신호처리 등에 매우 중요한 기법으로 사용된다. 일반적인 음성검출방식은 통계적인 모델을 기반으로 하여 likelihood ratio test (LRT)를 하게 된다. 그리고 이 값을 임계값과 비교하여 음성인지 아닌지 판단하게 된다. 본 논문에서는 가우시안 (Gaussian) 분포를 기반으로 하고 uniformly most powerful (UMP) 테스트를 이용하여 새로운 음성검출기법을 제안한다. 새로운 음성검출기법의 결정규칙은 기존 LRT에 기반하여 UMP 테스트를 통해 식을 유도하였다. UMP 테스트를 이용하면, 입력음성에 대한 절대값의 확률 분포를 Rayleigh 분포 형태로 얻을 수 있으며, 이 분포에 따라 최종적으로 음성검출을 하게 된다. 이 새로운 방식의 음성검출기는 기존의 방식에서 필요한 a priori signal-to-noise ratio (SNR) 값을 구하지 않고도 음성 유무를 판단할 수 있다는 장점이 있다. 실제로 다양한 음성검출에 대한 성능 평가결과, 제안된 기법이 기존 방식에 비해 우수한 성능을 나타내었다.

핵심용어: 음성검출기, 가우시안 분포, UMP 테스트, Rayleigh 분포, Likelihood ratio test, A priori SNR

투고분야: 음성처리 분야 (2.3)

Voice activity detectors (VADs) are important in wireless communication and speech signal processing. In the conventional VAD methods, an expression for the likelihood ratio test (LRT) based on statistical models is derived. Then, speech or noise is decided by comparing the value of the expression with a threshold. We propose a new method with the modified decision rule based on the Gaussian distribution and the uniformly most power (UMP) test. This method requires the distribution of the absolute value of the incoming speech signal. Then we can obtain the final decision through the relation between the Rayleigh distributions. This VAD method can detect speech without a priori signal-to-noise ratio (SNR) which is required in the conventional VAD algorithms. Additionally, in the various VAD performance tests, the proposed VAD method is shown to be more effective than the traditional scheme.

Key words: Voice activity detection, Gaussian distribution, UMP test, Rayleigh distribution, Likelihood ratio test, A priori SNR

ASK subject classification: Speech Signal Processing (2.3)

I. 서 론

음성검출기 (voice activity detector, VAD)는 현재 이동 통신 및 음성신호처리 기술 등에 있어서 매우 중요한 기술로써 사용되고 있다. 음성검출기는 잡음이 섞여 있는 음성신호에서 음성이 존재하는 부분과 잡음만 존재하는 부분을 판별하는 기술로, 현재까지 활발히 연구가 이루어지고 있다 [1]-[4], [8], [9], [11], [13].

책임저자: 김 동 국 (dkim@chonnam.ac.kr)
500-757 광주광역시 북구 용봉동 300번지 전남대학교
전자컴퓨터 공학부
(전화: 062-530-1794; 팩스: 062-530-1759)

최근에 연구되고 있는 여러 가지 통계적 기반 음성검출 기술들의 특징을 살펴보면 다음과 같다. 우선, 입력된 신호들의 전력스펙트럼 (power spectrum) 분석을 한 후, 이를 가우시안 (Gaussian) 분포나 라플라시안 (Laplacian) 분포, 혹은 감마 (gamma) 분포 형태의 확률밀도함수 (probability density function, pdf)을 갖는다고 가정을 한다 [9]. 이러한 분포는 이산 푸리에 변환 (discrete Fourier transform, DFT) 영역에서의 통계모델로 이용하게 된다. 음성의 DFT 영역에서 통계적 분포를 분석하여 잡음 혹은, 잡음에 오염된 음성에 대한 likelihood ratio test (LRT)식을 세우며, 이를 통하여

최종 결정 규칙을 도출하여 음성을 검출하게 된다 [1], [2], [4], [8], [9].

이러한 통계모델에 근거한 음성검출 방법은 [2]논문에서 처음 시도되었는데, 음성을 검출하는 설계기술은 Ephraim과 Malah의 연구에서 시작되었다 [5]. 그 이후에 나온 여러 통계적 모델을 바탕으로 한 음성검출기에서도 Ephraim과 Malah의 연구 기술이 기본 토대로 이용되었다. 통계모델을 기반으로 하는 음성검출기에서 결정 규칙을 도출할 때, 중요한 파라미터로 선행 신호대 잡음비 (a priori signal-to-noise ratio ; SNR)과 사후 신호대 잡음비 (a posteriori SNR) 이 필요하게 된다. [5]논문에서 제안한 기술은, 바로 이 두 개의 신호대 잡음비 변수를 추정하는데 DD (decision-directed) 추정을 통하여 하는 것이다. DD 방법은 기존의 maximum likelihood (ML) 방식을 이용하여 추정하는 것보다 더 효율적인 추정치를 제공하며, 뮤지컬 (musical) 잡음도 줄이는 효과까지 가져왔다 [2]. 음성검출과 음성향상을 위하여 더 나은 기술들이 제안되었는데, 대표적인 것으로 soft decision 기반으로 전력스펙트럼을 추정하는 방식을 들 수 있다 [3]. 이 방식은 기존의 음성검출방식과 비교하여 뛰어난 성능 향상을 가져올 수 있었다. 그리고 잡음을 추정하는 데에 있어, 스무딩 방식과 최소 통계치 방식을 적용하는 방법도 제안되었다 [10].

기존의 여러 음성검출 방식에서의 핵심은 a priori SNR 변수추정에 있다. 이 변수의 정확한 추정이 효과적인 음성검출 성능 향상을 가져오기 때문이다. 하지만, 본 논문에서는 a priori SNR 값을 구하지 않고서도 결정규칙 식을 구하는 새로운 방식을 제안하였다. 이는 LRT를 하는 데에 있어 새로운 결정 규칙을 세웠음을 뜻하는데, 이 식을 유도하기 위하여 uniformly most powerful (UMP) 테스트를 이용한다. UMP 테스트 방식은 어떠한 변수를 결정하는데 있어, 특정 변수 값이 항상 0보다 크다는 가정이 성립할 때, 통계치 추정식을 보다 간략히 하는 방식이다. 본 논문에서는 이 방식을 음성검출 결정규칙을 세우는 데에 적용하여 새로운 알고리즘을 제안한다.

본 논문에서는 통계적 음성검출 방식에 있어, DFT 계수 분포에 대한 특성으로 가장 널리 사용되는 가우시안 확률밀도함수를 기반으로 하였다. 그리고 결정 규칙을 세우는 데에 있어, LRT에 기반하여 UMP 테스트를 적용하는 새로운 접근을 시도하였다. 이러한 접근의 결과, 가우시안 분포기반의 가정에서 출발하여, UMP 테스트에 근거하여 음성의 크기에 대해 Rayleigh 분포형태의 결정 규칙을 얻게 되었다. 이러한 결정 규칙을 적용하는 데에 있어, 또 하나의 중요한 변수인 임계값을 추정하는 방식도 제안하였다. 임계값은 오탐율 확률 (false alarm probability)을 설정하여 이로부터 적응적으로 잡음에 따라 구할 수 있는 방법을 소개한다.

이러한 이론을 바탕으로 한 새로운 음성검출기를 실제 잡음이 섞인 음성신호에 적용하였고, 기존의 통계적 음성검출기와 비교하여 실험하였다. 실험 결과, 본 논문에서 제안한 UMP 테스트를 이용한 음성검출기 방식은 기존의 방식보다 성능 향상을 가져왔으며, A priori SNR 값을 추정하지 않아도 음성검출을 할 수 있음을 보여주었다.

본 논문의 II장에서는 통계적 음성검출기에 대해 소개하고, III장에서는 UMP테스트를 통한 새로운 음성검출기 방식에 대하여 논하였다. IV장에서는 기존의 방식과의 비교실험 결과를 보여주며, 마지막으로 V장에서 결론을 맺어 본 논문을 마친다.

II. 통계적 음성검출기

음성신호를 분석하기위해 음성신호는 비상관 잡음 (uncorrelated noise)이 섞여있다는 가정을 한다. 그로부터 각각의 프레임별 신호의 형태를 두 가지로 가정할 수 있는데, 그것은 다음과 같다.

H_0 : speech absent : $X(t) = N(t)$

H_1 : speech present : $X(t) = N(t) + S(t)$

여기서 $X(t) = [X_0(t), X_1(t), X_2(t), \cdots, X_{M-1}(t)]^T$,

$N(t) = [N_0(t), N_1(t), N_2(t), \cdots, N_{M-1}(t)]^T$, 그리고

$S(t) = [S_0(t), S_1(t), S_2(t), \cdots, S_{M-1}(t)]^T$는 각각 잡음에 오염된 음성신호, 잡음신호, 원래의 음성신호의 DFT 계수 벡터를 나타낸다. 그리고 M은 전체 주파수대역의 개수이고 T는 전치행렬을 나타낸다.

X_k 는 잡음에 오염된 신호의 k번째 bin의 DFT 계수를 뜻하는데, $X_k = X_{k(R)} + jX_{k(I)}$와 같이 실수부와 허수부의 합으로 표현할 수 있으며, 이는 각각 독립이라고 가정한다 [8].

실수부와 허수부의 분포가 가우시안분포 형태를 갖는다고 가정하고, 두 부분이 독립이기 때문에 두 부분을 곱하면 다음과 같은 X_k의 분포를 얻을 수 있다.

$$P_{\{X_{k(R)},X_{k(I)}\}}(u,v)=\frac{1}{\pi\lambda_{x,k}}exp(-\frac{u^2+v^2}{\lambda_{x,k}}) \quad (1)$$

$$P_{X_k}(z)=\frac{1}{\pi\lambda_{x,k}}exp(-\frac{|z|^2}{\lambda_{x,k}}),\ \ (z=u+jv) \quad (2)$$

여기에서 $\lambda_{x,k}$는 X_k의 분산 값을 뜻하고, u,v 는 각각 실수부와 허수부의 값을 뜻한다.

기존의 가우시안 분포를 가정한 VAD 방식은 H_0, H_1 상태에서의 X_k의 확률밀도함수를 통하여 LRT 식을 구하게 된다. 이러한 방식을 통하여 구한 식은 다음과 같다 [2].

$$\Lambda_k=\frac{p(X_k|H_1)}{p(X_k|H_0)}=\frac{\lambda_{n,k}}{\lambda_{s,k}+\lambda_{n,k}}exp\left\{\frac{\lambda_{s,k}|z|^2}{\lambda_{n,k}(\lambda_{s,k}+\lambda_{n,k})}\right\} \quad (3)$$

$$=\frac{1}{1+\xi_k}exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \quad (4)$$

여기서, $\lambda_{n,k}$는 잡음신호에 대한 분산이며, $\xi_k=\lambda_{s,k}/\lambda_{n,k}$ 이고, $\gamma_k=|z|^2/\lambda_{n,k}$ 으로 표현하며, 각각 a priori SNR과 a posteriori SNR을 뜻한다.

(4) 식을 보면, 음성검출을 하는 데에 필요한 파라미터는 a priori SNR과 a posteriori SNR이다. A priori SNR을 구하기 위해서는 DD 방식을 이용해서 구해야하며, a posteriori SNR은 입력된 신호에서 잡음과 전체 신호를 통해 구할 수 있다.

III. UMP 테스트를 이용한 통계적 음성검출기

3.1. UMP 테스트

제안된 알고리즘을 제시 하기에 앞서 UMP 테스트에 대하여 예를 들어 간단히 설명하겠다 [7].

$H_0: x[n]=w[n], \qquad n=0,1,...,N-1$

$H_1: x[n]=A+w[n], \qquad n=0,1,...,N-1$

의 두가지 가설상태에서 A의 값은 알지 못하나 $A>0$ 인 상태를 갖는다고 하자. 여기에서 $w[n]$은 분산이 σ^2인 백색가우시안잡음 (white Gaussian noise; WGN)이다. 이를 통하여 H_1의 결정식을 유도하면 다음과 같다.

$$\frac{p(x;A,H_1)}{p(x;H_0)}=\frac{\frac{1}{(2\pi\sigma^2)^{N/2}}exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]}{\frac{1}{(2\pi\sigma^2)^{N/2}}exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]}>\gamma \quad (5)$$

여기에 양변에 로그를 취하면

$$-\frac{1}{2\sigma^2}\left(-2A\sum_{n=0}^{N-1}x[n]+NA^2\right)>\ln\gamma \quad (6)$$

또는

$$A\sum_{n=0}^{N-1}x[n]>\sigma^2\ln\gamma+\frac{NA^2}{2} \quad (7)$$

앞에서 A의 값은 0보다 크기 때문에 양변을 A로 나눌 수 있다. 그리고 양변을 $1/N$으로 나누어주면,

$$\frac{1}{N}\sum_{n=0}^{N-1}x[n]>\frac{\sigma^2}{NA}\ln\gamma+\frac{A}{2}=\gamma' \quad (8)$$

와 같이 표현이 가능하다. 여기서 중요한 점은 A의 정확한 값을 알지 못하더라도 어떠한 상태인지를 결정할 수 있다는 것이다. 이런 테스트의 형태를 UMP 테스트라 부른다 [7]. 그러나 UMP 테스트는 A의 값이 $-\infty<A<\infty$ 인 상태에서는 적용될 수 없다. 즉, UMP 테스트는 특정한 A의 값이 0보다 큰 값만을 갖거나 혹은 0보다 작은 값만 갖는 가정인 one-sided 테스트에서만 적용할 수 있다 [7].

3.2. UMP 테스트를 이용한 음성검출기

본 절에서는 UMP 테스트를 이용하여 새로운 음성검출기법을 유도한다. 새로운 기법에서도 음성검출을 하는데 있어 출발은 기존의 방식과 같은 가정을 따르게 된다. 즉, 음성과 잡음의 분포는 가우시안 분포를 따른다고 가정한다.

II장에서 유도한 (3)식에서부터 다시 살펴보자.

$$\Lambda_k=\frac{p(X_k|H_1)}{p(X_k|H_0)}=\frac{\lambda_{n,k}}{\lambda_{s,k}+\lambda_{n,k}}exp\left\{\frac{\lambda_{s,k}|z|^2}{\lambda_{n,k}(\lambda_{s,k}+\lambda_{n,k})}\right\} \quad (9)$$

여기에서 우리는 입력된 음성신호의 처음 시작되는 묵음구간을 통해 잡음의 분산인 $\lambda_{n,k}$의 값을 구할 수 있다. 그리고 음성의 분산값인 $\lambda_{s,k}$의 값은 그 값이 항상 0보다 큰 값을 갖고 있음을 알고 있다. 즉, 우리는 $\lambda_{n,k}$의 값을 안다고 가정하고, 알려지지 않은 변수 $\lambda_{s,k}$는 $\lambda_{s,k}>0$ 상태를 갖는다 할 수 있다. 즉, 두 값을 통하여 UMP 테스트를 설계할 수 있게 된다.

위 (3)식을 UMP 테스트를 통하여 정리하면 다음과 같다.

$$\frac{\lambda_{n,k}}{\lambda_{s,k}+\lambda_{n,k}} exp\left\{\frac{\lambda_{s,k}|z|^2}{\lambda_{n,k}(\lambda_{s,k}+\lambda_{n,k})}\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_k \tag{10}$$

UMP 테스트를 이용하여 $\frac{\lambda_{n,k}}{\lambda_{s,k}+\lambda_{n,k}}$ 는 상수 처리하여 우변으로 넘길 수 있으며, 남은 식에 양변에 로그를 취하면,

$$\frac{\lambda_{s,k}|z|^2}{\lambda_{n,k}(\lambda_{s,k}+\lambda_{n,k})} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'_k \tag{11}$$

이다.

UMP 테스트를 토대로, 양변을 또다시 정리하면,

$$|z|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma''_k \tag{12}$$

위 식에서 $z = X_k$를 뜻하기 때문에 최종적으로 다음과 같이 정리될 수 있다.

$$|X_k|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma''_k \quad = \quad |X_k| \underset{H_0}{\overset{H_1}{\gtrless}} \sqrt{\gamma''_k} \tag{13}$$

마지막으로 이를 프레임 단위로 산술평균 (arithmetic mean)을 구하여, 최종 결정 식은 얻을 수 있다.

$$\frac{1}{M}\sum_{k=1}^{M}|X_k| \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{M}\sum_{k=1}^{M}\sqrt{\gamma''_k} \tag{14}$$

위 식에서 $|X_k|$의 값을 구하기 위하여, 본 논문에서는 입력된 잡음에 오염된 음성신호 X_k를 실수부와 허수부로 나누어 다루는 대신, 크기 분포를 기반으로 시작을 하였다. 즉, 음성을 $X_k = |X_k|e^{j\theta_{x,k}}$와 같이 크기 $|X_k|$와 위상 $\theta_{x,k}$로 표현하고, 이를 바탕으로 $|X_k|$의 분포를 식으로 구해보면, k번째 주파수 bin의 크기 $|X_k|$의 분포는, X_k가 가우시안 분포를 따르므로, 아래 (15)식과 같은 Rayleigh 분포를 갖게 됨을 알 수 있다 [6].

$$P_{|X_k|}(v) = \frac{2v}{\lambda_{x,k}} exp(-\frac{v^2}{\lambda_{x,k}})\, U(v) \tag{15}$$

여기서 $U(v)$는 단위계단함수 (unit-step function)이다.

또한, 각각의 주파수 bin에 대한 크기의 제곱인, $|X_k|^2$의 분포를 구하면, 지수분포형태를 갖는다 [6]. 이를 식으로 표현하면 다음과 같다.

$$P_{|X_k|^2}(v) = \frac{1}{\pi\lambda_{x,k}} exp(-\frac{v}{\lambda_{x,k}})\, U(v), \tag{16}$$

식 (15)를 이용하여 잡음만 존재하는 경우와 잡음과 음성이 섞여 있는 두가지 가설에 의해 식을 재정리하면 다음과 같다.

$$H_0 ; P_{|X_k|}(v) = \frac{2v}{\lambda_{n,k}} exp\left\{-\frac{v^2}{\lambda_{n,k}}\right\} U(v) \tag{17}$$

$$H_1 ; P_{|X_k|}(v) = \frac{2v}{\lambda_{s,k}+\lambda_{n,k}} exp\left\{-\frac{v^2}{\lambda_{s,k}+\lambda_{n,k}}\right\} U(v) \tag{18}$$

위 식 (17)와 식 (18)을 이용하여 입력 음성신호의 분포형태와 잡음만 존재하는 경우와 잡음에 오염된 음성의 분포를 나누어 분석할 수 있다. 즉, 위 두 식을 통하여 최종결정 식 (14)를 수행할 수 있다. 식 (14)에서 임계값 γ_k가 존재하는데, 이 값을 결정하는 것이 음성검출에 있어서 또 하나의 중요한 요소이다.

3.3. 임계값 결정

식 (14)에서 음성검출에 대한 최종 결정을 하기 위해서는 γ''_k값을 구해서 비교해야 한다. 이 임계값을 찾기 위해서는 오탐율 (false alarm, FA) 확률을 통하여 구한다 [7].

$$\begin{aligned} P_{FA} &= P_r\left\{|X_k| > \sqrt{\gamma''_k} \;:\; H_0\right\} \\ &= P_r\left\{\frac{|X_k|}{\sqrt{\lambda_{n.k}}} > \frac{\sqrt{\gamma''_k}}{\sqrt{\lambda_{n.k}}} \;:\; H_0\right\} \\ &= P_r\left\{\frac{|X_k|^2}{\lambda_{n.k}} > \frac{\gamma''_k}{\lambda_{n.k}} \;:\; H_0\right\} \quad = Q_{\chi_2^2}(\frac{\gamma''_k}{\lambda_{n.k}}) \end{aligned} \tag{19}$$

P_{FA} 값은 2-degree Chi-squared 분포 형태로 이루어지는데, 이러한 오탐율을 통해서 임계값을 얻을 수 있다.

임계값을 구하기 위하여 $\sqrt{\gamma''_k} = \gamma_k$ 라 두고 정리하면

$$P_{FA} = P_r\{|X_k| > \gamma_k : H_0\} = Q_{\chi_2^2}(\frac{\gamma_k^2}{\lambda_{n,k}}) = \exp(-\frac{\gamma_k^2}{2\lambda_{n,k}})$$

이고, 이식은 $\gamma_k^2 = -2\lambda_{n,k}\ln P_{FA}$와 같이 되며 우리가 찾고자 하는 임계값은 최종적으로 다음과 같이 나온다.

$$\gamma_k = \sqrt{-2\lambda_{n,k}\ln P_{FA}} \tag{20}$$

(20)식에서, 임계값을 결정하는데 중요한 점은, 임계값은 고정된 false alarm 확률 상태에서 오직 잡음의 분산에만 의존한다는 것이다.

IV. 실험 및 고찰

본 논문에서 제시한 UMP 테스트를 이용한 음성검출기의 성능을 테스트하기 위하여 여러 실험을 하였다. 실험방식은 음성신호에 Babble 잡음과 White 잡음, 두 가지의 잡음 신호를 섞어서 이를 제안한 음성검출기를 이용하여 테스트하였다. 그리고 각각의 잡음이 섞인 정도를 달리하였는데, 이를 각각 0dB, 5dB, 10dB, 15dB로 주어서 실험하였다. 본 논문의 음성검출기와 기존의 가우시안분포를 기반으로 하는 음성검출기와의 비교를 통하여 제안한 음성검출기의 성능을 객관적으로 알아보았다. 성능을 테스트하기 위하여 detection과 false-alarm 확률 (P_D, P_{FA})을 알아야 한다. P_D는 음성프레임 결과에 대하여 실제로 정확하게 음성이라고 판단한 확률을 뜻하며, P_{FA}는 비음성을 음성이라 잘못 판단한 확률을 뜻한다. P_D 와 P_{FA} 의 값을 계산하기 위하여 우리는 32초의 음성을 10 ms 단위로 수동 레이블링하여 기준으로 삼았다. 수동 표시된 실제 음성의 비율은 54.97% 이고, 이 중에 46.46%는 유성음 (voiced sounds) 이며 8.41%는 무성음 (unvoiced sounds) 이었다. 잡음에 오염된 음성신호를 만들기 위해, White, Babble 잡음을 NOISEX-92 [12] 잡음으로부터 SNR을 변화하면서 원래의 음성신호에 첨가하였다. 음성검출 테스트는 10 ms의 프레임에 대하여 수행하였다.

4.1. $|X_k|$의 분포

먼저, 본 논문에서 제시한 UMP 테스트를 통하여 구한 LRT식에서 $|X_k|$의 분포가 Rayleigh 분포를 따르는지를 확인하였다. 먼저, 주파수 대역 (M)의 개수를 16개로 설정하였으며, Rayleig 분포를 확인하기 위하여 Babble

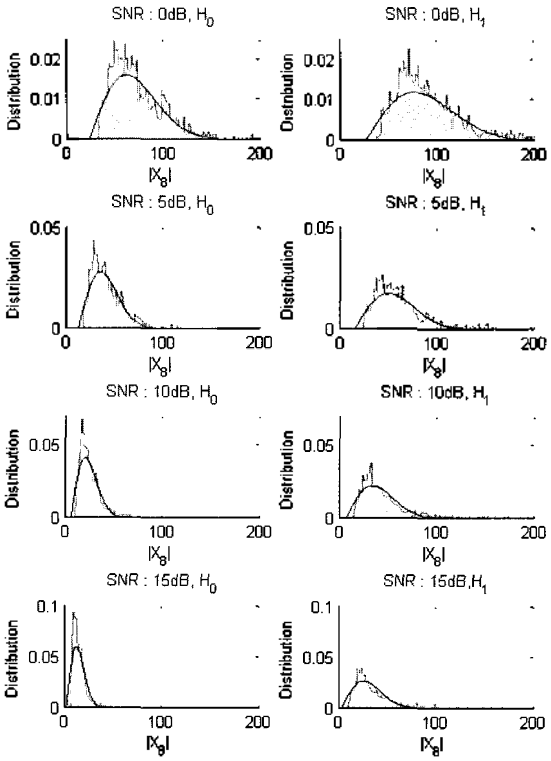

그림 1. Babble 잡음이 섞인 상태에서의 $|X_k|$의 분포도. SNR은 위에서부터 0dB, 5dB, 10dB, 15dB의 경우. 좌측: H_0, 우측: H_1

Fig. 1. Histogram of the magnitude in Babble noisy environment for $|X_k|$. SNR is 0dB, 5dB, 10dB, 15dB from top to bottom. Left: H_0, Right: H_1.

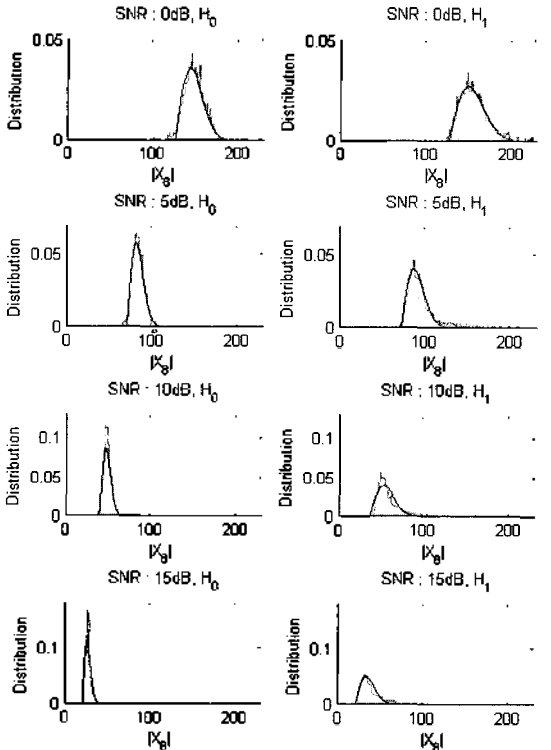

그림 2. White 잡음이 섞인 상태에서의 $|X_k|$의 분포도. SNR은 위에서부터 0dB, 5dB, 10dB, 15dB의 경우. 좌측: H_0, 우측: H_1

Fig. 2. Histogram of the magnitude in White noisy environment for $|X_k|$. SNR is 0dB, 5dB, 10dB, 15dB from top to bottom. Left: H_0, Right: H_1.

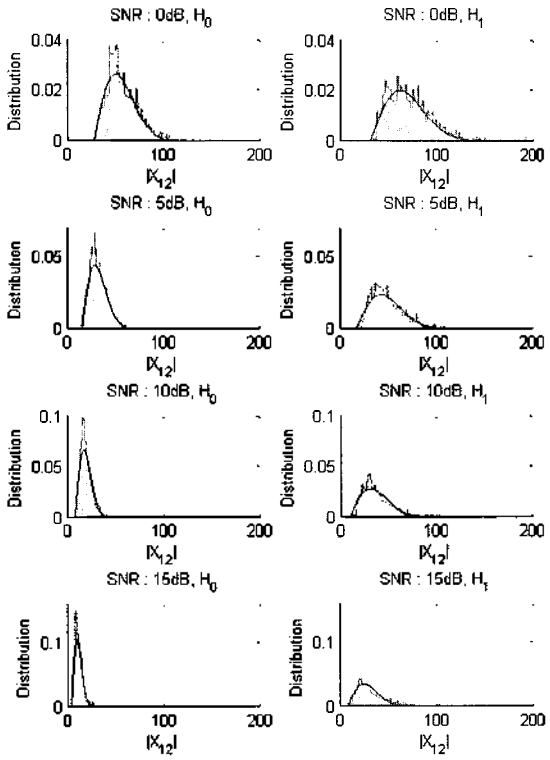

그림 3. Babble 잡음이 섞인 상태에서의 $|X_{12}|$의 분포도, SNR은 위에서부터 0dB, 5dB, 10dB, 15dB의 경우. 좌측: H_0, 우측: H_1

Fig. 3. Histogram of the magnitude in Babble noisy environment for $|X_{12}|$. SNR is 0dB, 5dB, 10dB, 15dB from top to bottom. Left: H_0, Right: H_1.

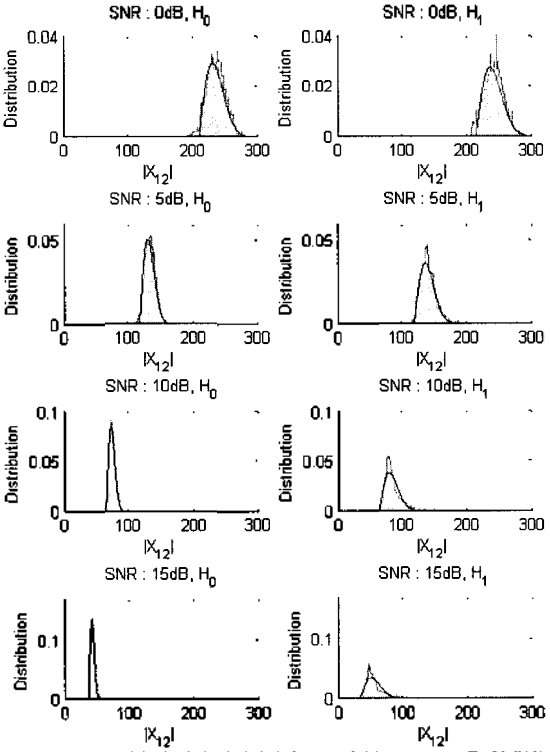

그림 4. White 잡음이 섞인 상태에서의 $|X_{12}|$의 분포도, SNR은 위에서부터 0dB, 5dB, 10dB, 15dB의 경우. 좌측: H_0, 우측: H_1

Fig. 4. Histogram of the magnitude in White noisy environment for $|X_{12}|$. SNR is 0dB, 5dB, 10dB, 15dB from top to bottom. Left: H_0, Right: H_1.

잡음과 White 잡음에 오염된 음성신호에서 $|X_k|$ 분포 값을 구하였다. 그리고 수동으로 확인한 음성과 비음성인 구간별로 나눈 후, 각각의 경우의 분포를 그려보았다. 16가지 경우 중 임의로 k=8인 경우와 k=12인 경우, 두 가지의 히스토그램을 제시하였다.

그림 1은 Babble 잡음이 섞인 음성신호에서의 분포를 보여준다. 그림1의 가장 위의 분포는 입력 신호대 잡음비가 0dB일 때의 분포이며, 차례대로 5dB, 10dB, 15dB의 분포를 나타내었다. 좌측의 분포는 입력신호가 잡음만이 존재하는 가설 H_0상태의 분포이며 우측의 분포는 H_1의 분포이다. y축의 값은 각각의 분포를 정규화 하여 나타내었다. 실선은 $|X_k|$를 토대로 그린 Rayleigh 분포이다.

그림 2는 White 잡음에 오염된 음성신호에서의 $|X_k|$의 분포를 그린 것이며, 실험방법은 그림1의 경우와 동일하다. 그리고 그림 3과 그림 4는 k=12일때의 경우의 $|X_k|$ 값의 히스토그램을 그린 것이고, 각각 Babble 잡음과 White 잡음에 오염되었을 경우를 그려보았다. 각각의 그림에서 x축은 $|X_k|$의 값을 나타낸 것이고, y축은 해당하는 값의 크기의 빈도를 나타낸 것이다. k=8인 경우와 12인 경우를 대표적으로 제시하였으나, 다른 구간에서도 제시한 히스토그램과 비슷한 분포를 갖음을 실험을 통하여 확인할 수 있었다.

앞에서 제시한 히스토그램들을 살펴보면, $|X_k|$의 분포가 레일리 분포 형태를 갖음을 확인할 수 있다. H_0의 분포는 상대적으로 적은 값을 많이 갖은 상태이며 H_1의 분포는 H_0분포에 비하여 큰 값으로 분포가 더 퍼져 있음을 알 수 있다. 이 두 개의 분포를 비교하여 우리가 제안한 음성검출을 할 수 있게 된다.

4.2. 음성검출

본 논문에서 제시한 방법을 통하여 음성검출을 하는 과정을 제시하고자 한다. 그림 5와 그림 6은 제안된 음성검출기의 실제 수행 과정을 나타낸 것이다. 첫 번째 그림은 실험에 사용한 잡음에 오염된 음성신호로 그림5는 Babble 잡음에서, 그리고 그림6은 White 잡음에서의 음성신호 파형을 나타낸다. 두 번째 그림은 $|X_k|$값과 임계값과의 관계를 그린 것이고 마지막 세 번째 그림은 두 값의 비교 후 음성의 유무를 판별한 것이다. 음성일 경우 1을 비음성일 경우 0의 값을 갖는다.

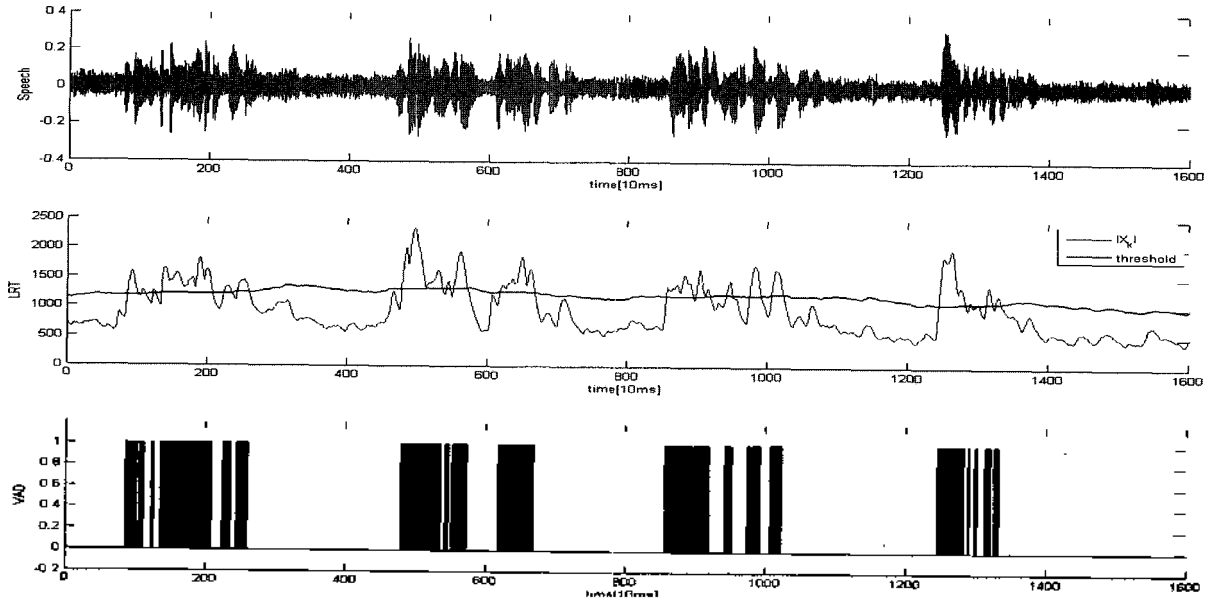

그림 5. Babble 잡음이 더해진 신호에서 입력 SNR이 5dB 일경우와 VAD 결과
첫 번째 그림은 입력된 음성신호를 나타내며, 두 번째는 $|X_k|$의 값과 임계값을 보여주고
마지막은 제안된 알고리즘에 의한 음성검출 결과임 (1:음성, 0:비음성)

Fig. 5. VAD result of the Babble noisy signal (input SNR 5dB)
first fig. is input signal, second fig. is the value of $|X_k|$ and threshold
last fig. is VAD result by the proposed algorithm (1: speech, 0: non speech).

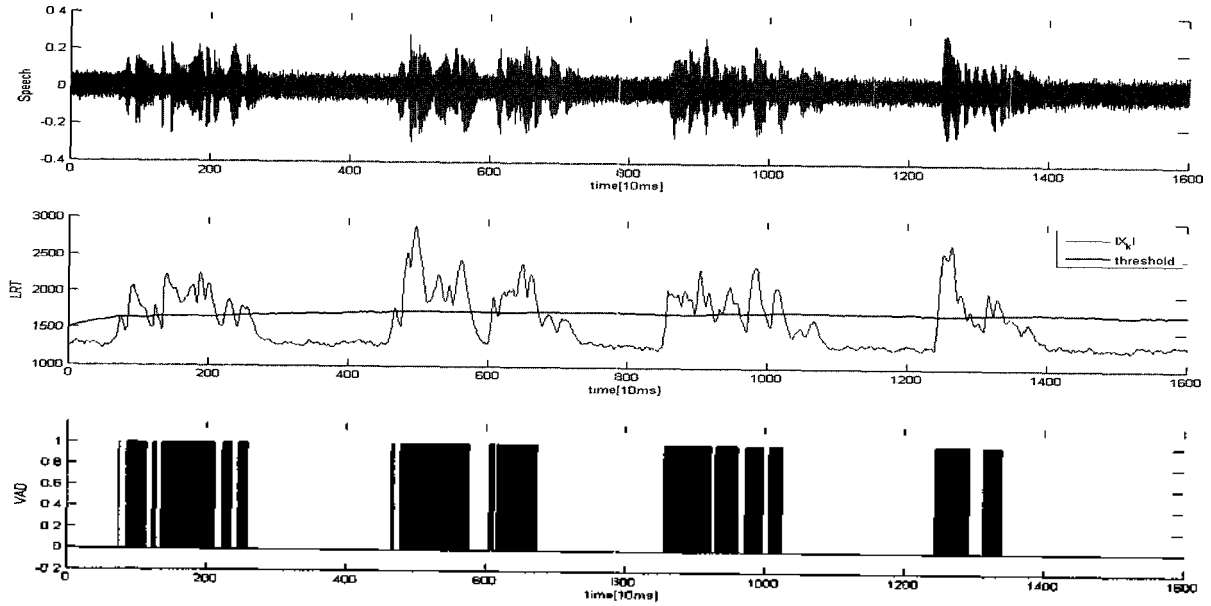

그림 6. White 잡음이 더해진 신호에서의 VAD 결과. (입력 SNR : 5dB)
첫 번째 그림은 입력된 음성신호를 나타내며, 두 번째는 $|X_k|$의 값과 임계값을 보여주고
마지막은 제안된 알고리즘에 의한 음성검출 결과임 (1:음성, 0:비음성)

Fig. 6. VAD result of the White noisy signal (input SNR 5dB)
first fig. is input signal, second fig. is the value of $|X_k|$ and threshold
last fig. is VAD result by the proposed algorithm (1: speech, 0: non speech).

4.3. ROC 곡선

본 논문에서 제안한 음성검출기의 성능을 알아보기 위해서는 $|X_k|$의 분포와 γ_k(임계값)을 통하여 P_D확률과 P_{FA}확률을 구할 수 있다. 이 두 개의 값의 관계를 표현한 것이 receiver operating characteristic (ROC) 곡선이다. 다음의 그림7는 Babble 잡음이 더해진 음성신호에서의 ROC 곡선을 그린 것이다. 각각 SNR이 0dB에서 15dB까지의 경우를 나타내주고 있다. ROC 곡선을 통하여 음성검출기의 성능을 알아볼 수가 있는데, 본 논문에서 제안한 방식의 음성검출기와 기존의 음성검출기와의

비교를 하기 위하여 가우시안 분포를 가정한 음성검출기의 ROC 곡선과 함께 제시하였다. 기존의 가우시안분포 기반의 음성검출기의 ROC 곡선은, a priori SNR 값을 추정함에 있어, 최적화된 a priori SNR 추정치를 토대로 제시하였다.

x축은 P_{FA}확률이고, y축은 P_D확률을 나타낸다. Babble 잡음이 섞인 경우의 ROC 곡선을 살펴보면 기존의 가우시안 방식에 비하여 성능이 우수함을 보여준다. 입력신호의 SNR 값이 작을 경우는 확연히 더 좋은 성능을 보임을 확인할 수 있다. SNR이 클 경우에는 P_{FA} 값이 낮을 때, 기존의 가우시안의 음성검출기에 비해 뒤처지는 경향이 있으나, 기존의 방식보다 보다 더 빠르게 P_D확률이 1에 근접함을 알 수 있다. 입력 SNR이 낮을수록 더 나은 성능 개선을 가져오는데, 이는 기존의 음성검출기에서 a priori SNR과 a posterioir SNR을 추정하는데 있어, 잡음과 음성의 분산 값에 의존을 많이 하게 된다. 이로 인하여 SNR이 낮을 경우, 두 파라미터 추정 및 음성검출이 영향을 많이 받으나, 제안한 방식은 이

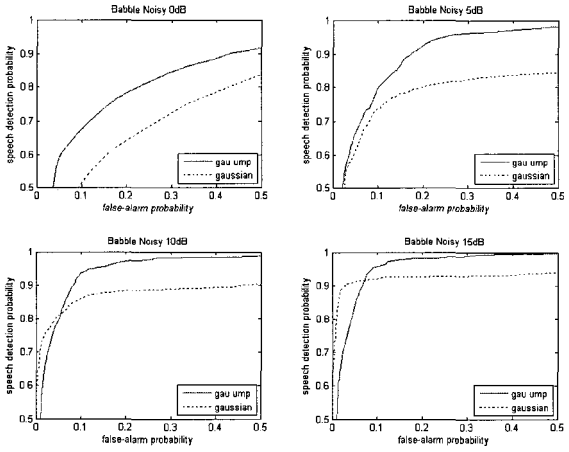

그림 7. Babble 잡음이 더해진 신호의 ROC 곡선
Fig. 7. Roc curve of the Babble noisy signal.

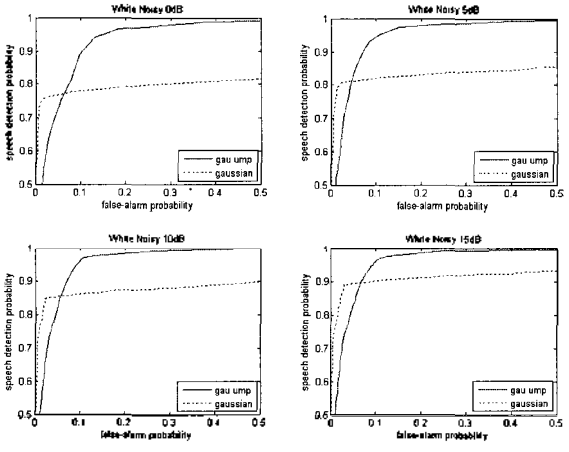

그림 8. White 잡음이 더해진 신호의 ROC 곡선
Fig. 8. Roc curve of the White noisy signal.

두 값의 추정이 없이도 음성검출을 하게 되므로 보다 나은 성능 개선을 얻을 수 있었다.

다음으로 그림8은 White 잡음이 더해진 경우의 ROC 곡선를 나타낸다. P_{FA}값이 매우 낮은 경우를 제외하면 전반적으로 보다 우수한 성능을 갖음을 알 수 있다.

V. 결 론

본 논문에서는 잡음이 더해진 음성신호가 입력되었을 때, 잡음과 음성을 판별하는 음성검출 방식에 대하여 새로운 기법을 제안하였다. 여러 음성검출기에서 사용하는 가우시안분포를 따른 다는 전제하에서, 입력신호의 절대값에 대한 분포를 분석하여 새로운 결정규칙을 도출해 내었다. 이러한 분포를 유도하는데 UMP 테스트 방식을 사용하였으며, 이 방식을 통하여 새로운 분포인 레일리 분포를 얻을 수 있었다. 입력신호의 절대값의 분포를 가지고 LRT 식에 적용하였으며, 이 값과 비교하기 위한 임계값은 P_{FA}확률 식으로부터 유도하였다.

이러한 설계방식에 따라 구현한 새로운 음성검출기는 기존의 가우시안 분포 기반의 음성검출기에 비하여 P_{FA} 값이 매우 낮은 경우 (약 0.05이하)에는 성능이 뒤처지는 경향이 있으나, 그 이외의 경우는 기존의 방식보다 더 우수한 성능을 나타내주었다. 입력 SNR을 다양하게 제시하였고, 잡음의 종류도 Babble과 White로 달리하여 실험하였다. ROC 커브를 그려본 결과, 입력 SNR이 낮을수록 더 높은 성능향상을 보여주었으며, 다양한 잡음 환경에서도 보다 나은 성능개선을 가져옴을 확인하였다.

이러한 성능의 개선뿐만 아니라, 본 논문에서 제시한 음성검출기는 또 하나의 장점을 가지고 있는데, 그것은 UMP 테스트의 도입으로 인한 계산상의 번거로움을 줄여주는 것이다. 이는 기존의 음성검출기에서 중요하게 사용되는 a priori SNR 값을 구하지 않고도 음성유무를 판별할 수 있음을 보여주었다. 여러 실험결과로부터, 우리는 잡음에 오염된 음성신호의 스펙트럼을 모델링하는데 있어, 가우시안의 경우에 비하여 UMP 테스트를 통한 음성검출기의 우수성을 알 수 있었다.

감사의 글

본 논문은 2004년도 전남대학교 학술연구비 지원에 의하여 연구되었습니다. (This study was financially

supported by research fund of Chonnam National University in 2004)

참 고 문 헌

1. A. Dvis, S. Nordholm and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," IEEE Trans. Audio, Speech, and Language Processing, 14 (2) 412-424, March 2006.
2. J. S. Sohn, N. S. Kim and W. Y. Sung, "A Statistical Model-Based Voice Activity Detection," IEEE Signal Process. Lett., 6 (1) 1-3, 1999.
3. N. S. Kim, and J. -H. Chang, "Spectral Enhancement Based on Global Soft Decision," IEEE Signal Process. Lett,, 7 (5) 108-110, 2000.
4. J. -H. Chang, J. W. Shin and N. S. Kim "Voice Activity Detector Employing Generalized Gaussian Distribution," IEEE Electronics Lett. 40 (24) 1561 - 1563, Nov. 2004.
5. Y. Ephraim and D. Malah, "Speech Enhancement Using A Minimum Mean-square Error Short-time Spectral Amplitude Estimator," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-32, 1109-1121, Dec. 1984.
6. P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, (John Wiley & Sons Inc., 2006)
7. S. M. Kay *Fundamentals of Statistical Signal Processing*, (Volume 2: Detection Theory, Prentice Hall, 1998)
8. J. -H. Chang and N. S. Kim, "Voice Activity Detection Based on Complex Laplacian Model," IEEE Electronics Lett., 39 (7) 632 - 634, April 2003.
9. J. -H. Chang, N. S. Kim and S. K. Mitra "Voice Activity Detection Based on Multiple Statistical Models" IEEE Trans. Signal Processing, 54 (6) 1965 - 1976, June 2006.
10. R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," IEEE Trans. Speech and Audio Processing 9 (5) 504 - 512, Jul. 2001.
11. Y. D. Cho and A. Kondoz "Analysis and Improvement of A Statistical Model-Based Voice Activity Detector," IEEE Signal Processing, Lett. 8 (10) 276-278, Oct. 2001.
12. A.Varga and H.J.M. Steeneken, "Assessment for Automatic Speech Recognition: II.NOISEX-92: A Database and An Experiment to Study The Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, 12 (3) 247-251, Jul.1993.
13. L. R. Rabiner and M. R. Sambur, "Voiced-Unvoiced-Silence Detection Using Itakura LPC Distance Measure," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. 2 323-326, May 1977.

저자 약력

• 장 근 원 (Keun Won Jang)

2006년 2월: 전남대학교 전자컴퓨터정보통신공학부 학사
2006년 3월~현재: 전남대학교 전자공학과 석사과정
※ 관심분야: 음성처리

• 김 동 국 (Dong Kook Kim)

1989년 2월: 전남대학교 전자공학과 학사
1991년 2월: 포항공과대학 전자전기공학과 석사
2003년 2월: 서울대학교 전기컴퓨터공학부 박사
1991년 2월~1993년 3월: 삼성전자 정보통신 연구원
1993년 3월~1999년 2월: 삼성종합기술원 전문연구원
2003년 4월~2004년 2월: 한국전자통신연구원 선임연구원
2004년 2월~현재 : 전남대학교 전자컴퓨터공학부 조교수
※ 관심분야: 음성처리, 음성인식, 패턴인식

• 장 준 혁 (Joon-Hyuk Chang)

1998년 2월: 경북대학교 전자공학과 학사
2000년 2월: 서울대학교 전기공학부 석사
2004년 2월: 서울대학교 전기컴퓨터공학부 박사
2000년 3월~2005년 4월: ㈜넷더스 연구소장
2004년 5월~2005년 4월: 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원
2005년 5월 ~2005년 8월: 한국과학기술연구원 (KIST) 연구원
2005년 9월~현재: 인하대학교 전자전기공학부 조교수
※ 관심분야: 음성 및 오디오 부호화, 음성향상, 음성인식, 적응신호처리