한국ITS학회논문지
제8권, 제3호 (2009년 6월)
pp. 64~70

잡음에 강인한 음성인식을 위한 Generalized Gamma 분포기반과 Spectral Gain Floor를 결합한 음성향상기법

Speech Estimators Based on Generalized Gamma Distribution and Spectral Gain Floor Applied to an Automatic Speech Recognition

김 형 국* 신 동** 이 진 호***
(Hyoung-Gook Kim) (Dong Shin) (Jin-Ho Lee)

요 약

본 논문은 잡음에 강인한 음성인식 성능을 획득하기 위해 generalized Gamma 분포기반의 음성향상 기법을 제안한다. 우수한 음성향상을 위해서 제안된 방식에서는 generalized Gamma 분포와 spectral gain floor를 이용한 음성추정 기법에 스펙트럼 최소잡음성분에 의한 회귀적인 평균 스펙트럼 값으로부터 유도되는 잡음추정을 결합하여 음질을 향상시켜 음성인식에 적용하였다. Spectral component, spectral amplitude 그리고 log spectral amplitude에 기반하여 제안된 음성향상 기법을 잡음환경에서의 음성인식에 적용하여 그 성능을 측정하였다.

Abstract

This paper presents a speech enhancement technique based on generalized Gamma distribution in order to obtain robust speech recognition performance. For robust speech enhancement, the noise estimation based on a spectral noise floor controled recursive averaging spectral values is applied to speech estimation under the generalized Gamma distribution and spectral gain floor. The proposed speech enhancement technique is based on spectral component, spectral amplitude, and log spectral amplitude. The performance of three different methods is measured by recognition accuracy of automatic speech recognition (ASR).

Key words: Speech estimators, generalized gamma distribution, noise estimation, speech recognition

Ⅰ. 서 론

최근에 자동차 환경에서 연구 및 개발되고 있는 지능형 교통 시스템에는 교통 상황, 도로 고조, 상황에 다른 항법 가이드 등 다양한 정보를 운전자가 쉽게 조작하기 위한 인간-기계 인터페이스로서 음성인식 기술을 적용하려는 연구가 활발히 진행되고 있다.

특히 자동차의 경우는 잡음의 크기가 크며 잡음의

† 이 논문은 2009년도 광운대학교 교내 학술연구비 지원에 의해 연구되었음.
* 주저자 : 광운대학교 전파공학과 부교수
** 공저자 : 광운대학교 전파공학과 석사과정
*** 공저자 : 광운대학교 전파공학과 학사과정
† 논문접수일 : 2009년 4월 14일
† 논문심사일 : 2009년 5월 20일
† 게재확정일 : 2009년 5월 21일

종류 및 특성이 다양할 뿐 아니라, 깨끗한 음성신호를 왜곡시켜 음성인식 성능을 저하시킨다. 이러한 자동차 환경에서 발생되는 잡음에 의해 왜곡되는 음성신호로부터 강인한 음성인식 성능을 획득하기 위해서는 전처리 영역에서의 효과적인 음성향상 알고리즘이 필요하다.

단일 마이크로폰를 사용하는 음성인식엔진에 적용되는 DFT(discrete Fourier transform)기반에서의 대부분의 음성향상 알고리즘은 잡음추정과 음성추정의 2가지 요소로 구성된다.

비정상(non-stationary) 잡음 환경에서 효과적인 잡음추정을 위해 DFT를 통해 획득된 파워스펙트럼의 회귀적인 평균값에서 추적된 최소값 통계(minimum statistics)를 이용해 Martin[1]과 Cohen[2]은 음성구간을 검출(voice activity detection)하여 잡음을 추정하는 방안을 제시하였다.

음성추정에 있어서는 Gaussian 분포기반의 log-spectral의 mean-square error[3]를 최소화할 수 있는 음성향상 이득기법이 널리 사용되어 오다가, 최근에 깨끗한 음성 스펙트럼의 확률적인 추정을 위한 DFT coefficients의 generalized Gamma분포[4]가 Gaussian 분포보다 음성향상에 있어서 우수한 성능을 보임이 입증되고 있다. 그러나 현재까지 generalized Gamma 분포기반의 minimum mean- square error(MMSE) amplitude 추정기법이 음성인식성능에 얼마만큼 효과적으로 기여하는지 보고된 바는 없다.

이에 따라, 본 논문에서는 DFT기반의 단일 마이크로폰에서의 음성향상을 위해 generalized Gamma distribution기반의 음성향상 이득(GGD)과 spectral gain floor (SGF)를 결합한 음성추정기법을 제안하고, 제안된 방식을 잡음환경에 왜곡된 음성데이터베이스를 이용해 음성인식 성능을 측정한다. 또한, 제안된 음성추정방식에 spectral component, spectral amplitude 그리고 log spectral amplitude를 적용한 세 가지 방식을 잡음환경에의 음성인식에 적용하여 음성인식 결과를 비교하고자 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 GGD와 SGF를 결합한 음성향상 알고리즘을 설명한다. 3장에서는 제시된 음성향상 알고리즘을 이용하여 음성인식 실험을 수행하고 실험 결과를 논의한다. 마지막으로 4장에서는 결론을 제시한다.

Ⅱ. 음성향상 알고리즘

본 논문에서 사용한 음성향상 알고리즘은 <그림 1>에 나타난 바와 같이 크게 잡음추정과 음성추정의 2가지 요소로 구성된다.

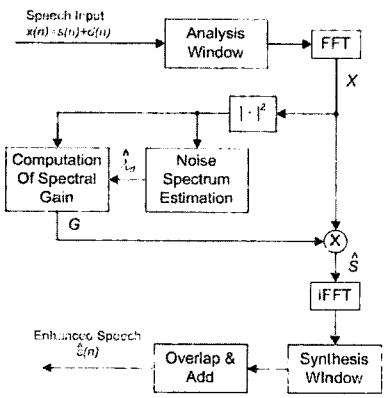
시간축 n상에서 원래의 음성신호 $s(n)$에 잡음신호 $d(n)$이 부과된 입력신호 $x(n)$을 윈도우 함수와 STFT (short-time Fourier transform)를 통해 주파수 축으로 변환하면 아래와 같이 표현된다.

$$X(k,l)=\sum_{n=0}^{N-1}x(n+lM)h(n)e^{-j(\frac{2\pi}{N})nk} \quad (1)$$

여기서 k는 k번째 스펙트럼 성분, l은 시간 프레임 지수, h는 N의 크기를 가지는 분석 윈도우이며, M은 프레임 스텝을 각각 나타낸다.

주어진 시간 프레임 지수 l에 대하여 오염된 입력음성 스펙트럼 $X(k,l)$는 $X(k,l)=S(k,l)+D(k,l)$로 나타낸다. $S(k,l)$와 $X(k,l)$는 크기성분 $A(k,l)$와 $R(k,l)$, 위상성분 $\phi(k,l)$과 $\theta(k,l)$를 통해 $S(k,l)=A(k,l)\exp(j\phi(k,l))$, $X(k,l)=R(k,l)\exp(j\theta(k,l))$로 표현된다.

입력 음성 스펙트럼으로부터 추정된 잡음성분에 GGD와 SGF를 결합한 음성추정 이득을 적용하여 잡음이 제거된 음성 스펙트럼 추정치 $\tilde{S}(k,l)$를 얻는다.



<그림 1> 음성향상 알고리즘의 구성도
<Fig. 1> Block diagram of speech enhancement

상세한 잡음추정과 음성추정에 대한 구체적인 설명은 다음과 같다.

1. 잡음추정

잡음추정은 <그림 2>와 같이 크게 6단계인 평균 스펙트럼 계산(short-term spectral averaging), 속지적 최소잡음성분 추적(local minimum tracking), 추적된 최소잡음성분을 이용한 음성구간 검출(VAD), 조건부 음성존재 확률 추정(conditional speech presence probability estimation), 스무딩 함수계산(smoothing parameter computation), 잡음성분 추정갱신(update noise spectrum estimation)으로 구성된다.

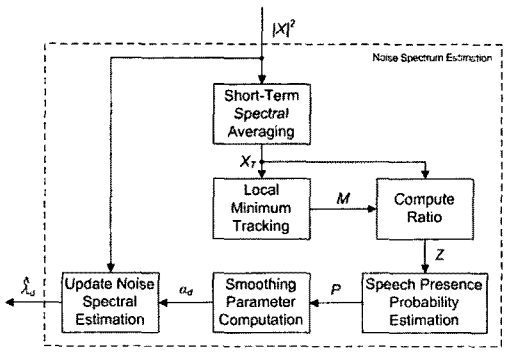
첫 번째 단계로, 입력 음성 스펙트럼에서 주파수 축과 시간 축에 대해 스무딩을 적용한 파워 성분의 평균 $X_T(k,l)$은 다음과 같은 일차 회귀 방정식에 의해 구해진다.

$$X_T(k,l)=\alpha_T X_T(k,l-1)+(1-\alpha_T)\left\{\frac{1}{2w+1}\sum_{i=-w}^{w}|X(k-i,l)|^2\right\} \quad (2)$$

여기서 w는 주파수 축에서의 스무딩을 위한 윈도우 함수이고, $\alpha_T(0<\alpha_T<1)$는 스무딩 파라미터이다.

두 번째 단계에서는 각 시간 축 프레임 지수 l로부터 구해진 평균 스펙트럼의 최소값을 프레임 수 $C(>l+1)$ 이내에서 비교함으로써 평균 스펙트럼 $X_T(k,l)$의 스펙트럼 최소잡음성분(local minimum) $M(k,l)$을 구한다.

$$M(k,l)=\min_{c=0\ldots C}\{M(k,l-c),X_T(k,l)\} \quad (3)$$



<그림 2> 잡음추정 알고리즘의 구성도
<Fig. 2> Block diagram of noise estimation

세 번째 단계는 입력된 음성신호의 평균 스펙트럼과 최소잡음성분 스펙트럼간의 비율을 이용하여 시간-주파수 성분 (k,l)에서의 음성존재구간과 비 음성존재구간을 구별하는 VAD $Z(k,l)$를 다음과 같이 계산한다.

$$Z(k,l)=\begin{cases}1 & \text{if} \quad X_T(k,l)/M(k,l)\geq\psi\\ 0 & otherwise\end{cases} \quad (4)$$

식 (4)에서 주파수 성분 k번째에서 $Z(k,l)=1$은 음성존재구간이며, $Z(k,l)=0$은 비 음성 존재구간이라 가정한다.

네 번째 단계로, 신호대 최소잡음비에 대한 문턱값에 영향을 받는 VAD를 기반으로 음성존재 확률 $p(k,l)$를 추정하고, 추정된 $p(k,l)$를 이용하여 잡음추정을 위한 최적 스무딩 함수 $a_d(k,l)$를 다음과 같이 고려한다.

$$p(k,l)=\begin{cases}\alpha_p+(1-\alpha_p)p(k,l-1) & \text{if} \quad Z(k,l)=1\\ (1-\alpha_p)p(k,l-1) & else\end{cases}, \quad (5)$$

$$\alpha_d(k,l)=\begin{cases}1 & \text{if} \quad Z(k,l)=1\\ \alpha_a+(1-\alpha_a)p(k,l) & else\end{cases}, \quad (6)$$

여기서 $\alpha_p(0<\alpha_p<1)$, $\alpha_a(0<\alpha_a<1)$는 시간에 따라 변하는 최적의 스무딩 파라미터를 나타낸다.

음성존재구간에서 잡음추정을 위한 최적의 스무딩 함수 $\alpha_d(k,l)=1$이면 잡음추정은 즉시 정지함으로써, 실제적인 음성존재구간에서 잘못된 잡음추정을 방지할 수 있다. 만약 VAD $Z(k,l)$가 0이라면, 음성존재확률은 높은 스무딩 함수 $\alpha_d(k,l)$를 가지고 회귀적으로 감소되며 잡음추정이 천천히 시작되어 음성구간에서 약한 음성요소를 보호할 수 있다.

마지막으로 잡음전력은 음성존재구간과 비 음성존재구간에서 스무딩 파라미터 $\alpha_d(k,l)$을 이용하여 다음과 같이 추정된다.

$$\tilde{\lambda}_d(k,l)=\alpha_d(k,l)\tilde{\lambda}_d(k,l-1)+(1-\alpha_d(k,l))|X(k,l)|^2 \quad (7)$$

즉, 변화하는 환경에 따른 잡음의 파워를 추정하기 위해 현재 프레임의 음성검출 결과를 기준으로 잡음구간이라고 판단될 경우에만 잡음의 파워가 갱신된다.

2. 음성추정

효과적인 잡음제거를 위해서는 MMSE (Minimum Mean-Square Error)기반의 음성추정방식이 사용되며, 음성추정은 <그림 3>과 같이 구성된다.

먼저, 추정된 잡음전력을 이용하여 사전 신호대잡음비(a priori SNR) $\xi(k,l)$와 사후 신호대잡음비(a posteriori SNR) $\gamma(k,l)$를 다음과 같이 구한다.

$$\gamma(k,l)=\begin{cases}\dfrac{|X(k,l)|^2}{\lambda_d(k,l)}, & \text{if}\,|X(k,l)^2>\lambda_d(k,l)| \\ 1 & else\end{cases} \tag{8}$$

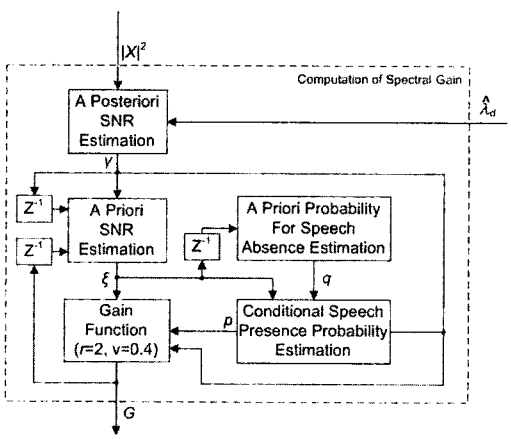
$$\upsilon(k,l)=\left(\frac{\xi(k,l)}{1+\xi(k,l)}\right)\gamma(k,l) \tag{9}$$

$$\xi(k,l)=\beta G^2(k,l-1)\gamma(k,l)+(1-\beta)\xi(k,l-1) \tag{10}$$

여기서 $\beta(0<\beta<1)$은 스무딩 파라미터를 나타낸다.

$$\zeta_{local}(k,l)=\sum_{i=-1}^{i=1}b(i)\zeta(k,l-1) \tag{11}$$

$$\zeta(k,l)=\sigma\zeta(k,l-1)\gamma(k,l)+(1-\sigma)\xi(k,l-1) \tag{12}$$



<그림 3> 스펙트럼 이득 계산의 구성도

<Fig. 3> Block diagram of computation of spectral gain

A priori SNR의 시간 및 주파수 축에 대한 스무딩을 식 (11)과 (12)를 이용하여 구한 평균 a priori SNR $\zeta(k,l)$로부터 현재 프레임에 대한 음성존재 비율 $p_s(k,l)$을 다음과 같이 획득한다.

$$p_s(k,l)=\begin{cases}0 & \text{if}\ \zeta_{local}(k,l)\le -10dB \\ 1 & \text{if}\ \zeta_{local}(k,l)\ge -5dB \\ \sin^2\left(2\pi\dfrac{\zeta_{local}(k,l)-\zeta_{\min}}{\zeta_{\max}-\zeta_{\min}}\right) & otherwise\end{cases} \tag{13}$$

구해진 $p_s(k,l)$을 이용하여 음성부재 추정에 대한 사전확률 $q(k,l)=1-p(k,l)$을 구하고, 식 (14)를 통해 조건부 음성존재 추정 비율을 아래와 같이 얻는다.

$$p(k,l)=\frac{1}{1+\dfrac{q(k,l)}{1-q(k,l)}(1+\xi(k,l)\exp(-\upsilon(k,l)))} \tag{14}$$

목표하는 음성향상 스펙트럼은 $p(k,l)$, $q(k,l)$과 결합된 잡음제거이득과 오염된 음성신호의 곱을 통해 다음과 같이 획득된다.

$$\tilde{S}(k,l)=X(k,l)\cdot G(k,l) \tag{15}$$

$$G(k,l)=(G_G(k,l))^{p_s(k,l)}(G_{\min})^{q(k,l)} \tag{16}$$

위에서 기술된 바와 같이 잡음제거이득 $G(k,l)$은 MMSE 음성추정에서 최적의 이득을 계산하기 위해 Rayleigh 분포보다 우수한 성능을 보이는 generalized Gamma분포기반의 음성추정 이득 $G_G(k,l)$과 SGF G_{min}로 구성된다.

식 (18)에 나타난 Generalized Gamma분포기반의 음성추정기법을 통해 음성향상 스펙트럼은 식 (17)과 같이 표현되며, $f_{S|X}(S_r,S_i|X)$는 오염된 음성 신호 X가 존재할 때 실수부 S_r와 허수부 S_r를 가지는 음성신호 S의 조건부확률분포를 나타낸다.

$$\tilde{S}=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(S_r+jS_i)f_{S|X}(S_r,S_i|X)dS_rdS_i \tag{17}$$

$$f_A(a)=\frac{\kappa\beta^{\nu}}{\Gamma(\nu)}a^{\kappa\nu-1}\exp(-\beta a^{\kappa}),\quad \beta\ge 0,a\ge 0,\nu\ge 0,\kappa\ge 0 \tag{18}$$

여기서 κ는 scale 파라미터, ν, β는 shape 파라미터를 나타내며, 랜덤변수 a는 음성 신호의 스펙트럼 크기를 나타낸다.

추정된 음성 $\hat{S}$는 오염된 음성 신호와 잡음제거이득의 곱으로 표현되고, 잡음제거 이득은 베이즈 법칙을 통해 식 (19)와 같이 유도된다.

$$G_G = \frac{1}{r}\frac{\int_0^{\infty}\int_{-\pi}^{+\pi} a e^{j(\phi-\theta)} f_{X|S}(x|a,\phi) f_A(a) d\phi da}{\int_0^{\infty}\int_{-\pi}^{+\pi} f_{X|S}(x|a,\phi) f_A(a) d\phi da} \quad (19)$$

식 (18)로부터 특정한 값 κ, ν의 closed-form solution와 Bessel 함수를 추정함으로써 실질적으로 음성향상에 generalized Gamma 분포를 적용시킨 잡음제거이득 $G_G(k,l)$을 최종적으로 식 (20)과 같이 구할 수 있다.

$$G_G(k,l) = \frac{\nu\xi(k,l)}{\nu+\xi(k,l)} \frac{M\left(\nu+1;2;\frac{\gamma(k,l)\xi(k,l)}{\nu+\xi(k,l)}\right)}{M\left(\nu;1;\frac{\gamma(k,l)\xi(k,l)}{\nu+\xi(k,l)}\right)} \quad (20)$$

여기서 $M(a;b;z)$은 CHF (confluent hypergeometric function)을 나타내며, a와 b는 CHF의 coefficient, z는 랜덤변수이다. $M(a;b;z)$는 아래와 같이 표현된다.

$$M(a,b,z) = 1 + \frac{az}{b} + \frac{(a)_2 z^2}{(b)_2 2!} + \ldots + \frac{(a)_n z^n}{(b)_n n!} + \ldots \quad (21)$$

G_{min}은 spectral floor constant로서 $G_G(k,l)$에 의한 잡음제거로 인해 음성스펙트럼의 일부가 제거되지 않도록 방지하는 trade-off 문턱값을 의미한다. 즉, 현재 시간 프레임-주파수 축 상에서 음성성분이 부재되면 음성향상 이득은 문턱값 G_{min}보다 커지게 함으로써 음성검출영역 성분을 보존하도록 한다.

Ⅲ. 실험 및 결과고찰

본 논문에서 제안된 음질향상 알고리즘과 기존의 잘 알려진 참고문헌 [2, 3]의 알고리즘의 성능을 객관적으로 평가하기 위해 ETSI AURORA2 데이터베이스를 이용하여 음성인식 성능을 측정하였다.

음성 인식기는 HTK를 사용하여 단어당 16개의 상태를 가지고 단어 단위 HMM으로 모델링하였다. 또한 특징벡터로서 12개의 MFCC와 로그 프레임 에너지, 그것의 증분, 증분의 증분으로 정의되는 총 39차를 사용하였다. 성능평가는 모든 잡음에 대한 SNR 0dB부터 20dB사이에서의 인식률 평균으로 나타내었다.

<표 1>은 기존의 참고문헌 [2-4]의 방식들과 본 논문에서 제안된 알고리즘과의 성능을 비교한 도표이다.

WNR, MM, OM, GG은 각각 잡음제거가 없는 방식, [3]에서 제안된 Gaussian 분포기반의 log-spectral amplitude 음성향상방식, [2]에서 사용된 optimally-modified log-spectral amplitude 음성향상방식, 그리고 본 논문에서 제안된 방식을 나타낸다.

제안한 알고리즘은 Set A에서 85.72%, Set B에서 83.67%, Set C에서 83.50%로 기존 알고리즘에 비해 향상된 인식성능을 나타내었고, 세부적으로는 자동차와 기차역 배경잡음에서 높은 인식률을 보였다.

또한, 제안된 음성향상 알고리즘을 spectral component, spectral amplitude, 그리고 log-spectral amplitude에 적용하여 음성인식 성능을 측정하여 <표 2>에 나타내었다.

<표 2>의 결과를 통해 $\kappa=0.4$, $\kappa\nu-1=2$의 파라미터 값에서 log-spectral amplitude기반의 음성향상 추정기법은 spectral component와 spectral amplitude기반의 음성추정기법 보다 전체적으로 높은 음성 인식률을 보임을 알 수 있다. 5dB와 20dB 사이의 SNR에서는 3가지 방식들의 음성 인식률은 매우 비슷하고, 특히 0dB SNR이하의 $\kappa=0.4$, $\kappa\nu-1=2$에서 log-spectral amplitude 추정기법이 다른 추정기법의 인식률보다 높음을 알 수 있다. 반면에, $\kappa=0.4$, $\kappa\nu-1=2$를 갖는 spectral amplitude 추정기법은 $\kappa=2$, $\kappa\nu-1=1$에서

<표 1> 인식률 성능
<Table 1> Accuracy of the recognition rates

잡음제거 방식	Set A	Set B	Set C	평균
WNR	61.37	56.20	66.58	61.38
MM	79.28	78.82	81.13	79.74
OM	80.34	79.03	81.23	80.20
GG	85.72	83.67	83.50	84.30

<표 2> 인식률 성능
<Table 2> Accuracy of the recognition rates

	Spectral Component		Spectral Amplitude		Log-Spectral Amplitude	
SNR	κ=2, $\kappa\nu$-1=1	κ=0.4, $\kappa\nu$-1=2	κ=2, $\kappa\nu$-1=1	κ=0.4, $\kappa\nu$-1=2	κ=2, $\kappa\nu$-1=1	κ=0.4, $\kappa\nu$-1=2
20	98.69	98.62	98.71	98.74	98.73	98.80
15	98.62	98.68	98.34	98.62	98.62	98.59
10	97.94	98.31	97.54	98.43	98.07	97.94
5	94.57	94.96	93.86	95.61	94.07	95.24
0	72.46	76.48	73.26	77.25	76.05	79.06
-5	26.04	31.90	27.42	31.56	30.06	36.17
Average	81.39	81.16	81.52	83.39	82.60	84.3

log-spectral amplitude 추정기법보다 약간 우수한 인식률을 나타내었다.

V. 결 론

본 논문에서는 generalized Gamma 분포와 최적이득인 SGF를 결합하여 자동차 환경에서의 음성인식 성능을 향상시키기 위한 음성향상 알고리즘을 제안하였다. 제안된 알고리즘은 기존에 사용된 MM, OM 방식 뿐만 아니라 단순한 generalized Gamma 분포기반의 음성향상 알고리즘보다 음성인식 성능이 우수함을 나타내었다. 그리고 제안한 음성추정 기법을 log-spectral amplitude에 적용한 경우가 spectral component와 spectral amplitude에 적용한 음성향상 알고리즘 보다 $\kappa=0.4$, $\kappa\nu-1=2$ 파라미터 값에서 보다 나은 음성인식률의 성능을 보여주었다.

향후 계획으로는 generalized Gamma 분포와 SGF를 결합한 log-amplitude 음성추정기반의 음성인식엔진을 영상검색 시스템에 적용해 보고자 한다.

참 고 문 헌

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 5, pp. 504-512, July 2001.

[2] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments," *Signal Processing*, Elsevier, vol. 81, no. 11, pp. 2403-2418, Nov. 2001.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 33, no. 2, pp. 443-445, Dec. 1985.

[4] R. C. Hendriks, J. S. Erkelens, J. Jensen, and R. Heusdens, "Minimum mean-sqaure error amplitude estimators for speech enhancement under the generalized Gamma distribution," *Proc. International Workshop on Acoustic Echo and Noise Control(IWAENC)*, vol. 10, pp. 1-4, Sept. 2006.

저자소개

김 형 국 (Kim, Hyoung-Gook)

2007년 3월 ~ 현재 : 광운대학교 전파공학과 부교수

신 동 (Shin, Dong)

2009년 3월 ~ 현재 : 광운대학교 전파공학과 석사과정
2009년 : 광운대학교 전파공학과 공학사

이 진 호 (Lee, Jin Ho)

2003년 3월 ~ 현재 : 광운대학교 전파공학과 학사과정