한국음향학회지 제30권 제7호 pp. 383~389 (2011)

환경잡음분류 기반의 향상된 음성부재확률 추정

An Improved Speech Absence Probability Estimation based on Environmental Noise Classification

손 영 호, 박 윤 식, 안 홍 섭, 이 상 민
(Young-Ho Son, Yun-Sik Park, Hong-Sub An, Sangmin Lee)

인하대학교 전자공학부
(접수일자: 2011년 6월 25일; 수정일자: 2011년 8월 8일; 채택일자: 2011년 8월 29일)

본 논문에서는 음성향상을 위하여 환경잡음분류를 적용한 향상된 음성부재확률 추정방법을 제안한다. 기존의 음성부재확률 추정방법에서는 마이크로폰 입력신호와 추정된 잡음신호 기반의 a *posteriori* SNR값에 문턱값을 적용하여 음성부재확률을 구하는데 필요한 음성부재의 a *priori* 확률을 도출하였다. 본 논문에서 제안된 알고리즘은 보다 효과적인 음성부재확률 추정을 위하여 고정된 문턱값과 스무딩 (smoothing)파라미터를 사용하는 기존의 방법과는 달리 잡음분류 알고리즘인 가우시안 혼합 모델 (Gaussian mixture model)을 사용하여 잡음마다 최적화된 파라미터를 적용한다. 제안된 음성 향상 기법은 ITU-T P.862 PESQ (perceptual evaluation of speech quality)와 composite measure를 이용하여 다양한 환경에서 평가하였으며, 제안된 알고리즘이 기존의 음성부재확률 추정방법보다 향상된 결과를 보였다.
핵심용어: 음성부재확률 추정, 가우시안 혼합 모델, 잡음분류
투고분야: 음성처리 분야 (2.3)

In this paper, we propose a improved speech absence probability estimation algorithm by applying environmental noise classification for speech enhancement. The previous speech absence probability required to seek a *priori* probability of speech absence was derived by applying microphone input signal and the noise signal based on the estimated value of a *posteriori* SNR threshold. In this paper, the proposed algorithm estimates the speech absence probability using noise classification algorithm which is based on Gaussian mixture model in order to apply the optimal parameter each noise types, unlike the conventional fixed threshold and smoothing parameter. Performance of the proposed enhancement algorithm is evaluated by ITU-T P.862 PESQ (perceptual evaluation of speech quality) and composite measure under various noise environments. It is verified that the proposed algorithm yields better results compared to the conventional speech absence probability estimation algorithm.
Keywords: Speech absence probability, Gaussian mixture model (GMM), Noise Classification
ASK subject classification: Speech Signal Processing (2.3)

I. 서 론

최근 스마트폰과 차량 내비게이션 등 실제적인 음성신호처리 시스템이 필요한 환경이 늘어나면서 음성인식과 음성향상에 관련된 관심이 증가 하였으며, 많은 알고리즘이 연구되었다. 전통적인 잡음추정 방법으로는 음성검출기 (VAD, voice activity detector)에 의존하여 음성 부재 구간에서 잡음의 평균을 구하는 방법이 사용되었다 [1-4]. 이러한 방법은 조정이 어려우며 신호 대 잡음비가 낮은 응용분야에 사용될 경우 정확한 추정이 불가능하여 왜곡된 음성 신호가 출력되는 오류가 발생하기 쉽다는 단점을 가지고 있다. 음성검출기의 대안으로 나온 최소 통계잡음 추정 (MMSE, minimum statistics noise estimation) [5]의 경우에는 음성검출기를 사용하지 않으며 최적으로 신호의 파워 스펙트럼을 스무딩 (smoothing)하여 최소전력을 계산한다. 전력 스펙트럼 스무딩 알고리즘은 시간과 주파수에 종속관계가 있는 스무딩 매개변수로 1차 회귀 시스템을 사용하는데 이때 스무딩 매개변수의 최적화를 통하여 조건적인 평균 자승 오차를 최소로 하여 비정상 잡음신호를 추정한다. 하지만 기존의 방법들보다 두 배의 분산을 가져 특이점에서만 민감하게 반응

책임저자: 이 상 민 (sanglee@inha.ac.kr)
402-751 인천시 남구 용현동 253 인하대학교 전자공학부
(전화: 032-860-7420; 팩스: 032-868-3654)

하고 최소 탐색 윈도우를 작게 하였을 시 작은 에너지를 갖는 음소를 약하게 만드는 단점을 지니고 있다 [6]. 단점들을 보완하여 계산량을 줄이고 효율적인 최소 추적방법이 제안되기도 하였다 [7]. 하지만 잡음신호의 에너지가 급상승하는 경우엔 잡음추정이 느려지고 신호를 소멸시킨다는 문제점을 지니고 있다 [8]. 다른 방법으로는 soft decision이 있으며, soft decision은 음성 영역에서도 잡음 신호의 파워 스펙트럼을 추정한다 [9-10].

한편 기존의 Malah가 제안한 음성존재 부정확성 추적방법은 soft decision에서 음성부재확률 (SAP, speech absence probability)을 구할 때 사용되는 음성부재의 a *priori* 확률을 a *posteriori* SNR과 특정 문턱값과의 비교를 통하여 음성인지 비음성인지 판별하여 음성부재의 a *priori* 확률을 각 프레임과 주파수 밴드마다 다르게 적용, 음성부재확률의 성능을 향상 시켰다 [11]. 하지만 음성부재의 a *priori* 확률을 얻기 위해 사용되는 문턱값과 스무딩 파라미터가 고정되어 있어 다양한 잡음 환경에 적용될 경우 음성 향상의 결과가 좋지 않을 수 있다.

본 논문에서는 향상된 음성부재확률을 구하기 위해 잡음 분류 알고리즘인 가우시안 혼합 모델 (GMM, Gaussian mixture model)을 적용하여 각 잡음 종류에 따라 최적화된 파라미터를 적용하여 음성부재확률에 신뢰도를 높이는 방법을 제안한다 [12-13]. 그 결과 객관적 음질 평가 방법인 ITU-T P.862 PESQ (perceptual evaluation of speech quality) [14]와 composite measure [15] 그리고 MOS (mean opinion score)를 테스트 하여 기존의 음성부재확률 추정방법보다 향상된 성능을 보였다.

II. 음성존재 부정확성 추적방법 고찰

먼저 오염된 음성신호 $y(i)$는 원래의 음성신호 $s(i)$에 잡음신호 $n(i)$가 더해져서 만들어 졌다고 가정하며, 각각의 성분을 discrete Fourier transform (DFT)를 통해서 주파수 축으로 다음과 같이 나타낼 수 있다.

$$Y(t,k) = S(t,k) + N(t,k) \tag{1}$$

여기서 t는 프레임, k는 주파수를 의미하며 음성의 부재와 존재에 대한 기본 가설로 사용되는 $H_0(t,k)$와 $H_1(t,k)$는 다음과 같은 식으로 정의 내릴 수 있다.

$$\begin{aligned} H_0(t,k) &: Y(t,k) = N(t,k) \\ H_1(t,k) &: Y(t,k) = S(t,k) + N(t,k) \end{aligned} \tag{2}$$

여기서 $S(t,k)$와 $N(t,k)$는 제로 평균 복소 가우시안 분포를 가진다고 가정하며, 음성과 잡음신호의 스펙트럼을 각각 나타내고 있다. 주어진 두 가설을 조건으로 한 확률 밀도 함수는 다음과 같이 나타낼 수 있다 [10].

$$\begin{aligned} p(Y(t,k)|H_0) &= \frac{1}{\pi\lambda_d(t,k)} \exp\left[-\frac{|Y(t,k)|^2}{\lambda_d(t,k)}\right], \\ p(Y(t,k)|H_1) &= \frac{1}{\pi\left[\lambda_s(t,k)+\lambda_d(t,k)\right]} \\ &\quad \cdot \exp\left[-\frac{|Y(t,k)|^2}{\lambda_s(t,k)+\lambda_d(t,k)}\right] \end{aligned} \tag{3}$$

여기서 $\lambda_s(t,k)$와 $\lambda_d(t,k)$은 t번째 프레임과 k번째 주파수의 음성과 잡음의 분산을 나타낸다. 위의 가설로부터 입력신호 $Y(t,k)$의 음성 부재 확률은 아래와 같이 주어진다.

$$\begin{aligned} p(H_0|Y(t,k)) &= \frac{p(Y(t,k)|H_0)p(H_0)}{p(Y(t,k)|H_0)p(H_0)+p(Y(t,k)|H_1)p(H_1)} \\ &= \frac{1}{1+\frac{1-q}{q}\Lambda(Y(t,k))} \end{aligned} \tag{4}$$

여기서 $\Lambda(Y(t,k))$는 우도비 (likelihood ratio, LR)이며, q는 $p(H_0) = (1-p(H_1))$를 의미하며 보통 고정된 q로 0.5로 설정 되었다 [11]. 하지만 음성존재 부정확성 추적방법은 0.5로 고정된 q를 각 프레임과 주파수 밴드에 의존하는 q를 $q(t,k)$라 가정하면 음성부재확률은 다음과 같이 구할 수 있다.

$$p(H_0|Y(t,k)) = \frac{1}{1+\frac{1-q(t,k)}{q(t,k)}\Lambda(Y(t,k))} \tag{5}$$

여기서 $q(t,k)$는 현재 신호의 음성존재의 판별을 통하여 현재 프레임에 대한 가중치를 결정하며, 다음과 같이 표현된다.

$$q(t,k) = \alpha_p q(t,k-1) + (1-\alpha_p) I(t,k) \tag{6}$$

여기서 α_p는 스무딩 파라미터로 0.95로 설정되며, $I(t,k)$는 아래의 식을 통해 구해진다.

$$\begin{aligned} I(t,k) &= 0 \,;\, \text{if } \gamma(t,k) < \gamma_{TH} \\ I(t,k) &= 1 \,;\, \text{if } \gamma(t,k) > \gamma_{TH} \end{aligned} \tag{7}$$

여기서 $\gamma(t,k)$는 a *posteriori* SNR을 의미하며, γ_{TH}는 문턱값이며 0.8로 설정되었다. $\gamma(t,k)$가 γ_{TH}보다 작다고 판단될 경우 0으로 $I(t,k)$로 설정되며, $\gamma(t,k)$가 γ_{TH}보다

크다고 판단될 경우 1로 $I(t,k)$가 설정되어 $q(t,k)$를 구한다.

III. Gaussian Mixture Model 기반의 환경 잡음 분류

지금까지 우리는 음성존재 부정확성 추적방법에서의 음성부재확률을 구하는 방법에 대해 알아보았다. 하지만 기존의 음성존재 부정확성 추적방법에서는 q를 구하기 위하여 고정된 문턱값과 스무딩 파라미터를 사용하여 수시로 변하는 잡음환경에서 정확한 음성 부재 확률을 추정하지 못하였다. 하지만 잡음 분류 알고리즘인 GMM을 적용하여 고정된 문턱값과 스무딩 파라미터 값을 잡음환경에 최적화하여 향상된 음성 부재확률을 도출하였다. 최적화된 문턱값과 스무딩 파라미터를 찾기 위해서 0.1에서 1.0까지 0.01단위로 변화시켜가며 테스트를 진행하였다. 각 잡음별로 가장 음질이 우수하도록 결정하기 위해 객관적인 음질평가인 Composite measure를 사용 하였으며, 다음과 같이 구성되어 있다.

$$C_{ovl} = 1.549 + 0.805 PESQ - 0.512 LLR - 0.007 WSS \tag{8}$$

여기서 로그 우도 비 (LLR, log-likelihood ration)은 깨끗한 신호와 잡음 처리가 된 신호의 각각에 대해 추출된 LPC를 이용하여 복원된 신호의 차이를 로그 스케일로 측정하는 측정법을 나타내고, weighted-slpoe spectral distance (WSS)는 정해진 프레임 내에서 인접한 주파수 밴드 사이의 관계의 왜곡도를 측정하는 측정법을 나타낸다. Composite measure에서 사용하는 PESQ는 기존의 PESQ에서 음성의 왜곡과 잡음의 왜곡에 대한 측정치에 가중치를 더 주도록 수정된 측정법을 나타내며, 실험에 사용한 잡음은 NOISEX-92의 대표적인 잡음인 babble, car, office, white을 사용하여 테스트를 진행하였다. 테스트 결과, 주어진 잡음 종류에 따른 최적의 값을 표 1에서 보여주며 잡음 환경에 따라 다른 최적의 값을 갖는 것을 볼 수 있다. 표 1을 통해서 비정상상태 잡음인 babble, office의 경우 기존의 파라미터 값보다 큰 값을 가졌으며, 정상상태 잡음인 car와 white잡음의 경우 기존의 파라미터 값보다 작은 값을 갖는 것을 알 수 있다.

그림 1은 문턱값 γ_{TH}와 스무딩 파라미터 α_p가 변함에 따라 발생하는 composite measure값을 3D mesh 곡선을 보여준다. 그림 1은 비 정상상태 잡음인 babble 10 dB로 composite measure 결과 기존의 0.8과 0.95로 고정되어

표 1. 다양한 잡음 환경에서 composite measure 수치 비교를 통한 최적화된 문턱값, 스무딩 파라미터 (frame)
Table 1. Optimal threshold and smoothing parameter (frame) is selected by comparing composite measure score.

Noise type	Threshold γ_{TH}	Smoothing parameter α_p
Babble noise	0.98	0.98
Car noise	0.20	0.27
office noise	0.97	0.96
White noise	0.28	0.25

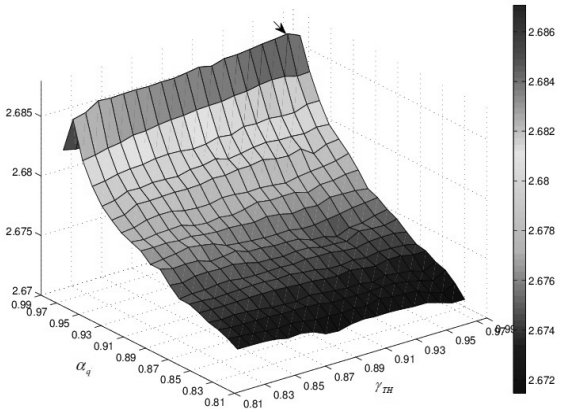

그림 1. Babble 잡음 (SNR = 10 dB) 에서의 최적의 동작 값
Fig. 1. 3D mesh curve of the optimal operating values for the babble noise 10 dB SNR.

있는 문턱값과 스무딩 파라미터와 달리 0.98과 0.98에서 최적화된 값을 갖는 것을 그림 1을 통해 볼 수 있다. 3D mesh 곡선을 통하여 잡음마나 최적화된 파라미터가 존재 한다는 것을 알 수 있다.

잡음 분류를 위해 사용된 GMM은 패턴 인식기로 주어진 표본데이터 집합의 분포 밀도를 단 하나의 확률 밀도 함수로 모델링 하는 방법을 개선한 밀도 추정 방법으로 복수 개의 가우시안 확률 밀도 함수로 데이터의 분포를 모델링하는 방법이다. GMM은 훈련부와 인식부로 나누어지며, 훈련부에서 사용되는 잡음구간이 인식부에서 다시 사용되는 것을 방지하기 위하여 잡음의 구간을 구분하여 특성잡음에서 특성화 되지 않도록 분류하여 사용하였다. 훈련부에서는 잡음의 모델을 각각 만들고 인식부에서는 이 모델을 이용하여 잡음을 인식한다. 분류 시스템에서 사용되는 GMM은 가우시안 밀도의 혼합성분 가중치 합인 함수로서 다음과 같이 표현된다.

$$P(\vec{x}|\lambda) = \sum_{i=1}^{M} \alpha_i P_i(\vec{x}) \tag{9}$$

$$P_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{L}{2}} \left|\sum_i\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T (\sum_i)^{-1} (\vec{x}-\vec{\mu_i})\right\} \tag{10}$$

여기서 Σ_i는 공분산 행렬, α_i는 혼합 성분의 가중치를 나타내고, $\vec{\mu_i}$는 평균 벡터를 나타낸다.

$$\lambda = \left\{ \alpha_i, \vec{\mu_i}, \Sigma_i \right\}, \quad i = 1,...,M \tag{11}$$

여기서, 훈련부는 식 (11)와 같은 파라미터를 가지고 Expectation Maximization (EM)알고리즘 기반의 학습을 통하여 잡음에 대한 혼합 가우시안 모델 λ를 추정하고 인식부는 훈련부에 만들어진 λ를 이용해서 입력된 음성신호에 대한 사후 확률을 구하여 가장 큰 확률을 갖는 모델을 찾는다. 실제로 구성된 모델 외에 판별이 불가능한 잡음의 대안으로 Universal background models (UBM)을 두어 잡음 판별이 안 될 경우 기존의 문턱값 0.8, 스무딩 파라미터 0.95로 설정하였다. 잡음 분류를 위해 실시간으로 데이터의 특징벡터를 입력을 받으며 사용된 특징벡터는 총14차이며 그 특징벡터는 자기 상관함수 (Autocorrelation Function)와 반사계수 (Reflection Coefficients)를 사용한 Levinson-Durbin 알고리즘을 사용하여 구한 Linear Prediction Coding (LPC) 계수 10차, LPC 분석에서의 에러 성분에 대한 잔류 에너지 1차, 프레임 에너지의 이동평균 1차, 최소값 10을 가진 프레임의 에너지 1차, 잔류에너지의 이동평균1차이다. 이것을 기반으로, 각각의 잡음에서 GMM 파라미터 λ는 $\lambda_n (= \lambda_1, \lambda_2, \lambda_3, \lambda_4)$로 설정하여 $n=1$인 경우 babble 잡음, $n=2$인 경우 car 잡음, $n=3$인 경우 office 잡음, $n=4$인 경우 white 잡음, $n=5$인 경우 UBM (universal background noise) 잡음으로 선택되도록 하였다. 또한, 잡음 업데이트 구간에서 갑작스러운 변화를 막기 위하여 long-term 스무딩을 다음과 같이 한다.

$$\log_p(\vec{x}(t)|\lambda_n) = \zeta \log_p(\vec{x}(t-1)|\lambda_n) + (1-\zeta)\log_p(\vec{x}(t)|\lambda_n) \tag{12}$$

여기서, ζ는 스무딩 파라미터로 0.9로 설정되며, 다음과 같이 추정된 구간별 최적모델에 따른 우도 (likelihood)값을 비교하여 가장 큰 우도를 갖는 것으로 분류한다.

$$\hat{n}(t) = \underset{n=1,2,3,4,5}{\arg max} \sum_{t=1}^{N} \log p(\vec{x}(t)|\lambda_n), \quad n = 4(1: babble, 2: car, 3: office, 4: white, 5: UBM) \tag{13}$$

그림 2는 기존의 음성부재확률을 구하는 구조에 GMM 기반의 잡음 분류 알고리즘을 통합한 구조에 대한 블록도를 보여주고 있다. 통합 알고리즘에서 배경잡음 $d(t)$, 근단 화자신호 $s(t)$라 하고 $Y(t,k)$를 $y(t)$의 t번째 프레임의 k번째 주파수 성분이라 하면 입력신호 $Y(t,k)$로부터 잡음 분류 알고리즘인 GMM을 통하여 분류된 잡음 종류에 최적화된 파라미터 값을 적용하여 향상된 음성부재확률을 추정한다. 기존의 음성존재 부정확성 추적방법에서 사용되는 문턱값과 스무딩 파라미터는 잡음 환경에 관계없이 고정값이 적용되기 때문에 a *posteriori* SNR과의 문턱값의 비교에서 잡음을 음성으로 혹은 음성을 잡음으로 인식하여 음성부재확률의 신뢰도가 떨어진다. 또한 스무딩 파라미터는 전 프레임에 가중치에 따라 q값이 결정되기 때문에 최적화된 파라미터적용이 매우 중요하다. 이러한 기존의 음성존재 부정확성 추적방법을 보완하기 위하여 분류된 잡음 정보 $\hat{n}(t)$를 기반으로, 식 (7)의 문턱값 γ_{TH}와 식 (6)의 스무딩 파라미터 α_p를 $\hat{\gamma}_{TH}(t)$와 $\hat{\alpha}_p(t)$로 표 1을 기반으로 아래와 같이 적용한다.

$$\begin{aligned} I(t,k) &= 0; \text{ if } \gamma(t,k) < \hat{\gamma}_{TH}(t) \\ I(t,k) &= 1; \text{ if } \gamma(t,k) > \hat{\gamma}_{TH}(t) \end{aligned} \tag{14}$$

$$\hat{q}(t,k) = \hat{\alpha}_p(t)\hat{q}(t,k-1) + (1-\hat{\alpha}_p(t))I(t,k) \tag{15}$$

여기서, $\hat{\gamma}_{TH}(t)$의 갑작스러운 변화를 방지하기 위하여 long-term 스무딩을 해주며 수식적으로 다음과 같이 표현된다.

$$\hat{\gamma}_{TH}(t) = \epsilon\hat{\gamma}_{TH}(t-1) + (1-\epsilon)\hat{\gamma}_{TH}(t) \tag{16}$$

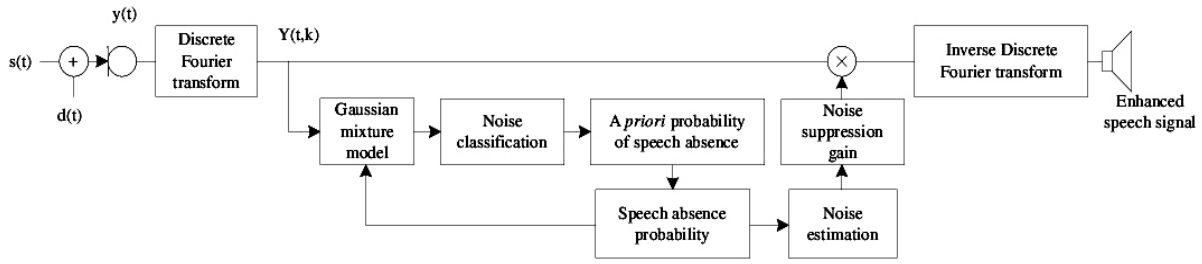

그림 2. 제안된 음성부재확률 추정방법에 잡음분류 알고리즘을 적용한 블록도
Fig. 2. Illustration of estimation of prior distribution in speaker space.

여기서 ϵ는 스무딩 파라미터이며 0.9값을 갖는다. 우리는 $\hat{q}(t,k)$를 통하여 새로운 음성 부재 확률을 아래와 같이 도출하였다.

$$p(H_0|Y(t,k)) = \frac{1}{1+\frac{1-\hat{q}(t,k)}{\hat{q}(t,k)}\Lambda(Y(t,k))} \tag{17}$$

여기서 $\Lambda(Y(t,k))$는 다음과 같이 표현된다.

$$\Lambda(Y(t,k)) = \frac{p(Y(t,k)|H_1)}{p(Y(t,k)|H_0)} = \frac{1}{1+\xi(t,k)} \exp\left[\frac{\gamma(t,k)\xi(t,k)}{1+\xi(t,k)}\right] \tag{18}$$

여기서 $\gamma(t,k)$, $\xi(t,k)$는 각각 a *posteriori* SNR와 a *priori* SNR로 아래와 같이 정의된다 [1].

$$\xi(t,k) = \frac{\lambda_s(t,k)}{\lambda_d(t,k)}, \tag{19}$$

$$\gamma(t,k) = \frac{|Y(t,k)|^2}{\lambda_d(t,k)}, \tag{20}$$

스무딩 파라미터에 의한 갱신으로 잡음전력을 추정하는 soft decision 기반의 잡음전력 추정은 long-term 스무딩된 전력 스펙트럼 $\lambda_d(t,k)$는 다음과 같이 업데이트 되어 추정된다.

$$\lambda_d(t,k) = \alpha_d\lambda_d(t-1,k) + (1-\alpha_d)|Y(t,k)|^2 \tag{21}$$

여기서 α_d는 스무딩 파라미터로 0.99로 설정되었으며, 성능평가를 위하여 MMSE (minimum mean square error) shore-time spectral amplitude 기반의 잡음제거 이득 $G(\xi(t,k),\gamma(t,k))$을 가지는 잡음제거기에 적용된다 [1].

또한, GMM 패턴 인식기를 사용하는데 있어서 음성이 섞인 구간에서의 잡음정보 분류를 할 경우 다른 잡음으로 인식하기 때문에 이 같은 오류를 막기 위하여 음성 부재 확률을 이용하여 잡음으로 인식된 구간에서만 GMM의 결과를 사용하였다 [12-13]. 그림 3의 (c)는 (a)와 (b)신호에 대하여 각 구간의 잡음전력 추정의 변화를 보여주고 있다. (c)로부터 잡음 전력 추정에 있어서 기존의 고정된 q와 Malah방법의 경우 잡음 추정 과정에서 왜곡이 발생하여 입력된 잡음의 전력을 정확히 추정하지 못하는 것을 볼 수 있으며, 제안된 방법은 기존이 방법들 보다 잡음 전력 추정에 있어서 오류가 줄어 보다 정확히 추정되는 것을 볼 수 있다. 그림 4의 (c)는 (a)와 (b)신호에 대하여 각 구간의 음성존재확률 변화를 보여주고 있다. (c)로부터 마이크로폰 입력신호에 대한 음성존재확률 $\{1-p(H_0|Y(t,k))\}$

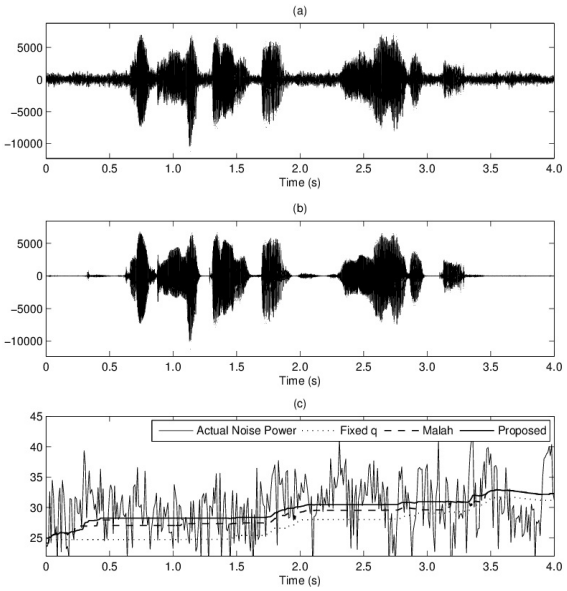

그림 3. White 잡음 (SNR = 10 dB) 에서의 잡음전력 추정 비교 (a) 잡음 섞인 음성 파형, (b) 깨끗한 음성 파형 (c) 실제 잡음전력 (실선), 고정된 q값 기반 추정된 잡음 전력 (점선), Malah 방법 기반 추정된 잡음전력 (일점선) 그리고 제안된 방법 기반 추정된 잡음전력 (굵은선)

Fig. 3. Comparison of noise power estimation (k=3) under White noise (SNR = 10 dB). (a) Noisy speech, (b) Clean speech, (c) Actual noise power (solid line), estimated noise power based on fixed q (dotted line), estimated noise power based on Malah algorithm (dashed line) and estimated noise power based on proposed method (dark line).

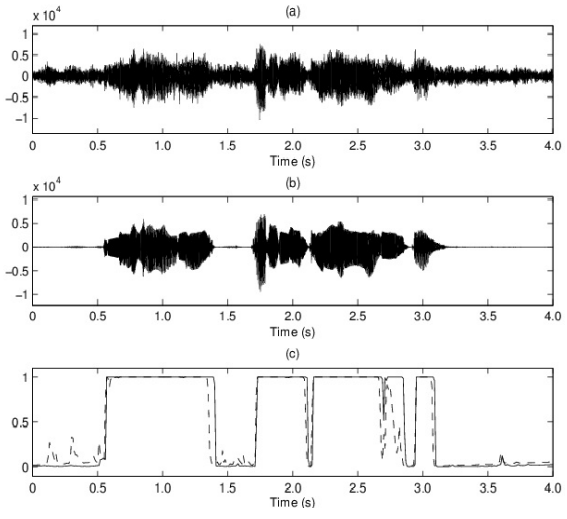

그림 4. Babble 잡음 (SNR = 5 dB) 에서의 음성존재 확률 비교 (a) 잡음 섞인 음성 파형, (b) 깨끗한 음성 파형 (c) 실시간 프레임에서의 음성존재 확률 : 기존 알고리즘의 확률 (일점선), 제안된 알고리즘의 확률 (실선)

Fig. 4. Comparison of speech presence probability (k=3) under Babble noise (SNR = 5 dB). (a) Noisy speech, (b) Clean speech, (c) Speech presence probability in short-time frames probability of conventional algorithm (dashed line), probability of proposed algorithm (solid line).

는 잡음 구간에서는 1에 가까운 값을 나타내지만 그와 반대의 경우 0에 가까운 값을 나타낸다. 기존의 음성부재확률은 잡음이 많이 들어간 경우 잡음과 음성의 구분이 불분명하여 음성임에도 불구하고 음성이 아니라고 판단되는 것을 보여주며 제안된 방법은 기존의 방법보다 음성과 잡음의 판별의 오류를 줄여 주는 것을 볼 수 있다.

IV. 실험 조건 및 결과

본 논문에서는 제안된 음성 향상 알고리즘의 성능을 평가하기 위해 널리 적용되고 있는 ITU-T P.862 PESQ, composite measure 그리고 MOS (mean opinion score) 테스트를 통하여 성능 평가를 하였다. 표 2, 3의 PESQ와 composite measure 테스트를 위해 남성, 여성화자 각각이 35개의 문장을 발음하도록 한 총 70개의 음성 데이터를 한 프레임의 크기를 10 ms에서 8 kHz로 샘플링 하여 네 가지 형태의 잡음이 부가된 오염된 음성을 사용하였고 잡음은 NOISEX-92의 babble, car, office, white 잡음을 사용 하였으며 SNR을 5, 10, 15 dB 세 가지로 나누어 테스트 하였다. 또한 주관적 테스트인 MOS 테스트를 위해 남성, 여성화자 각각이 4개의 문장을 발음하도록 한 총 8개의 음성 데이터를 PESQ와 composite measure 실험과 동일한 방법으로 잡음이 부가된 오염된 음성을 사용하였다. 테스트를 위하여 기존의 고정된 q는 프레임과 주파수 밴드에서 변함없이 $q=0.5$값으로 설정 하였으며, Malah 방법은 각 프레임과 주파수 밴드에 따라 가변하는 q로 설정하였으며, 기존 실험에서 사용된 값인 $\alpha_p=0.95$, $\gamma_{TH}=0.8$로 설정하였다. 또한 Malah 방법보다 최근 알고리즘인 MCRA를 사용하여 제안된 방법과 비교하였다. 표 2, 3 그리고 4는 기존의 음성부재확률 추정방법보다 논문에서 제안한 환경잡음분류 기반의 음성부재확률 추정방법이 PESQ, composite measure 그리고 MOS 테스트 결과 향상된 수치를 보여주고 있다. 이는 제안된 알고리즘의 음성향상 기법이 기존 알고리즘의 음성향상 기법보다 깨끗한 음성신호를 출력 할 수 있도록 음성부재확률을 더 잘 추정함에 따라 음성향상기법의 성능의 향상이

표 2. 잡음 환경에서 고정된 q와 Malah가 제안한 방법과 MCRA 그리고 제안된 알고리즘의 PESQ 수치 비교
Table 2. PESQ score of the fixed q, Malah, MCRA and proposed algorithm.

Noise type	Method	SNR (dB)		
		5	10	15
Babble noise	fixed q	2.331	2.656	2.940
	Malah	2.342	2.669	2.953
	MCRA	2.349	2.677	2.963
	Proposed	**2.365**	**2.690**	**2.981**
Car noise	fixed q	3.488	3.747	3.975
	Malah	3.520	3.766	3.991
	MCRA	3.532	3.778	4.007
	Proposed	**3.559**	**3.795**	**4.026**
Office noise	fixed q	2.326	2.628	2.947
	Malah	2.337	2.634	2.951
	MCRA	2.349	2.651	2.972
	Proposed	**2.368**	**2.667**	**2.985**
White noise	fixed q	2.099	2.434	2.781
	Malah	2.103	2.442	2.786
	MCRA	2.117	2.451	2.794
	Proposed	**2.131**	**2.467**	**2.802**

표 3. 잡음 환경에서 고정된 q와 Malah가 제안한 방법과 MCRA 그리고 제안된 알고리즘의 Composite Measure 수치 비교
Table 3. Composite Measure score of the fixed q, Malah, MCRA and proposed algorithm.

Noise type	Method	SNR (dB)		
		5	10	15
Babble noise	fixed q	2.682	3.065	3.361
	Malah	2.710	3.082	3.397
	MCRA	2.721	3.095	3.412
	Proposed	**2.742**	**3.113**	**3.431**
Car noise	fixed q	3.801	4.087	4.315
	Malah	3.820	4.101	4.336
	MCRA	3.839	4.121	4.339
	Proposed	**3.869**	**4.158**	**4.358**
Office noise	fixed q	2.751	3.084	3.436
	Malah	2.764	3.103	3.451
	MCRA	2.788	3.124	3.469
	Proposed	**2.808**	**3.151**	**3.494**
White noise	fixed q	2.284	2.682	3.064
	Malah	2.305	2.698	3.078
	MCRA	2.329	2.712	3.088
	Proposed	**2.352**	**2.742**	**3.117**

표 4. 잡음 환경에서 고정된 q와 Malah가 제안한 방법과 MCRA 그리고 제안된 알고리즘의 MOS 수치 비교
Table 4. MOS score of the fixed q, Malah, MCRA and proposed algorithm.

Noise type	Method	SNR (dB)		
		5	10	15
Babble noise	fixed q	2.042	2.331	2.718
	Malah	2.071	2.356	2.749
	MCRA	2.089	2.375	2.764
	Proposed	**2.103**	**2.392**	**2.781**
Car noise	fixed q	3.143	3.677	3.975
	Malah	3.177	3.704	4.007
	MCRA	3.198	3.727	4.013
	Proposed	**3.225**	**3.756**	**4.035**
Office noise	fixed q	2.204	2.501	2.859
	Malah	2.241	2.548	2.887
	MCRA	2.253	2.562	2.898
	Proposed	**2.266**	**2.557**	**2.911**
White noise	fixed q	1.839	2.159	2.572
	Malah	1.855	2.174	2.599
	MCRA	1.871	2.190	2.611
	Proposed	**1.902**	**2.212**	**2.628**

있음을 확인 할 수 있었다. 실험에서 사용된 모든 잡음 환경에서 기존의 방법들 보다 제안된 잡음분류 기반의 음성부재확률 추정방법에 있어서 향상된 결과를 갖는 것을 보여준다.

V. 결 론

본 논문에서는 기존의 음성부재확률 추정방법에 잡음 분류 정보를 이용하여 음성부재확률의 신뢰도를 높이는 새로운 알고리즘을 제안하였다. 기존의 방법에서 적용되던 고정된 문턱값과 스무딩 파라미터를 잡음 분류정보를 이용하여 잡음 환경에 따라서 최적화된 파라미터를 프레임마다 선택하도록 하였다. 그 결과, 실험에 사용된 모든 잡음 환경과 신호 대 잡음 비 환경에서 기존의 음성부재확률 추정방법보다 우수한 성능을 보였다.

감사의 글

본 연구는 서울시 산학연 협력사업 (SS100022)의 일환으로 수행하였음.

참 고 문 헌

1. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator." *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.
2. Y. Epharim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 443-445, 1985.
3. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113-120, 1979.
4. J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection" *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, 1999.
5. R. Martin. "Spectral subtraction based on minimum statistics," in *Proc.*, vol. ASSP-27, no. 2, pp. 113-120, 1979.
6. I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403-2418, 2001.
7. G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. 4th EUROSPEECH'95*, Madrid, Spain, pp. 1513-1516, 1995.
8. J. Meyer, K. U. Simmer and K. D. Kammeter, "Comparison of one-and two channel noise-estimation techniques," in *Proc. 5th IWAENC'97*, London, U.K, pp. 137-145, 1997.
9. R. J. McAualy and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 137-145, 1980.
10. N. S. Kim and J. H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, pp. 108-110, 2000.
11. D. Malah, R. Cpx, and A. Accardi, "Tracking speech presence uncertainty to improve speech enhancement in non-stationary noise environments," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 789-792, 1999.
12. G. Xuan, W.Zhang, and P. Chai, "EM algorithm of Gaussian mixture model and hidden Markov model," *Proc. IEEE International Conference on Image Processing*, vol. 1, pp. 145-148, 2001.
13. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.
14. ITU-T P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Feb. 2001.
15. Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement, " *IEEE Transactions on Audio, Speech and Language Processing*, pp. 229-238, 2008.

저자 약력

•손 영 호(Young-Ho Son)

2010년 2월: 수원대학교 전자공학과 학사
2010년 3월 ~ 현재: 인하대학교 전자공학부 석사 과정

•박 윤 식(Yun-Sik Park)

2006년 2월: 인하대학교 전자공학과 학사
2008년 2월: 인하대학교 전자공학부 석사
2008년 3월 ~ 현재: 인하대학교 전자공학부 박사 과정

•안 홍 섭(Hong-Sub An)

2010년 2월: 인하대학교 전자공학과 학사
2010년 3월 ~ 현재: 인하대학교 전자공학부 석사 과정

•이 상 민(Sangmin Lee)

1987년: 인하대학교 전자공학과 학사 졸업
1989년: 인하대학교 전자공학과 석사 졸업
2000년: 인하대학교 전자공학과 박사 졸업
1989년 1월 ~ 1994년 7월: LG 이노텍 선임연구원
1995년 1월 ~ 2002년 3월: 삼성종합기술원 책임 연구원
2002년 4월 ~ 2005년 2월: 한양대학교 의공학교실 연구교수
2005년 3월 ~ 2006년 8월: 전북대학교 생체정보공학부 조교수
2006년 9월 ~ 현재: 인하대학교 전자전기공학부 부교수