한국음향학회지 제29권 제3호 pp. 200~208 (2010)

최소 통계법과 Short-Term 예측계수 코드북을 이용한 Non-Stationary/Mixed 배경잡음 추정 기법

Non-Stationary/Mixed Noise Estimation Algorithm Based on Minimum Statistics and Codebook Driven Short-Term Predictor Parameter Estimation

이 명 석*, 노 명 훈*, 박 성 주**, 이 석 필**, 김 무 영*
(Myeongseok Lee*, Myunghoon Noh*, Sung-Joo Park**, Seok-Pil Lee**, Moo Young Kim*)

*세종대학교 정보통신공학과, **전자부품연구원 디지털미디어연구센터

(접수일자: 2010년 2월 8일; 수정일자: 2010년 3월 23일; 채택일자: 2010년 4월 10일)

본 논문에서는 배경잡음에 강인한 잡음제거 알고리즘 설계를 위해서 minimum statistics (MS) 기법을 codebook driven short-term predictor parameter estimation (CDSTP) 기법에 접목하는 방법을 제안한다. MS는 stationary 배경잡음에는 강인하지만, non-stationary 배경잡음에는 상대적으로 취약하다. CDSTP는 non-stationary 배경잡음에 강인한 특성을 보이지만, 코드북에 없는 배경잡음 환경에는 취약하다. 따라서 non-stationary 배경잡음에 강인한 CDSTP 방법과 별도의 코드북 학습 과정이 필요 없는 MS를 결합해서 다양한 배경잡음에 강인한 알고리즘을 제안한다. 제안방법은 MS나 CDSTP 방법에 비해서 전체적으로 향상된 perceptual evaluation of speech quality (PESQ) 성능을 나타냈으며, 특히 stationary 배경잡음과 non-stationary 배경잡음이 섞여 있는 mixed 배경잡음 환경에서 강인한 특성을 보였다.

핵심용어: Minimum statistics, 잡음추정, 음성향상, non-stationary 배경잡음, mixed 배경잡음
투고분야: 음성처리 분야 (2,3)

In this work, the minimum statistics (MS) algorithm is combined with the codebook driven short-term predictor parameter estimation (CDSTP) to design a speech enhancement algorithm that is robust against various background noise environments. The MS algorithm functions well for the stationary noise but relatively not for the non-stationary noise. The CDSTP works efficiently for the non-stationary noise, but not for the noise that was not considered in the training stage. Thus, we propose to combine CDSTP and MS. Compared with the single use of MS and CDSTP, the proposed method produces better perceptual evaluation of speech quality (PESQ) score, and especially works excellent for the mixed background noise between stationary and non-stationary noises.

Keywords: Minimum statistics, noise estimation, speech enhancement, non-stationary noise, mixed noise
ASK subject classification: Speech Signal Processing (2,3)

I. 서 론

음성향상 기법은 휴대기기를 이용한 통신 및 음악정보처리 (music information retrieval) 분야는 물론이고 로봇제어를 위한 음성인식 등 다양한 분야에서 필요한 기술이다. 음성향상 기법의 성능을 향상시키기 위해서는 잡음을 제거하는 기술 자체도 중요하지만, 근본적으로는 잡음을 추정하는 알고리즘의 성능이 더욱 중요하다.

잡음을 제거하는 기술에는 잡음으로 오염된 음성이 깨끗한 음성과 배경 잡음의 합이라는 가정에서 시작된 spectral subtraction (SS) 방법이 있다 [1]. 또한, 잡음의 변화가 주파수 밴드 별로 독립적이라는 가정을 이용하여 성능을 개선한 multi-band spectral subtraction (MBSS) 방법 등이 있다 [2,3].

잡음 추정 알고리즘은 voice activity detection (VAD)

책임저자: 김 무 영 (mooyoung@sejong.ac.kr)
143-747 서울시 광진구 군자동 세종대학교 정보통신공학과
(전화: 02-3408-3695; 팩스: 02-3408-4330)

을 이용한 방법이나 minimum statistics (MS)를 이용한 방법 등 다양한 알고리즘들이 연구되어 왔다 [4,5]. MS 잡음 추정 알고리즘은 오염된 입력 신호의 최근 D-frame 구간에 대해서 최소 전력 크기를 가지는 프레임이 잡음만 포함하고 있는 프레임이라고 가정을 한다. MS는 stationary 배경잡음 환경에서는 비교적 잘 작동 한다고 알려져 있으나, non-stationary 배경잡음 환경에서는 강인하지 못한 성능을 보인다.

반면, codebook driven short-term predictor parameter estimation (CDSTP)과 같은 방법은 stationary 배경잡음 환경뿐만 아니라 non-stationary 배경잡음 환경에서도 대체적으로 강인한 성능을 보인다 [6]. CDSTP 알고리즘은 음성과 잡음의 스펙트럼을 linear predictive coding (LPC) 계수로 표현하고, 대표적인 스펙트럼 shape들을 LPC 형태로 코드북에 저장한다. 이후 maximum-likelihood estimates 방식을 사용하여 음성과 잡음의 코드북 파라미터와 각각의 gain 값을 추정하여 잡음을 추정하는 알고리즘이다.

하지만, CDSTP는 코드북에 저장된 shape 이외의 잡음 환경에는 취약하므로 본 논문에서는 CDSTP를 MS와 결합하여 사용하는 방법에 대해서 살펴보았다. 또한, 현실에서의 잡음은 대부분 두 가지 혹은 그 이상의 잡음이 섞여 있는 형태로 존재한다. 예를 들어, 전쟁터의 경우 총소리나 비행기 소리 등이 함께 있기 때문에, 본 논문에서는 이러한 mixed 잡음 환경에 대한 음성 향상 기법도 살펴보았다. 이러한 mixed 잡음을 제거 하는데 있어서는 기존의 방법들을 한 가지만 사용 하는 것보다는 서로 다른 특성의 잡음 추정 알고리즘을 섞어서 사용하는 것이 효과적이다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 알고리즘인 MBSS와 MS에 대한 설명이 있겠고, 3장에서는 non-stationary 배경잡음 환경에서도 강인한 CDSTP 알고리즘에 대하여 설명할 것이다. 4장에서는 제안하는 알고리즘에 대한 자세한 설명을 하도록 하겠다. 5장에서는 기존 알고리즘들과 제안된 알고리즘에 대한 비교 실험을 보여준다. 마지막으로 6장에서는 최종 결론을 내리도록 하겠다.

II. Multi-Band Spectral Subtraction (MBSS) and Minimum Statistics (MS)

MBSS는 잡음의 변화가 주파수 밴드 별로 독립적이라는 가정에 근거하며 각 주파수 밴드마다 독립적인 subtraction factor를 이용한다. 한 프레임에서 다음 프레임으로 넘어갈 때 각 주파수 밴드별 잡음이 급격하게 변화한다면 SS 방식은 한계를 가지게 되며 부분적인 음성 왜곡으로 이어지게 된다. MBSS 방식은 음성 스펙트럼을 N개의 오버랩 되지 않은 밴드로 나누고, 각 밴드 별로 독립적으로 SS를 적용해서 잡음을 제거한다.

MBSS가 잡음을 제거하는 과정을 살펴보면, 먼저 잡음으로 오염된 음성이 들어오면 FFT를 통해서 주파수 도메인으로 바꾸고, 아래 식과 같이 주파수 영역에서 윈도우를 씌우는 스무딩 과정을 거치게 된다.

$$|\overline{Y}_j(w_k)| = \sum_{i=-M}^{M} W_i |Y_{j-i}(w_k)|, \quad b_i < w_k < e_i \qquad (1)$$

여기서, W_i는 i 번째 프레임의 가중치 윈도우 값을, M은 스무딩을 위해 필요한 과거와 미래 프레임의 수를, $w_k = 2\pi k/N$은 k 번째 주파수를 (k = 0, 1, ..., N-1), b_i와 e_i는 i 번째 프레임의 처음과 끝 주파수를 나타낸다. 또한, $|Y_i(w_k)|$는 잡음으로 오염된 음성의 i 번째 프레임, k 번째 주파수의 스펙트럼 magnitude를 나타내고, $|\overline{Y}_j(w_k)|$는 j 번째 프레임, k 번째 주파수의 스무딩된 스펙트럼 magnitude를 나타낸다. 원래 스펙트럼 대신 스무딩된 스펙트럼을 사용함으로써 확률적으로 원래 스펙트럼에 비해서 스펙트럼의 분포 범위가 감소되고 잔여 잡음의 변화가 감소되는 효과를 갖는다.

즉, 잡음이 제거된 i번째 프레임의 스펙트럼 magnitude $|\widehat{X}_i(w_k)|$는 다음과 같이 구할 수 있다.

$$|\widehat{X}_i(w_k)| = |\overline{Y}_i(w_k)| - \alpha_i |\widehat{D}_i(w_k)| \qquad (2)$$

여기서, $|\widehat{D}_i(w_k)|$는 추정된 잡음의 i 번째 프레임, k 번째 주파수의 스펙트럼 magnitude이고, α_i는 SNR에 따라 결정되는 subtraction factor 값이다.

MS 잡음 추정 방식에서는 오염된 입력 신호의 최근 D-frame 구간에 대해서 최소 전력크기를 가지는 프레임이 잡음만 포함하고 있는 프레임이라고 가정하기 때문에, D-frame 구간은 음성 신호가 비활성화 되는 묵음 구간을 포함할 만큼 충분히 큰 윈도우를 선택하여야 하며, 잡음의 통계적 특성이 변하지 않을 정도의 충분히 작은 윈도우를 선택하여야 한다. MS 잡음 추정 알고리즘의 세부적인 절차는 다음과 같다.

먼저 잡음으로 오염된 음성의 periodogram $|Y_j(w_k)|^2$을 구한다. 구해진 잡음으로 오염된 음성의 periodogram은 스무딩 과정을 통해서 급격히 변하는 문제점을 보완한다. 스무딩 factor $\alpha(j,w_k)$를 이용하여 스무딩된 파워 스펙트럼 $P(j,w_k)$를 구한다.

$$P(j,w_k)=\alpha(j,w_k)P(j-1,w_k)+(1-\alpha(j,w_k))|\overline{Y}_j(w_k)|^2 \quad (3)$$

그 후 D-frame 윈도우에서 최소 파워 스펙트럼 $P_{\min}(j,w_k)$을 구하게 된다. 이때 D값은 150으로 하였고, 최소 파워스펙트럼은 다음 식을 이용하여 구한다.

$$P_{\min}(j,w_k)=\min\{P(j,w_k),P(j-1,w_k),...,P(j-D-1,w_k)\} \quad (4)$$

위의 식을 이용해서 구한 최소 파워 스펙트럼은 bias 정정 요소로 보상 과정을 거쳐서 현재 프레임의 잡음을 추정하는데 쓰이게 된다.

III. Codebook Driven Short-Term Predictor Parameter Speech Enhancement (CDSTP)

최근 들어서 non-stationary 배경잡음 처리에 대해서도 많은 연구들이 진행되어 왔다. 그 중 한 가지로 CDSTP 알고리즘이 있다 [6]. CDSTP 알고리즘은 음성과 잡음의 스펙트럼 shape들을 LPC 계수로 표현하고, 대표적인 스펙트럼 shape들을 LPC 형태로 코드북에 지장한 후, 해당 코드북을 이용하여 잡음을 추정하는 방식이다. 잡음으로 오염된 음성신호는 음성과 잡음 신호로 분리될 수 있다고 가정하고, 음성과 잡음 신호 각각의 코드북 파라미터 추정에는 maximum-likelihood estimates 방식을 사용하였다 [7]. 또한 잡음과 음성의 스펙트럼 shape 이외에 gain을 계산하는 과정을 추가적으로 포함하고 있다.

CDSTP에서는 음성과 잡음 스펙트럼 shape 코드북에서 음성과 잡음 shape 코드벡터 후보를 선정하여 likelihood 값을 계산하게 되고, likelihood 값을 최소화 하는 코드북 엔트리와 gain을 추정하여 최종적으로 Wiener 필터의 계수로 사용하게 된다. CDSTP에서는 Itakura-Saito distortion을 이용하여 음성과 잡음에 대한 적절한 코드벡터를 찾는다 [8]. 프레임마다 이와 같은 과정의 반복을 통해서 파라미터들을 측정하게 된다.

CDSTP에서 사용하는 파라미터는 LPC 계수와 gain이다. 잡음과 음성 shape 코드벡터에 따른 gain 추정은 non-stationary 배경잡음 추정에 중요한 역할을 한다. 그림 1은 음성과 잡음의 shape 코드벡터를 통해 gain값을 구하고, Itakura-Saito distortion을 이용하여 최적의 파라미터를 찾는 과정을 나타낸다. 잡음으로 오염된 입력신호의 스펙트럼과 코드북에 저장되어 있는 정보들을 이용해 잡음과 음성의 shape과 gain을 측정한다.

음성과 잡음 코드북에서 각각 i 번째 음성 코드벡터 $\{a_x^i\}$와 j 번째 잡음 코드벡터 $\{a_w^j\}$를 선택하는 과정을 나타내면

$$i^*,j^*=\underset{i,j}{\arg\max}\ \max_{\sigma_x^2,\sigma_w^2} p_y(y|a_x^i,a_w^j;\sigma_x^2,\sigma_w^2) \quad (5)$$

와 같다. 여기서 σ_x^2과 σ_w^2는 음성과 잡음의 excitation variance를 나타낸다. (5) 식의 pdf $p_y(y|a_x^i,a_w^j;\sigma_x^2,\sigma_w^2)$가 Gaussian이라 가정하고 likelihood 값을 로그 도메인에서 표현하면

$$L=\int_0^{2\pi}-\frac{P_y(w)}{\frac{\sigma_x^2}{|A_x^i(w)|^2}+\frac{\sigma_w^2}{|A_w^j(w)|^2}}+\ln\left(\frac{1}{\frac{\sigma_x^2}{|A_x^i(w)|^2}+\frac{\sigma_w^2}{|A_w^j(w)|^2}}\right)dw \quad (6)$$

로 나타낼 수 있는데, $P_y(w)$는 입력신호의 스펙트럼, $A_x^i(w)$는 i번째 음성 코드벡터의 스펙트럼, $A_w^j(w)$는 j번째 잡음 코드벡터의 스펙트럼이다.

(5) 식과 (6) 식을 결합을 하게 되면,

$$i^*,j^*=\underset{i,j}{\arg\min}\left\{\min_{\sigma_x^2,\sigma_w^2} d_{IS}\left(P_y(w),\hat{P}_y(w)\right)\right\} \quad (7)$$

와 같이 나타낼 수 있다. 여기서 $\hat{P}_y(w)$는 음성과 잡음

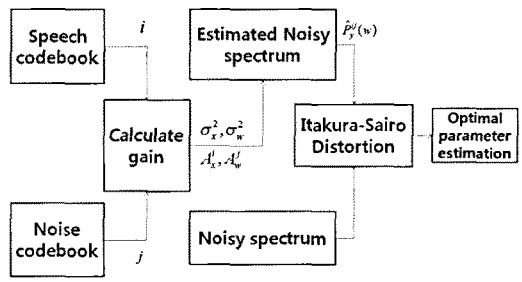

그림 1. CDSTP의 기본 블록도
Fig. 1. Basic block diagram of CDSTP.

코드벡터를 통해 표현한 스펙트럼으로 $\frac{\sigma_x^2}{|A_x^i(w)|^2}+\frac{\sigma_w^2}{|A_w^j(w)|^2}$와 같이 표현할 수 있다. $d_{IS}\left(P_y(w),\hat{P}_y(w)\right)$는 입력신호와 합성신호 스펙트럼의 Itakura-Saito distortion을 의미하며, 다음 식과 같이 표현할 수 있다.

$$d_{IS}\left(P_y(w),\hat{P}_y(w)\right)=\frac{1}{2\pi}\int_0^{2\pi}\left(\frac{P_y(w)}{\hat{P}_y(w)}-\ln\left(\frac{P_y(w)}{\hat{P}_y(w)}\right)-1\right)dw \quad (8)$$

Gain은 (8) 식을 최소화 하는 과정에서 구할 수 있으며

$$C\begin{bmatrix}\sigma_x^2\\ \sigma_w^2\end{bmatrix}=D \quad (9)$$

와 같은 식을 풀면 얻어진다. 이 때, C와 D는

$$C=\begin{bmatrix}\left\|\frac{1}{P_y^2(w)|A_x^i(w)|^4}\right\| & \left\|\frac{1}{P_y^2(w)|A_x^i(w)|^2|A_w^j(w)|^2}\right\|\\ \left\|\frac{1}{P_y^2(w)|A_x^i(w)|^2|A_w^j(w)|^2}\right\| & \left\|\frac{1}{P_y^2(w)|A_w^j(w)|^4}\right\|\end{bmatrix}$$

$$D=\begin{bmatrix}\left\|\frac{1}{P_y(w)|A_x^i(w)|^2}\right\|\\ \left\|\frac{1}{P_y(w)|A_w^j(w)|^2}\right\|\end{bmatrix} \quad (10)$$

로 표현할 수 있다 ($\|f(w)\|=\int|f(w)|dw$). 잡음과 음성의 SIP 측정에서 몇 가지 이론을 적용할 수 있는데 여기서는 파형을 강화하는데 초점을 맞춘다.

$[i^*,j^*,\sigma_x^2,\sigma_w^2]$와 같이 최적의 LPC와 gain이 선택이 되었다면, 다음과 같은 Wiener filter를 구현하여 잡음 제거에 적용할 수 있다 [9].

$$H(w)=\frac{\frac{\sigma_x^2}{|A_x^{i^*}(w)|^2}}{\frac{\sigma_x^2}{|A_x^{i^*}(w)|^2}+\frac{\sigma_w^2}{|A_w^{j^*}(w)|^2}} \quad (11)$$

IV. 제안하는 잡음제거 알고리즘

MS는 stationary 배경잡음에는 강인하지만, non-stationary 배경잡음에는 상대적으로 취약하다. CDSTP는 non-stationary 배경잡음에도 강인한 특성을 보이지만 코드북에 저장된 shape 이외의 잡음환경에는 취약하므로, 본 논문에서는 CDSTP와 MS를 결합하여 사용하고자 한다.

실생활 환경에서는 단일 잡음만을 듣게 되는 경우는 거의 없다. 대부분의 경우에 우리는 두 가지 혹은 세 가지 이상의 잡음에 노출되고, 이러한 잡음 환경에서 음성에 대한 여러 가지 기술들, 즉, 휴대기기를 이용한 통신 및 음악정보처리 (music information retrieval) 등을 수행해야 한다. 따라서 기존 알고리즘들이 고려했던 단일 잡음 제거에 대한 연구와 함께 mixed 잡음에 대한 처리도 고려하여 MS 와 CDSTP를 접목시킨 알고리즘을 구현 하였다. 본 논문에서는 그림 2에서 보듯이 잡음을 제거하는데 있어서 한가지의 잡음 추정 방법만을 사용 하지 않고, 두 가지의 잡음 추정 방법을 연결하여 사용하여 기존의 알고리즘보다 더욱 강인한 성능을 보이는 알고리즘을 제안한다. 제안한 알고리즘에서는 두 방법을 연결할 때 한 알고리즘을 먼저 사용하여 1차적인 향상된 음성을 얻고, 이 음성을 두 번째 알고리즘을 이용하여 최종적으로 향상된 음성을 얻도록 한다.

4.1. MS + MS

MS나 CDSTP 등의 잡음 추정 알고리즘을 사용하여 잡음을 제거할 경우 잡음으로 오염된 원음성에 비해서 향상된 음성을 얻을 수 있다. 하지만, 대부분의 알고리즘이 매 프레임 마다 잡음을 정확하게 추정 할 수는 없기 때문에 잔여 잡음이 남게 된다. 실제의 잡음보다 추정된 잡음

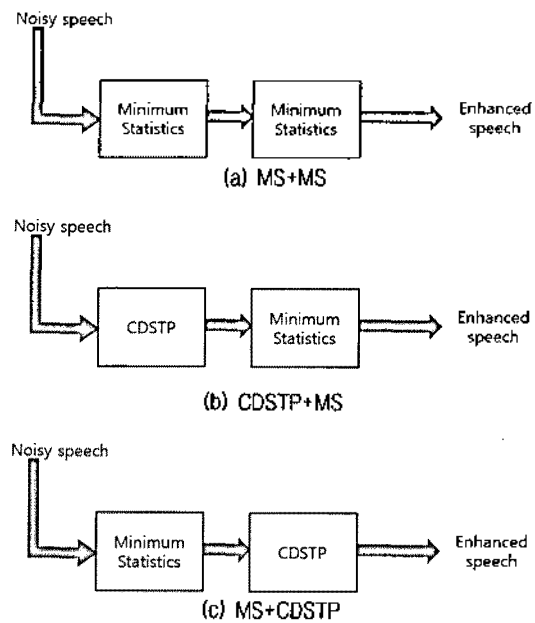

그림 2. 제안하는 알고리즘의 블록도.
Fig. 2. Block diagram of the proposed algorithms.

이 클 경우에는 잡음이 부분적으로 지나치게 제거되어서 musical noise가 생기게 되고, 실제의 잡음보다 추정된 잡음이 작을 경우에는 너무 적게 제거가 되어 잔여잡음(residual noise)이 남게 된다. 따라서 유럽 전기통신 표준협회 (European Telecommunications Standards Institute: ETSI)에서는 음성인식을 위한 음성향상 기법으로 다음과 같은 방법을 권고한다 [10]. 입력 신호에 대해서 우선적으로 Wiener filter를 이용하여 1차적인 잡음을 제거하도록 하고, 다시 Wiener filter를 재사용하여 잔여 잡음에 대한 처리를 하도록 한다.

본 논문에서는, 음질 향상 관점에서, ETSI에서 권고한 Wiener filter를 multi-stage로 사용하고자 한다. 이 때, Wiener filter의 noise estimation 방법으로 MS 알고리즘을 이용하였다. 앞서 설명 하였던 MS 알고리즘이 다양한 잡음 환경에 대해서 우수한 성능을 보임이 보고되었지만, 역시 잔여 잡음의 문제가 있으며, 따라서 그림 2 (a)와 같이 MS + MS 형태로 잔여잡음을 제거하고자 한다.

4.2. CDSTP + MS

MS + MS 알고리즘은 MS를 이용한 1차적인 잡음제거 후 남아 있는 잔여잡음에 대한 처리로 MS를 한 번 더 사용함으로써 기존의 MS를 한 번만 사용한 경우보다 강인한 효과를 기대하였다. 하지만, MS + MS 는 기대했던 효과와는 다르게 perceptual evaluation of speech quality (PESQ) [11] 관점의 개선 효과는 없었다. MS + MS의 경우에는 한번 제거한 잡음에 대하여 또 한 번 같은 방법으로 잡음을 재거를 하였기 때문에 성능 향상이 미미했다.

따라서 잡음 제거 알고리즘들을 접목 시킬 경우에는 특성이 서로 다른 잡음 추정 알고리즘을 결합하는 것이 유리하다고 생각할 수 있다. 본 논문에서는 CDSTP 알고리즘을 기준으로 두 가지 방법을 제안 하도록 한다. 먼저 첫 번째로 제안 하는 알고리즘은 그림 2 (b)와 같이 CDSTP와 MS를 결합한 알고리즘이다. CDSTP의 경우 코드북으로 해당 잡음의 특성을 training 하여 갖고 있기 때문에 MS 알고리즘에 비해 더 강인한 성능을 기대할 수 있다. 하지만, 코드북을 이용하여 잡음을 제거하였다 하더라도 잔여 잡음이 존재하기 때문에 이에 대해서는 다른 특성의 잡음 추정 알고리즘인 MS를 사용하도록 하였다. 즉, CDSTP + MS 방법을 사용하면 기존의 방법인 MS나 CDSTP의 단독 사용보다는 더 강인한 성능을 보일 것으로 기대할 수 있다.

4.3. MS + CDSTP

CDSTP + MS 방법은 코드북으로 1차적인 잡음추정을 하고 MS를 이용하여 잔여 잡음을 추정하도록 설계되었다. 본 방식은 테스트 음성의 잡음 환경이 CDSTP 코드북 학습 환경과 일치하는 경우에는 우수한 성능을 보일 것으로 기대할 수 있다. 하지만, 일반적인 경우에 테스트 환경은 학습 환경과 불일치가 발생하므로, MS + CDSTP 방식을 제안하였다. MS + CDSTP 방법은 우선적으로 MS를 이용하여 코드북으로 모델링할 수 없는 잡음을 제거하고, 남은 잡음을 CDSTP로 모델링하여 제거하고자 하였다. 따라서 CDSTP의 단점이라고 할 수 있는 outside 잡음에 대해서도 처리가 가능하다는 점이 MS + CDSTP 알고리즘의 강점이라 할 수 있다. 또한, CDSTP + MS나 MS + CDSTP 알고리즘은 서로 다른 두 가지 이상의 잡음이 섞여 있는 mixed 잡음 환경에서 다른 알고리즘보다 더욱 강인한 성능을 보일 것으로 기대된다.

V. 실험 및 결과

본 논문에서는 제안된 알고리즘의 성능을 평가하기 위해서 기존의 알고리즘들을 포함하여 총 6가지의 실험을 진행 하였다. 실험에 이용된 알고리즘은 다음과 같다.

1) MS
2) CDSTP
3) MS + MS
4) CDSTP + MS
5) MS + CDSTP
6) MS + CDSTPv2

기본적인 실험조건은 다음과 같다. 음성과 배경잡음의 sampling rate는 8 kHz, window는 hamming window를 사용했으며, 프레임의 길이는 20 ms로 설정하였다. CDSTP의 음성 코드북 트레이닝을 위해서는 TIMIT database를 이용하였다. TIMIT에서 168명이 녹음한 총 1680개의 음성을 가지고 (각 사람당 10문장씩 추출) 10차 LPC 계수를 추출한 후, line spectral frequencies (LSF) 계수로 변환하여 Generalized Lloyd Algorithm (GLA)을 이용하여 10 bit 음성 코드북을 생성하였다. CDSTP의 잡음 코드북 트레이닝을 위해서는 white, volvo, machine-gun, babble noise를 inside 잡음으로 사용하였다. 각 잡음의 스펙트로그램은 그림 3과 같으며, 각 잡음 코드북별 LPC 차수는 6, 6, 16, 10으로 하고 비트할당은 3, 3, 4, 2 bits로 하였다. LPC 차수와 비트 할당을 높일 경우

spectral envelop 표현은 보다 정확해지지만 복잡도가 증가하므로, 성능과 복잡도 관점에서 실험에 의한 최적치를 선정하였다.

알고리즘의 성능 평가를 위해서는 트레이닝 과정에서 사용되지 않은 음성과 잡음을 이용하였다. 테스트 음성 신호로는 TIMIT database에서 추출한 50명의 화자 (남자 24명, 여자 26명)에 대해서 각각 10문장씩 추출하여 총 500문장의 신호를 사용하였다. 그림 4는 학습 시 고려되지 않고 테스트에만 사용된 outside 잡음인 F16, polyphonic ringtone, machinegun+white 잡음의 스펙트로그램이다. 음성신호에 대하여 해당 잡음들을 0 dB, 10 dB, 20 dB로 섞어서 사용 하였다.

표 1은 각 알고리즘의 음질을 perceptual evaluation of speech quality (PESQ) 점수로 계산한 결과이다 [10].

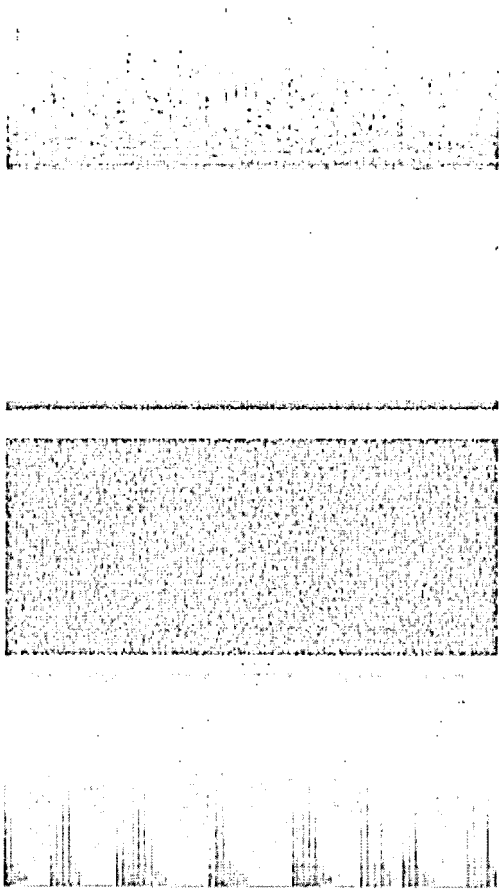

그림 3. Inside 잡음의 spectrogram: 위에서부터 차례대로 babble, volvo, white, machinegun noise (x축은 time, y축은 normalized frequency).

Fig. 3. Spectrogram for the inside noises: in order, babble, volvo, white, and machinegun noise (x and y axes denote time and normalized frequency, respectively).

또한 각 알고리즘에 대해 log spectral distortion (LSD)과 segmental speech-to-noise ratio (SSNR)값을 이용해 잡음 추정이 잘 되었는지를 확인 하였다. 표 2는 LSD를 이용한 결과 값이고, 표 3은 SSNR을 이용한 결과 값이다. LSD와 SSNR은 10dB의 volvo, babble 단일 잡음과 machinegun + white 의 혼합된 잡음을 사용하여 측정하였다.

5.1. MS vs MS+MS

MS 잡음 추정 알고리즘을 사용하여 잡음을 제거 하였을 경우 모든 잡음이 제거가 되지 않고 어느 정도의 잔여 잡음이 존재하는 것을 알 수 있다. 때문에 잔여 잡음을 처리하기 위해 MS를 두 번 사용하여 결과를 비교해 보았다. PESQ 관점에서 살펴보면, 표 1에서 보는 바와 같이 MS을 두 번 사용하였을 때는 MS를 한번만 사용하는 것에 비해 큰 개선효과가 없었다. 표 2와 3의 SD와 SSNR 결과를 보면, PESQ 관점에서 약간의 이득이 있었던 volvo 잡음에 대해서만 이득이 있고, 그 외의 잡음 환경에 대해서는 큰 개선 효과가 없음을 알 수 있었다.

5.2. MS vs CDSTP

MS는 대체적으로 stationary 배경잡음에 강인한 방법

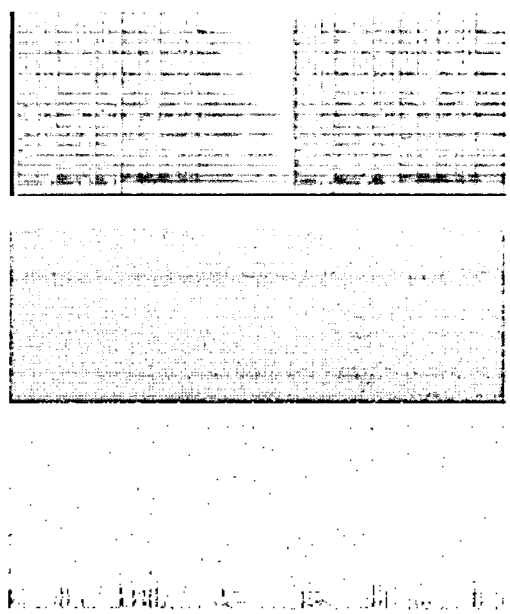

그림 4. Outside 잡음의 spectrogram: 위에서부터 ringtone, F16, machinegun+white noise (x축은 time, y축은 normalized frequency).

Fig. 4. Spectrogram for the outside noises: in order, ringtone, F16, and machinegun+white noise (x and y axes denote time and normalized frequency, respectively).

표 1. 기존 알고리즘과 제안한 알고리즘의 PESQ값 비교.
Table 1. PESQ scores of the conventional and proposed algorithms.

	Noise type (SNR)	MS	MS+MS	CDSTP	CDSTP+MS	MS+CDSTP	MS+CDSTPv2
Inside training noise	machinegun (0 dB)	2.43	2.41	2.76	2.72	2.44	2.44
	machinegun (10 dB)	3.06	3.04	3.32	3.30	3.06	3.08
	machinegun (20 dB)	3.56	3.55	3.72	3.70	3.56	3.60
	volvo (0 dB)	3.21	3.24	3.58	3.61	3.26	3.44
	volvo (10 dB)	3.82	3.84	4.03	4.05	3.82	4.00
	volvo (20 dB)	4.25	4.22	4.26	4.25	4.22	4.30
	white (0 dB)	1.70	1.70	1.76	1.77	1.85	1.86
	white (10 dB)	2.29	2.30	2.51	2.55	2.54	2.55
	white (20 dB)	2.93	2.94	3.15	3.2	3.19	3.19
	babble (0 dB)	1.91	1.90	1.88	1.88	1.94	1.92
	babble (10 dB)	2.57	2.57	2.61	2.63	2.67	2.67
	babble (20 dB)	3.21	3.21	3.27	3.30	3.32	3.33
Outside training noise	F16 (0 dB)	2.07	2.07	1.93	1.99	2.09	2.09
	F16 (10 dB)	2.70	2.70	2.66	2.72	2.73	2.74
	F16 (20 dB)	3.34	3.34	3.31	3.39	3.37	3.37
	ringtone (0 dB)	1.83	1.83	1.85	1.85	1.86	1.87
	ringtone (10 dB)	2.49	2.49	2.52	2.52	2.51	2.51
	ringtone (20 dB)	3.12	3.12	3.16	3.16	3.14	3.14
	machinegun + white (0 dB)	1.49	1.47	1.36	1.40	1.58	1.55
	machinegun + white (10 dB)	2.20	2.19	1.95	2.02	2.41	2.34
	machinegun + white (20 dB)	2.87	2.86	2.74	2.80	3.07	3.04

표 2. 각 알고리즘의 LSD 값 비교 (10 dB 잡음에 대한 결과)
Table 2. LSD scores of the conventional and proposed algorithms.

	log spectral distortion					
	MS	MS+MS	CDSTP	CDSTP+MS	MS+CDSTP	MS+CDSTPv2
volvo	1.36	1.07	0.96	0.98	1.17	0.90
babble	3.52	3.52	3.43	3.56	3.40	3.41
machinegun + white	6.61	6.65	7.39	7.48	5.05	5.55

표 3. 각 알고리즘의 SSNR 값 비교 (10dB 잡음에 대한 결과)
Table 3. SSNR scores of the conventional and proposed algorithms.

	segmental SNR					
	MS	MS+MS	CDSTP	CDSTP+MS	MS+CDSTP	MS+CDSTPv2
volvo	0.77	0.88	0.85	0.87	0.96	0.83
babble	5.17	5.18	4.95	4.93	5.06	5.01
machinegun + white	1.13	1.13	0.96	0.95	1.24	1.21

으로 알려져 있고, 잡음 신호에 대하여 별도의 학습 과정이 필요하지 않다. CDSTP는 stationary 배경잡음 뿐만 아니라 non-stationary 배경잡음에서도 강인한 성능을 보인다. CDSTP의 경우 잡음 코드북에 존재하는 잡음인 inside 잡음에 대해서는 대체로 MS 보다 좋은 성능을 보인다. 왜냐하면 제거하고자 하는 잡음에 대한 정보를 LP 스펙트럼 형태의 코드북으로 갖고 있기 때문에, 과거 D-frame만을 이용하여 잡음을 추정하는 MS에 비해 더 강인한 효과를 나타내게 된다. 특히, non-stationary 배경잡음인 machinegun 잡음에 대해서는 성능 향상 효과가 크게 나타난다.

반면에, 잡음 코드북에 존재하지 않는 outside 잡음 환경에 대해서는 CDSTP보다는 MS에서 보다 좋은 성능이 관찰된다. 특이할 점은 코드북 학습 시 고려하지 않은

ringtone 잡음의 경우 일부 프레임에서 보이는 LP 스펙트럼의 유사성 때문에 MS에 비해서 CDSTP의 성능이 더 좋게 관찰된다는 점이다. 즉, CDSTP를 실생활에서 적용할 경우 모든 종류의 잡음에 대해서 코드북을 학습할 필요는 없으며 대표적인 잡음군만을 선정하여 학습하더라도 MS 보다 우수한 성능을 기대할 수 있다. 그에 대한 결과는 표 1의 PESQ점수와 표 2의 LSD, 표 3의 SSNR의 값에서 확인 할 수 있다.

5.3. 제안하는 알고리즘

표 1,2,3에 보듯이 CDSTP만을 사용하여 음성 향상을 시키는 것 보다는 제안하는 알고리즘인 CDSTP의 앞 또는 뒤에 MS 알고리즘을 집목하여 잡음을 추정하였을 경우가 대체적으로 성능이 향상됨을 볼 수 있었다. MS를 접목시킴으로 CDSTP의약점인 outside 잡음에 대한 보완이 가능하기 때문이다. 이 때, inside 잡음에 대해서는 CDSTP를 먼저 수행하고 남은 산여 잡음을 MS로 처리하는 CDSTP+MS 방식의 성능이 우수하였고, outside 잡음에 대해서는 MS+CDSTP 방식의 성능이 우수하였다.

예외적으로 machinegun 잡음의 경우는 discrete 한 특성이 있기 때문에 MS와 같이 과거 D-frame이내의 window를 살펴서 잡음을 측정하는 방식으로는 제거가 불가능하고, MS의 접목 없이 CDSTP만을 사용 하는 경우가 가장 우수한 성능을 나타낸다.

또한, 제안 알고리즘은 mixed 잡음 환경인 machinegun+white 잡음 환경에서 다른 알고리즘보다 더욱 강인한 성능을 보이는 것을 알 수 있다. Mixed 잡음 환경에 대해서는 먼저 MS로 제거 가능한 잡음을 제거한 후 CDSTP를 이용하여 MS로는 제거하기 힘든 machinegun 잡음을 제거하도록 하기 때문에 기존의 알고리즘보다 더욱 강인한 성능을 볼 수 있다.

따라서 예측 가능한 잡음 환경에서는 CDSTP+MS 방식이, 그리고 mixed 잡음 환경을 포함한 예측이 힘든 다양한 잡음 환경을 고려하면 MS+CDSTP 방식이 가장 우수한 성능을 보인다고 할 수 있다.

MS+CDSTP 방식에서는 MS를 먼저 실행하여 CDSTP에 사용될 잡음 스펙트럼의 shape이 변하기 때문에 CDSTP 잡음 코드북 학습 시 MS로 처리한 잡음 신호를 이용하는 것이 바람직하다. 이 방식을 MS+CDSTPv2로 명명하였으며, 표 1에서 보는 바와 같이 MS+CDSTPv2 방법이 MS+CDSTP 방법에 비해서 inside 잡음에 대해서는 대체로 좋은 성능을 나타내며, outside 잡음에 대해서도 비슷한 성능을 나타냈다. Mixed 잡음에 대해서는 좋지 못한 성능을 보이는데 이것은 MS+CDSTPv2 방법이 machinegun과 white가 섞인 잡음을 고려해서 코드북 설계를 하지 않았기 때문이다. 이러한 결과는 표 2를 통해서도 볼 수 있다. 하지만 예외적으로 표 3의 SSNR 관점에서는 다른 두 결과와는 조금 다르게 MS+CDSTP가 MS+CDSTPv2에 비해 조금 더 좋은 성능을 가짐이 관찰되었다.

VI. 결론

본 논문에서는 mixed 잡음을 포함한 non-stationary 잡음 환경에 강인한 배경 잡음 추정 알고리즘으로 MS와 CDSTP를 결합하는 방식을 제안하였다. 잡음 코드북 학습시 고려한 잡음 환경에 대해서는 MS를 CDSTP의 후처리로 사용하는 경우가, 학습 시 고려하지 못한 잡음 환경에 대해서는 MS를 CDSTP의 전처리로 사용하는 경우가 각각의 알고리즘을 독립적으로 사용하는 경우보다 우수한 방식을 보임을 알 수 있었다. 동일한 잡음 추정 알고리즘인 MS를 두 번 연속적으로 사용하는 경우의 성능개선이 미미함에 비해, non-stationary 환경에 강인한 CDSTP와 별도의 학습 과정이 필요 없는 MS를 결합하는 제안 방식은 PESQ 관점에서 상당한 성능개선 효과를 보였다. 그리고 LSD와 SSNR의 값에서도 PESQ점수와 대체로 비슷한 결과를 볼 수 있었다. 특히, 본 논문에서 고찰한 mixed 잡음 환경에 대해서는 다른 기존 알고리즘보다 훨씬 강인한 성능 향상을 얻을 수 있었다. 향후에는 CDSTP 학습 시 필요한 잡음 군에 대한 연구를 진행할 예정이며, 제안 알고리즘을 음성코딩, 음성인식, 음악정보처리 등 다양한 분야에 적용해 볼 예정이다.

감사의글

이 논문은 2009년도 정부 (지식경제부)의 재원으로 정보통신 산업원천기술사업의 지원을 받아 수행된 연구임 (No. 2009-S-001-01).

참고문헌

1. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.
2. L. Singh, and S. Sridharan, "Speech enhancement using

critical band spectral subtraction," in *Proc. Intern. Conf. Spoken Lang. Processing*, pp. 2827–2830, 1998.
3. S. Kamath, and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, 2002.
4. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio processing*, vol. 9, no. 5, pp. 504–512, 2001.
5. 박윤식, 장준혁, "강인한 음성향상을 위한 Minimum Statistics와 Soft Decision의 확률적 결합의 새로운 잡음전력 추정기법," *한국음향학회지*, 26권, 4호, 153–158쪽, 2007.
6. S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Speech Audio processing*, vol. 14, issue 1, pp.163–176, 2006.
7. M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, col. 1, pp. 669–672, 2001
8. R. M. Gray, A. Buzo, A. H. G Jr. and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol 28, no. 4, pp. 367–376, 1980
9. T. Sreenicas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Processing*, col. 4, no. 5, pp. 383–389, 1996.
10. ETSI ES 202 050, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced ront-end feature extraction algorithm; Compression algorithms*, 2007.
11. A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, col. 2, pp. 749–752, 2001

저자 약력

•이 명 석 (Myeongseok Lee)

2008~2009: 세종대학교 정보통신공학과, 학사
2010~현재: 세종대학교 정보통신공학과, 석사과정

•노 명 훈 (Myunghoon Noh)

2003~2009: 세종대학교 정보통신공학과, 학사
2010~현재: 세종대학교 정보통신공학과, 석사과정

•박 성 주 (Sung-Joo Park)

1997.2 경북대학교 전자공학과, 석사
1997~1999 대우전자 영상연구소 연구원
2000~2002 디지털앤디지털 선임연구원
2002~2004 LSI Logic Korea 선임연구원
2004~현재 KETI 디지털미디어연구센터 선임 연구원

•이 석 필 (Seok-Pil Lee)

1990.2. 연세대학교 전기공학과, 공학사
1992.2. 연세대학교 대학원, 전기공학과, 공학석사
1997.2. 연세대학교 대학원, 전기전자공학과, 공학박사
1997~2002 대우전자 영상 연구소, 선임연구원
2002~현재 KETI 디지털미디어연구센터, 센터장

•김 무 영 (Moo Young Kim)

1989.3~1993.2 연세대학교 전자공학과, 학사
1993.3~1995.2 연세대학교 전자공학과, 석사
1995.2~2000.12 삼성종합기술원 전문연구원
2001.1~2004.11 Royal Institute of Technology (KTH, 스웨덴) Dept. Signals, Sensors, Systems, 박사
2004.12~2005.2 Royal Institute of Technology (KTH, 스웨덴) Dept. Signals, Sensors, Systems, PostDoc
2005.2~2006.8 Ericsson Research (스웨덴), Senior Research Engineer
2006.8~현재 세종대학교 정보통신공학과, 조교수
※ 관심분야: 음성/오디오/비디오 신호처리, 패턴인식, 정보이론.