논문 2007-44SP-6-14

Smoothed Global Soft Decision에 근거한 음성 향상 기법

(Speech Enhancement based on Smoothed Global Soft Decision)

조 규 행*, 박 윤 식*, 장 준 혁**

(Q-Haing Jo, Yun-Sik Park, and Joon-Hyuk Chang)

요 약

본 논문에서는 잡음 환경에서의 음성 향상을 위해 향상된 Global Soft Decision (GSD) 기법을 제안한다. 통계적 모델을 바탕으로 한 음성 향상과 관련한 연구에서 GSD는 음성의 꼬리 부분에서 취약하다고 알려져 있으며, 이를 개선하기 위해 Smoothed Global Likelihood Ratio (SGLR)를 바탕으로 한 새로운 음성 향상 기법을 GSD에 적용한다. 제안된 방법은 다양한 잡음 환경에서 MOS 실험을 바탕으로 기존의 연구와 비교하였으며 우수한 성능을 보여주었다.

Abstract

In this paper, we propose an improved global soft decision for speech enhancement in noise environments. From an examination of statistical model-based speech enhancement, it is shown that the global soft decision has a fundamental drawback at the offset region of speech signals. To overcome the drawback, we apply a new speech enhancement method based on a smoothed Global likelihood ratio to the global soft decision. Performances of the proposed method are evaluated by subjective tests under various environments and yield better results compared with the reported speech enhancement method.

Keywords: Speech Enhancement, Global Soft Decision, Smoothed Global Likelihood Ratio

I. 서 론

배경 잡음이 존재하는 경우 음성 인식과 부호화와 같은 시스템의 두드러진 성능 저하와 관련하여 최근 수년간 잡음 환경에서의 음성 향상과 관련된 관심이 증가하였으며[1~11], 또한 많은 알고리즘들이 연구되었다[12~20]. 기존의 알고리즘들은 일반적으로 깨끗한 음성과 잡음의 스펙트럼을 몇 개의 결합된 파라미터들을 이용한 비상관적 (uncorrelated) 통계 모델을 통해 특성화하였다. 하지만 음성 부재 구간에서 가정한 통계 모델은 음성 존재 구간에서의 통계적 모델과 상이하므로 스펙트럼의 추정은 음성의 부재와 존재를 고려하여 처리하여야 한다.

일반적으로, soft decision 이득 수정에 근거한 음성 향상 기법들이 각 프레임에 대해 음성의 존재 및 부재를 음성 검출기 (VAD, Voice Activity Detector)를 이용해 hard decision을 취하는 음성향상기법 보다 우수한 성능을 보이는 것으로 알려졌다[8, 16~17, 21]. 최근 각각의 스펙트럼 성분들을 독립적으로 다루는 대신 주어진 프레임을 전역적으로 수행하는 Global Soft Decision (GSD)을 기반으로 한 새로운 음성 향상 알고리즘이 제안되었다[12].

본 논문에서는 GSD의 성능 향상을 위해 스무딩된 전역 우도비 (SGLR, Smoothed Global Likelihood Ratio)를 기반으로 한 GSD를 제안한다. 전역 음성 부재 확률 (GSAP, Global Speech Absence Probability)을 계산하기 위한 강인한 방법을 고안하며, 스펙트럼 이득 수정과 기존의 제안된 음성 향상 기법에서 필요로 하는 단독의 VAD 알고리즘을 위한 잡음 스펙트럼 추정의

* 학생회원, ** 정회원, 인하대학교 전자전기공학부
(School of Electronic and Electrical Engineering, Inha University)

※ 본 연구는 정보통신부 및 정보통신연구진흥원의 IT 신성장동력핵심기술개발사업의 일환으로 수행하였음. [2005-S096-02, 신체장애인을 위한 착용형 단말 인터페이스 기술]

접수일자: 2007년4월2일, 수정완료일: 2007년10월24일

갱신에 적용한다. 제안된 알고리즘은 다양한 잡음 환경에서 Mean Opinion Score (MOS) 실험을 바탕으로 기존의 음성 향상 기법과 성능을 비교한다.

Ⅱ. Global soft decision의 이해

시간축 상에서 원래의 음성신호 $x(n)$에 잡음신호 $d(n)$이 부과된 입력신호 $y(n)$을 DFT (Discrete Fourier Transform)를 통해 주파수 축으로 변환하면 아래와 같이 표현된다.

$$Y_k(n)=X_k(n)+D_k(n) \quad \text{for} \quad k=1,\cdots,N \tag{1}$$

여기서 $Y_k(n)$, $X_k(n)$과 $D_k(n)$은 각각 $Y(n)$, $X(n)$과 $D(n)$의 k번째 스펙트럼 성분을 나타낸다. H_0, H_1이 각각 음성의 부재와 존재에 대한 가설이라고 하면 각 주파수 채널별로 다음과 같이 기술된다.

$$H_0\text{: speech absent} \ : \ Y(n)=D(n) \tag{2}$$

$$H_1\text{: speech present} : Y(n)=X(n)+D(n) \tag{3}$$

음성과 잡음신호의 스펙트럼이 zero-mean 복소 가우시안 분포의 특성을 가진다고 가정하면 주어진 가설 H_0와 H_1을 조건으로 한 확률 밀도 함수는 아래와 같이 주어진다.

$$p(Y_k(n)|H_0)=\frac{1}{\pi\lambda_{d,k}(n)}\exp\left\{-\frac{|Y_k(n)|^2}{\lambda_{d,k}(n)}\right\} \tag{4}$$

$$p(Y_k(n)|H_1)=\frac{1}{\pi[\lambda_{d,k}(n)+\lambda_{x,k}(n)]} \cdot \exp\left\{-\frac{|Y_k(n)|^2}{\lambda_{d,k}(n)+\lambda_{x,k}(n)}\right\} \quad \text{for} \quad k=1,\cdots,N \tag{5}$$

여기서 $\lambda_{x,k}(n)$와 $\lambda_{d,k}(n)$는 각각 k번째 주파수 채널별 음성과 잡음의 분산이며, 입력 신호 $Y(n)$의 음성 부재 확률 (SAP, Speech Absence Probability)은 아래와 같다.

$$p(H_0|Y(n))=\frac{p(Y(n)|H_0)p(H_0)}{p(Y(n)|H_0)p(H_0)+p(Y(n)|H_1)p(H_1)} \tag{6}$$

여기서 $p(H_0)\ (=1-p(H_1))$은 음성 부재에 대한 사전 확률이다. 각각의 주파수 채널별 성분이 통계적으로 독립이라는 가정으로부터, 식(6)은 아래와 같이 표현된다.

$$\begin{aligned} p(H_0|Y(n)) &= \frac{p(H_0)\prod_{k=1}^{N}p(Y_k(n)|H_0)}{p(H_0)\prod_{k=1}^{N}p(Y_k(n)|H_0)+p(H_1)\prod_{k=1}^{N}p(Y_k(n)|H_1)} \\ &= \frac{1}{1+q\Lambda_G(n)} \end{aligned} \tag{7}$$

여기서 $q=p(H_1)/p(H_0)$이며, $\Lambda_G(n)$는 전역 우도비 (GLR, Global Likelihood Ratio)를 나타내며 아래와 같이 각 주파수 채널별 우도비 (LR, Likelihood Ratio)의 곱으로 표현된다.

$$\begin{aligned} \Lambda_G(n) &= \prod_{k=1}^{N}\Lambda_k(Y_k(n)) \\ &= \prod_{k=1}^{N}\frac{p(Y_k(n)|H_1)}{p(Y_k(n)|H_0)} \end{aligned} \tag{8}$$

이 때, 식(4)와 (5)에서 가정한 확률 밀도 함수로 부터 각 주파수 채널별 LR은 아래의 식으로 유도된다.

$$\begin{aligned} \Lambda_k(Y_k(n)) &= \frac{p(Y_k(n)|H_1)}{p(Y_k(n)|H_0)} \\ &= \frac{1}{1+\xi_k(n)}\exp\left\{\frac{\gamma_k(n)\xi_k(n)}{1+\xi_k(n)}\right\} \end{aligned} \tag{9}$$

여기서 $\xi_k(n)$와 $\gamma_k(n)$는 각각 *a priori* signal-to-noise ratio (SNR)와 *a posteriori* SNR이며 다음과 같다[16].

$$\xi_k(n)=\frac{\lambda_{x,k}(n)}{\lambda_{d,k}(n)} \tag{10}$$

$$\gamma_k(n)=\frac{|Y_k(n)|^2}{\lambda_{d,k}(n)} \tag{11}$$

Ⅲ. Smoothed Likelihood Ratio를 적용한 향상된 Global Soft Decision

음성의 꼬리 부분에서 흔히 발생하는 검출 오류 문제점을 보완하기 위해 HMM을 기반으로 한 hang-over

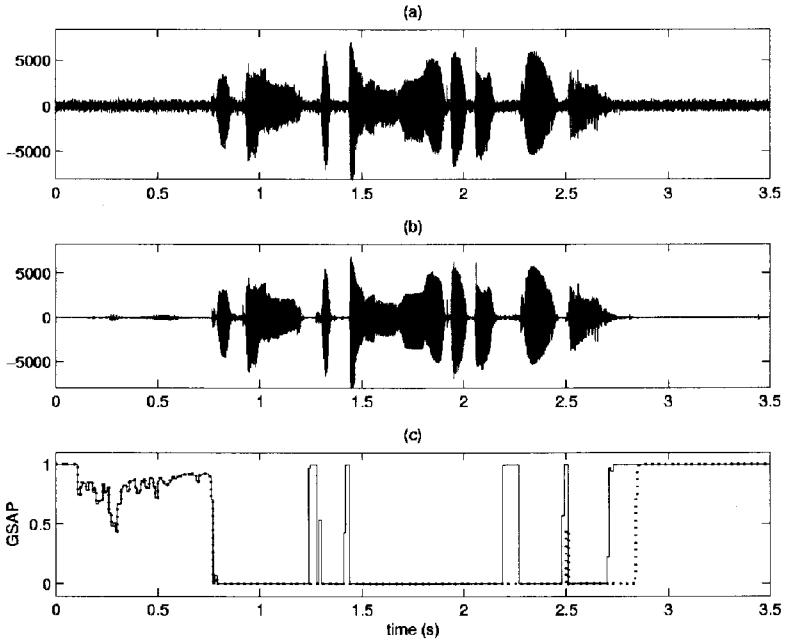


그림 1. GSAP의 예 (a) 잡음 섞인 음성 (b) 깨끗한 음성 (c) 기존의 GSAP (실선)와 SGLR을 이용한 GSAP (점섬)

Fig. 1. Examples of the GSAP (a) Noisy speech (b) Clean speech (c) conventional GSAP (solid line) and GSAP using the SGLR (dotted line)

같은 알고리즘들이 적용되고 있다. 최근에는 계산적으로 간단하면서도 효율적인 방법으로 알려진 강인한 음성 검출기의 파라미터인 LR이 음성의 offset 영역에서 *a priori* SNR의 지연으로 인해 급격히 변화하게 되므로 이를 보완하기 위한 스무딩된 우도비 (SLR, Smoothed Likelihood Ratio)가 소개되었다[9]. 비슷한 이유로 본 논문에서는 GSAP의 성능을 향상시키기 위해 SGLR을 제안한다.

먼저, 식(7)에 적용된 GSD 기법의 성능은 $\{\lambda_{x,k}\}$와 $\{\lambda_{d,k}\}$의 추정치의 신뢰성에 크게 의존한다. $D(n)$과 $X(n)$이 정상 상태 (stationarity)라는 가정 아래 음성 구간에서의 잡음 전력 갱신을 고려하여 배경 잡음과 음성 각각의 long-term 스무딩된 전력 스펙트럼을 사용하는데 이때 사용되는 잡음과 음성의 분산에 대한 추정치는 아래와 같이 표현된다.

$$\hat{\lambda}_{d,k}(n+1) = \zeta_d \hat{\lambda}_{d,k}(n) + (1-\zeta_d) E\left[|D_k(n)|^2 \mid Y_k(n)\right] \tag{12}$$

$$\hat{\lambda}_{x,k}(n+1) = \zeta_x \hat{\lambda}_{x,k}(n) + (1-\zeta_x) E\left[|X_k(n)|^2 \mid Y_k(n)\right] \tag{13}$$

여기서 $\hat{\lambda}_{d,k}(n)$와 $\hat{\lambda}_{x,k}(n)$가 각각 $\lambda_{d,k}$와 $\lambda_{x,k}$의 추정치이고, $\zeta_d(=0.99)$ 와 $\zeta_x(=0.97)$는 정상 상태의 가정을 고려한 스무딩 파라미터이다. 식(12)와 (13)에서 $X(n)$과 $D(n)$에 대한 통계적 가정을 바탕으로 아래의 식을 얻는다.

$$E\left[|D_k(n)|^2 \mid Y_k(n)\right] = E\left[|D_k(n)|^2 \mid Y_k(n), H_0\right] p(H_0 \mid Y_k(n)) + E\left[|D_k(n)|^2 \mid Y_k(n), H_1\right] p(H_1 \mid Y_k(n)) \tag{14}$$

$$E\left[|X_k(n)|^2 \mid Y_k(n)\right] = E\left[|X_k(n)|^2 \mid Y_k(n), H_0\right] p(H_0 \mid Y_k(n)) + E\left[|X_k(n)|^2 \mid Y_k(n), H_1\right] p(H_1 \mid Y_k(n)) \tag{15}$$

여기서

$$E\left[|D_k(n)|^2 \mid Y_k(t), H_0\right] = |Y_k(n)|^2 \tag{16}$$

$$E\left[|D_k(n)|^2 \mid Y_k(n), H_1\right] = \left(\frac{\hat{\xi}_k(n)}{1+\hat{\xi}_k(n)}\right) \hat{\lambda}_{d,k}(n) + \left(\frac{1}{1+\hat{\xi}_k(n)}\right)^2 |Y_k(n)|^2 \tag{17}$$

$$E\left[|X_k(n)|^2 | Y_k(t), H_0\right] = 0 \quad (18)$$

$$E\left[|X_k(n)|^2 | Y_k(n), H_1\right] = \left(\frac{1}{1+\hat{\xi}_k(n)}\right)\hat{\lambda}_{x,k}(n) + \left(\frac{\hat{\xi}_k(n)}{1+\hat{\xi}_k(n)}\right)^2 |Y_k(n)|^2 \quad (19)$$

여기서

$$\hat{\xi}_k(n) = \frac{\hat{\lambda}_{x,k}(n)}{\hat{\lambda}_{d,k}(n)}. \quad (20)$$

식(12)와 (13)으로부터 $\hat{\lambda}_{d,k}(n)$와 $\hat{\lambda}_{x,k}(n)$은 현재의 음성 신호 $Y_k(n)$에 의존하지 않으며 관련 파라미터들에 의한 이전 프레임으로부터 유추된 일종의 예측된 추정치를 의미한다. 실제로 예측 추정치가 GSAP를 추정할 때 *a priori* SNR보다 더 정확하다고 알려져 있지만 *a priori* SNR은 이전 프레임에서 유도되기 때문에 음성의 offset 영역에서 낮은 GLR로 인해 식(7), (9), (20)에 의해 주어진 GSAP는 매우 큰 값을 빈번하게 나타내게 된다. 이런 이유로 음성의 꼬리 부분에 대한 연속성을 강조하기 위해 SGLR을 위한 다음과 같은 식을 고려한다.

$$\tilde{\Lambda}_G(n) = \kappa\tilde{\Lambda}_G(n-1) + (1-\kappa)\prod_{k=1}^{N}\Lambda_k(Y_k(n)) \quad (21)$$

여기서 $\kappa(=0.9)$는 실험적으로 최적화된 값으로써 long-term 스무딩 파라미터이다. 그림 1에서 제안된 SGLR이 음성의 offset 영역에서 GSAP의 급격한 변화를 지연시킴으로서 상대적으로 강인한 성능을 보여주며, onset영역에서는 큰 변화를 보이지 않음을 알 수 있다.

$\hat{X}(n) = [\hat{X}_1(n), \hat{X}_2(n), \cdots, \hat{X}_N(n)]$을 n번째 프레임에서의 깨끗한 음성의 추정치라고 할 때, 기존의 스펙트럼 향상 기법은 오염된 음성 신호 $Y(n)$의 각각의 주파수 성분에 특정 이득을 적용하여 $\hat{X}(n)$을 추정한다. 본 논문에서는 스펙트럼 이득을 계산하기 위한 여러 가지 방법들 중 뮤지컬 잡음을 제거하는데 우수한 성능을 보이는 Ephraim과 Malah의 방법을 채택한다 [16, 22].

Ⅳ. 실 험

본 논문에서 제안된 SGLR의 음성 향상 알고리즘을 검증하기 위해 다양한 잡음 환경에서 주관적 음질 실험을 수행하였다. 남성과 여성 화자가 각각 5개씩 발음한 총 10개의 문장이 실험에 사용되었다. 잡음 환경을 만들기 위해 깨끗한 음성에 NOISEX-92 데이터베이스 중 white, babble, buccaneer의 세 종류의 잡음이 다양한 SNR로 부과되었다. 평가를 위해 기존의 [12]에서 적용된 GSAP 계산 모듈의 GLR을 SGLR으로 변환하여 적용하였으며, MOS 결과는 10명의 청자에 의해 평가되어진 점수를 평균하여 최종적으로 구하였다.

표 1은 다양한 잡음 환경에서의 MOS 결과를 보여준다. 결과로부터 제안된 SGLR이 대부분의 잡음 환경에서 기존의 SEGSD 음성 향상 알고리즘 보다 우수한 것을 확인할 수 있다[12].

표 1. 제안된 알고리즘과 SEGSD 기법을 적용한 음성 향상 알고리즘의 MOS 결과

Table 1. MOS results for the proposed enhancement algorithm (SGLR) and conventional SEGSD tech- nique

Noise	SNR (dB)	MOS results		
		None	SEGSD	SGLR
White	5	1.24	2.50	2.71
	10	1.56	3.20	3.32
	15	2.14	3.61	3.74
Babble	5	2.11	2.90	3.09
	10	2.25	3.44	3.63
	15	2.34	3.61	3.82
Buccaneer	5	1.30	2.50	2.65
	10	1.80	3.10	3.31
	15	2.01	3.59	3.72

Ⅴ. 결 론

본 논문에서는 SGLR 기법을 적용한 새로운 스펙트럼 향상 알고리즘을 제안하였다. 음성 변이 구간에서의 GSAP의 추정치의 성능 향상을 위해 간단하지만 매우 효율적인 GLR의 강인한 추정 방법을 제시하였다. 제안된 방법의 수행은 MOS 실험을 통해 기존의 음성 향상 기법보다 우수함을 알 수 있었다.

참 고 문 헌

[1] J.-H. Chang and N. S. Kim, "Voice activity

detection based on complex Laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632-634, Apr. 2003.

[2] J.-H. Chang, N. S. Kim and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 1965-1976, June 2006.

[3] J.-H. Chang and N. S. Kim, "A new structural approach in system identification with generalized analysis-by-synthesis for Robust Speech Coding," *IEEE Trans. Speech and Audio Processing*, vol. 14, no. 3, pp. 747-751, May 2006.

[4] J.-H. Chang, "Perceptual weighting filter for robust speech modification," *Signal Processing*, vol. 86, Issue 5, pp. 1089-1093, May 2006.

[5] J.-H. Chang, N. S. Kim and S. K. Mitra, "A statistical model-based V/UV decision under background noise environments," *IEICE Trans. on Info. and Systs.*, vol. E87-D, no. 12, pp. 2885-2887, Dec. 2004.

[6] J.-H. Chang, J. W. Shin and N. S. Kim, "Voice activity detection employing generalized Gaussian distribution," *Electronics Letters*, vol. 40, no. 24, pp.1561-1562, Nov. 2004.

[7] J.-H. Chang and N. S. Kim, "Distorted speech rejection for automatic speech recognition in wireless communication," *IEICE Trans. Info. and Systs.*, vol. E87-D, no. 7, pp. 1978-1981, July 2005.

[8] 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, Jan. 1999.

[9] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276-278, Oct. 2001.

[10] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC Residual domain," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 217-231, Mar. 2001.

[11] TIA/EIA/IS-127, "Enhanced variable rate codec, speech service option 3 for wideband spectrum digital systems," 1996.

[12] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision", *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, May 2000.

[13] J.-H. Chang and N. S. Kim, "Speech enhancement : new approaches to soft decision," *IEICE Trans. Inf. and Syst.*, vol. 27, E84-D, pp. 1231-1240, Sep. 2001.

[14] J.-H. Chang, "Warped discrete cosine transform-based noisy speech enhancement," *IEEE Trans. Circuit and Systems II*, vol. 52, issue 9, pp. 535-539, Sept. 2005.

[15] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selectied Areas in Communications*, vol. 16, no. 9, pp. 1818-1829, Dec. 1998.

[16] 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. 32, no. 6, pp. 1109-1121, Dec. 1984.

[17] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 365-368, 1998.

[18] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol 81, pp. 2403-2418, Nov. 2001.

[19] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12-15, Jan. 2002.

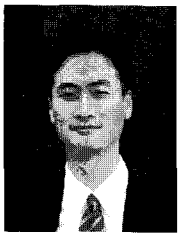
[20] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113-116, Apr. 2002.

[21] R. J. McAulary and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.28, pp. 137-145, Apr. 1980.

[22] O. Cappe, "Elimination of musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345-349, Apr. 1994.

저 자 소 개

조 규 행(학생회원)
2004년 인하대학교 전자공학과 학사 졸업
2004년~2006년 LG.Philips LCD 연구원
2006년~현재 인하대학교 전자공학과 석사과정
<주관심분야 : 음성검출, 잡음제거>



장 준 혁(정회원)
1998년 경북대학교 전자공학과 학사 졸업
2000년 서울대학교 전기공학부 석사 졸업
2004년 서울대학교 전기컴퓨터공학부 박사 졸업
2000년~2005년 ㈜넷더스 연구소장
2004년~2005년 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원
2005년~2005년 한국과학기술연구원 (KIST) 연구원
2005년~현재 인하대학교 전자전기공학부 조교수
<주관심분야 : 음성/오디오 신호처리, 통신 신호처리, 휴먼/컴퓨터 인터페이스 등>

박 윤 식(학생회원)
2006년 인하대학교 전자공학과 학사 졸업
2006년~현재 인하대학교 전자공학과 대학원 석사과정
<주관심분야 : 잡음제거, 음향학적 반향제거>