음성향상을 위한 2차 조건 사후 최대 확률기법 기반 Global Soft Decision

준회원 금 종 모*, 종신회원 장 준 혁**

Improved Global-Soft Decision Incorporating Second-Order Conditional MAP for Speech Enhancement

Jong-Mo Kum* *Associate Member*, Joon-Hyuk Chang** *Lifelong Member*

요 약

본 논문에서는 기존의 global soft decision 방법에서 음성부재확률의 고정 파라미터에 2차 조건 사후 최대 확률기법을 적용한 음성 향상 기법을 제안한다. 기존의 global soft decision 방법은 음성부재확률을 구하기 위해 가정한 가설에 따라 파라미터값을 고정하여 다양한 음성 환경 변화에 민감한 점을 고려하여 본 논문에서 제안한 알고리즘은 기존의 고정 파라미터 값에 직전 2 프레임에서의 음성 존재와 부재에 대한 조건을 부여해주어 음성과 음성사이의 상호 연관성을 고려해주고, 보다 유동적으로 현재 프레임의 음성부재확률을 추정하는 음성향상 기법이다. 제안된 방법의 성능평가를 위해 ITU-T P.862 perceptual evaluation of speech quality (PESQ)를 이용하여 평가하였고, 그 결과 제안된 2차 조건 사후 최대 확률기법을 적용한 global soft decision 방법은 기존의 Global soft decision 방법보다 향상된 결과를 나타내었다.

Key Words : Speech Enhancement, Global Soft Decision, Second-order Conditional Maximum *a posteriori* (Second-order CMAP)

ABSTRACT

In this paper, we propose a novel method to improve the performance of the global soft decision which is based on the second-order conditional maximum *a posteriori* (CMAP). Conventional global soft decision scheme has an disadvantage in that the speech absence probability adjusted by a fixed-parameter was sensitive to the various noise environments. In proposed approach using the second-order CMAP, speech absence probability value is more flexible which exploit not only the current observation but also the speech activity decisions in the previous two frames. Experimental results show that the proposed improved global soft decision method based on second-order conditional MAP yields better results compared to the conventional global soft decision technique with the performance criteria of the ITU-T P. 862 perceptual evaluation of speech quality (PESQ).

Ⅰ. 서 론

최근 이동통신 단말기나 차량 네비게이션 등 실제적인 음성신호처리 시스템이 필요한 환경이 늘어나면서 음성향상 기술에 대한 연구가 주목받고 있다. 실제 음성 향상 과정에서 잡음을 정확하게 추정하는

※ 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2008-C1090-0804-0007)
또한 본 연구는 지식경제부와 한국산업기술재단의 전략기술인력양성사업으로 수행된 연구결과임
* 인하대학교 전자공학과 DSP연구실(jmkum@dsp.inha.ac.kr),
** 인하대학교 전자공학과 조교수(changjh@inha.ac.kr)
논문번호 : KICS2009-03-086, 접수일자 : 2009년 3월 3일, 최종논문접수일자 : 2009년 5월 22일

것이 가장 중요한 요소이며, 특히 비상관 잡음신호를 처리할 수 있어야 한다. 실제로 많은 다양한 방법들이 음성 스펙트럼 향상을 위한 노력의 일환으로 시도되었다. 이러한 방법들 중에는 스펙트럼 차감법[1],[2], Wiener filtering[3], soft decision 추정[4], 최소 평균 자승 오차 (MMSE, Minimum Mean Square Error)[5]등이 주로 사용되고 있다. 이러한 방법들은 구현상의 이점과 다양한 배경잡음에 적용 가능한 장점을 지니고 으며, 특히 soft decision에 근거한 추정방법이 뛰어난 성능을 가진다는 것이 알려져 있다. 최근에 제안된 Global soft decision 방법에서는 기존의 채널별 음성부재확률 (LSAP, local speech absence probability)과 현재 프레임에서의 모든 데이터에 의해 결정 되어지는 전역 음성부재확률(GSAP, global speech absence probability)이 결합되어 새로이 통계적으로 견실한 음성부재확률을 도출하였다[6]. 하지만 음성부재확률을 구하기 위해 기존의 통계적 가정을 바탕으로 넣어준 고정 파라미터값을 취한 것은 다양한 음성 환경 변화에서 한계를 나타낸다.

본 논문에서는 음성과 음성사이의 강력한 상호 연관성이 있는 점을 고려하여 연구되어진 [7]을 기본으로 2차 조건 MAP (maximum a posteriori)를 사용하여 음성부재확률을 구하기 위해 사용된 고정 파라미터값 대신 직전 2 프레임에서의 음성의 존재, 부재의 조건을 부여해주는 유동적인 파라미터값을 사용하여[8] 주어진 환경에 따라 변화함은 물론 음성과 음성사이의 상호 연관성을 고려한 향상된 Global soft decision 기법을 제시한다. 제안된 음성 향상 기법은 ITU-T P.862 perceptual evaluation of speech quality (PESQ)[9]를 통해 평가 하였고 기존의 Global soft decision방법보다 향상된 결과를 나타내었다.

Ⅱ. Global Soft Decision 개요

먼저 오염된 음성신호 $y(t)$는 원래의 음성신호 $x(t)$에 잡음신호 $n(t)$가 더해져서 만들어졌다고 가정한다. 여기서 t는 이산시간을 나타낸다. 음성 향상 기법에서 사용되고 있는 기본가설 $H_0(k,l)$, $H_1(k,l)$이 각각 음성의 부재와 존재를 나타낸다고 하면 다음과 같이 표현된다.

$$\begin{aligned} H_0(k,l):\ & Y(k,l) = N(k,l) \\ H_1(k,l):\ & Y(k,l) = X(k,l) + N(k,l) \end{aligned} \tag{1}$$

여기서 $Y(k,l)$, $X(k,l)$ 그리고 $D(k,l)$은 각각 오염된 음성신호, 원래 음성 신호 그리고 잡음 신호의 퓨리에 변환 계수를 나타내고 l번째 프레임에서의 k번째 주파수 성분이 된다.

음성신호와 잡음의 스펙트럼이 복소가우시안 분포를 따른다는 가정으로부터 가설 $H_0(k,l)$, $H_1(k,l)$에 근거한 확률밀도함수는 다음과 같이 주어진다.

$$\begin{aligned} P(Y(k,l)|H_0(k,l)) &= \frac{1}{\pi\lambda_n(k,l)} exp - \frac{|Y(k,l)^2|}{\lambda_n(k,l)} \\ P(Y(k,l)|H_1(k,l)) &= \frac{1}{\pi(\lambda_x(k,l)+\lambda_n(k,l))} \\ &\quad \cdot \exp\left\{-\frac{|Y(k,l)^2|}{(\lambda_x(k,l)+\lambda_n(k,l))}\right\} \end{aligned} \tag{2}$$

위에서 $\lambda_x(k,l)$, $\lambda_n(k,l)$는 각각 음성과 잡음의 분산을 나타낸다.

음성의 존재와 부재에 관한 가설을 바탕으로 우선 주파수 채널별 음성부재확률은 다음과 같이 구해질 수 있다.

$$\begin{aligned} P(H_0(k,l)|Y(k,l)) &= \frac{P(Y(k,l)|H_0(k,l))P(H_0(k,l))}{P(Y(k,l))} \\ &= \frac{P(Y(k,l)|H_0(k,l))P(H_0(k,l))}{P(Y(k,l)|H_0(k,l))P(H_0(k,l)) + P(Y(k,l)|H_1(k,l))P(H_1(k,l))} \\ &= \frac{1}{1+\frac{P(H_1(k,l))}{P(H_0(k,l))}\Lambda(Y(k,l))} \end{aligned} \tag{3}$$

또한 한 프레임에서의 음성부재확률은 현재프레임의 관찰결과를 기반으로 다음과 같이 구할 수 있다.

$$\begin{aligned} P(H_0|Y(l)) &= \frac{P(Y(l)|H_0)P(H_0)}{P(Y(l))} \\ &= \frac{P(Y(l)|H_0)P(H_0)}{P(Y(l)|H_0)P(H_0)+P(Y(l)|H_1)P(H_1)}. \end{aligned} \tag{4}$$

각 주파수 성분들의 통계적인 독립성을 가정하면 한 프레임에서의 음성 부재 확률을 다음과 같이 표현할 수 있다.

$$P(H_0|Y(l)) = \frac{1}{1+\frac{P(H_1)}{P(H_0)}\prod_{k=1}^{M}\Lambda(Y(k,l))} \tag{5}$$

여기서 $P(H_0)$, $P(H_1)$은 음성 부재와 존재에 대한 a priori 확률값이 되고 $\Lambda(Y(k,l))$는 k번째 주

파수 채널에서의 우도비 (likelihood ratio) 로서 다음과 같이 나타낼 수 있다.

$$\Lambda(Y(k,l)) = \frac{P(Y(k,l)|H_1)}{P(Y(k,l)|H_0)} = \frac{1}{1+\xi(k,l)} exp\left[\frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}\right] \quad (6)$$

여기서 $\xi(k,l) \equiv \frac{\lambda_x(k,l)}{\lambda_n(k,l)}$, $\gamma(k,l) \equiv \frac{|Y(k,l)|^2}{\lambda_n(k,l)}$ 이 되고 $\xi(k,l)$, $\gamma(k,l)$는 각각 *a priori* SNR과 *a posteriori* SNR을 나타낸다[6].

Ⅲ. 2차 조건 MAP (maximum *a posteriori*)를 이용한 향상된 Global Soft Decision

지금까지 우리는 Global soft decision 방법에서의 음성부재 확률을 구하는 방법에 대해 알아보았다. 하지만 기존의 Global soft decision방법에서는 고정된 파라미터값 q(=$P(H_1)/P(H_0)$)를 사용하였기 때문에 수시로 변하는 잡음환경에서 정확한 음성 부재 확률을 추정하지 못하였다. 하지만 직전 2프레임의 음성 존재와 부재에 관한 조건을 부여해 주면서 음성과 음성사이의 상호 연관성까지 고려해주는 2차 조건 MAP를 이용한 향상된 Global soft decision을 제안한다.

음성 활동에서 인접한 프레임들의 상호 연관성을 고려하여 히든 마르코프 모델 (Hidden Markov Model, HMM)을 이용한 행오버를 사용함으로서 통계모델을 기반으로 한 VAD의 에러를 효과적으로 줄일 수 있다[7]. 즉, 음성 활동에서 프레임들간의 강력한 상호 연관성에 기반 하여 이전 2 프레임의 조건이 추가된 음성 부재 확률을 식 (7)과 같이 표현할 수 있다.

여기서, $\alpha = P(Y(k,l)|H(k,l)=H_0, H(k,l-1)=H_i, H(k,l-2)=H_j)P(H(k,l)=H_0|H(k,l-1)=H_i, H(k,l-2)=H_j)$, $\beta = P(Y(k,l)|H(k,l)=H_1, H(k,l-1)=H_i, H(k,l-2)=H_j)P(H(k,l)=H_1|H(k,l-1)=H_i, H(k,l-2)=H_j)$ 이다. 제안된 방법에서는 Global soft decision의 q 값 대신 $\hat{q} = (P(H(k,l)=H_1|H(k,l-1)=H_i, H(k,l-1)=H_j)/P(H(k,l)=H_0|H(k,l-1)=H_i, H(k,l-2)=H_j)$로 대체되어진다. 이것은 다음의 식처럼 음성과 음성 사이의 상호 연관성을 고려해 신뢰성을 높여준다.

$$P(H(k,l)=H_1|H(k,l-1)=H_1, H(k,l-2)=H_1) > P(H(k,l)=H_1) \quad (8)$$

이를 바탕으로 위의 제안된 식은 다음과 같이 표현할 수 있다.

$$P(H(k,l)=H_0|Y(k,l), H(k,l-1)=H_i, H(k,l-2)=H_j) = \frac{1}{1+\hat{q}\Lambda(Y(k,l))}, \quad i=0,\ 1,\ j=0,\ 1 \quad (9)$$

$\hat{q}$의 값은 직전 2프레임의 영향을 받아 다음과 같이 4가지의 값을 가지게 된다.

$$\begin{aligned}
\widehat{q_{00}} &(= \frac{P(H(k,l)=H_1|H(k,l-1)=H_0, H(k,l-2)=H_0)}{P(H(k,l)=H_0|H(k,l-1)=H_0, H(k,l-2)=H_0)}) \\
\widehat{q_{01}} &(= \frac{P(H(k,l)=H_1|H(k,l-1)=H_0, H(k,l-2)=H_1)}{P(H(k,l)=H_0|H(k,l-1)=H_0, H(k,l-2)=H1)}) \\
\widehat{q_{10}} &(= \frac{P(H(k,l)=H_1|H(k,l-1)=H_1, H(k,l-2)=H_0)}{P(H(k,l)=H_0|H(k,l-1)=H_1, H(k,l-2)=H_0)}) \\
\widehat{q_{11}} &(= \frac{P(H(k,l)=H_1|H(k,l-1)=H_1, H(k,l-2)=H_1)}{P(H(k,l)=H_0|H(k,l-1)=H_1, H(k,l-2)=H_1)})
\end{aligned} \quad (10)$$

$\widehat{q_{00}}$는 이전 프레임에 음성이 존재하지 않고 그 이전 프레임에도 음성이 존재하지 않을 때 이며, $\widehat{q_{01}}$은 이전 프레임에 음성이 존재하지 않고 그 이전 프레임에 음성이 존재할 때 이다. 또한 $\widehat{q_{10}}$은 이전 프레임에는 음성이 존재하고 그 이전 프레임에는 음성이 존재하지 않을 때 이며, $\widehat{q_{11}}$은 이전 프레임과 그 이전 프레임에 모두 음성이 존재 할 때 이다. 이렇게 함으로써 이전 2 프레임의 정보가 음성신호일 확률이 높을 때에는 음성 부재 확률값을 더 작게 만들어주고 이전 2 프레임의 정보가 잡음신호일

$$\begin{aligned}
&P(H(k,l)=H_0|Y(k,l), H(k,l-1)=H_i, H(k,l-2)=H_j) \\
&= \frac{P(Y(k,l)|H(k,l)=H_0, H(k,l-1)=H_i, H(k,l-2)=H_j)P(H(k,l)=H_0|H(k,l-1)=H_i, H(k,l-2)=H_j)}{P(Y(k,l))} \\
&= \frac{P(Y(k,l)|H)k,l)=H_0, H(k,l-1)=H_i, H(k,l-2)=H_j)P(H(k,l)=H_0|H(k,l-1)=H_i, H(k,l-2)=H_j)}{\alpha+\beta}, \ \text{i=0, 1, j=0, 1}
\end{aligned} \quad (7)$$

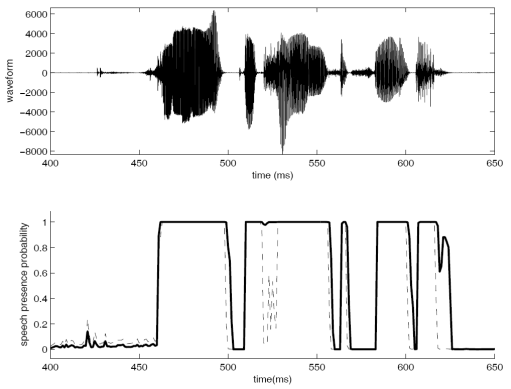

그림 1. F16 잡음 (SNR = 10 dB) 에서의 확률 비교 (a) 깨끗한 음성 파형 (b) 실시간 프레임에서의 음성 존재 확률: 기존의 Global soft decision의 확률 (점선), 제안된 알고리즘의 확률 (굵은선)

확률이 높을 때에는 음성 부재 확률값을 1에 가깝게 만들어준다.

이전의 고정된 파라미터 q값을 사용하던 Global soft decision보다 제안된 2차 조건 MAP를 이용한 방법이 음성부재확률을 구할 때 보다나은 성능을 보임을 그림 1에서 확인할 수 있다.

Ⅳ. 실험 결과

본 논문에서 제안한 알고리즘은 직전 2 프레임에서의 음성 존재와 부재에 대한 조건을 부여해주어 음성과 음성사이의 상호 연관성을 고려해주고, 보다 유동적으로 현재 프레임의 음성부재확률을 추정하는 음성 향상 기법이다. 제안된 음성 향상 알고리즘의 음질 평가를 위해 널리 적용되고 있는 ITU-T P.862 PESQ방법으로 음성 향상의 성능 비교를 하였다[9], [10], [11].

표 1의 ITU-T P.862 perceptual evaluation of speech quality (PESQ)[9] 테스트를 위해 남성, 여성 화자 각각이 100개의 문장을 발음하도록 한 샘플 음성 한 프레임의 크기가 10 ms에서 8 kHz로 샘플링 한 데이터에 세 가지 형태의 잡음이 부가 되었다. 잡음은 NOISEX-92 데이터베이스의 white noise, car noise, F16 noise에서 5, 10, 15 dB의 SNR을 가지고 테스트파일을 구성하였다. 또한 기존 Global soft decision에 의한 PESQ를 위해 고정 파라미터 q값은 1로 설정해 주었고, 제안된 방법에서의 4개의 경우의 파라미터 값은 긴 음성파일의 확률적 통계자료를 바탕으로 $\widehat{q_{00}}$=0.0246, $\widehat{q_{01}}$=0.0738, $\widehat{q_{11}}$=53.41, $\widehat{q_{10}}$=479로 설정하여 실험을 하였다.

표 1. PESQ 수치비교

Noise type	Method	SNR (dB)		
		5	10	15
White noise	Global	2.080	2.423	2.475
	Proposed	2.082	2.424	2.478
Car noise	Global	3.310	3.596	3.848
	Proposed	3.320	3.604	3.854
F16 noise	Global	2.148	2.540	2.847
	Proposed	2.196	2.554	2.858

표 1에서 보는 것과 같이 기존의 global soft decision 알고리즘과 제안된 알고리즘을 비교하기 위해 PESQ 테스트를 실시한 결과 모든 실험조건에서 제안된 방법의 결과가 좋은 것을 볼 수 있고 특히 낮은 SNR에서 보다 나은 성능을 보임을 알 수 있다. 이는 그림 1에서와 같이 고정된 파라미터 q값을 사용하던 Global soft decision보다 제안된 2차 조건 MAP를 이용한 방법이 다양한 잡음 환경에서 음성부재확률을 구할 때 보다 정확하게 추정할 수 있으므로 음성 향상 시스템에서의 성능이 좋음을 확인할 수 있다.

Ⅴ. 결 론

본 논문에서는 기존의 Global soft decision 알고리즘에서 음성부재확률의 고정 파라미터 대신 직전 2프레임 이전의 음성 존재와 부재의 정보를 부과하여 음성과 음성간의 상호 연관성을 고려하는 2차 조건 MAP를 적용하여 보다 유동적으로 음성부재확률을 구하였다. 이러한 파라미터의 조정으로 인하여 다양한 음성 환경에서의 정확한 잡음 추정을 가능하게 하며, 음성 향상 시스템에서 제안된 알고리즘이 기존의 방법보다 다양한 환경변화에 더욱 강인한 성능을 보였다.

참 고 문 헌

〔1〕 S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, Apr. 1979.

〔2〕 J. S. Lim, A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 67, pp. 1583-1604, Dec. 1979.

〔3〕 R. J. McAulary and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 137-145, Apr. 1980.

〔4〕 P. Scalart and J. Wieira Filho, "Speech Enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, Atlanta, U.S.A., pp. 629-632, May 1996.

〔5〕 Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

〔6〕 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.

〔7〕 J. W. Shin, H. J. Kwon, S. H. Jin and N. S. Kim, "Voice activity detection based on conditional MAP criterion," *IEEE Signal Processing Letters*, vol. 15, pp. 257-260, Feb. 2008.

〔8〕 J.-M. Kum, J.-H. Chang, "Speech Enhancement Based on Minima Controlled Recursive Averaging Incorporating Second-Order Conditional MAP Criterion," *IEEE Signal Processing Letters*, vol. 16, pp. 624-627, July, 2009.

〔9〕 ITU-T P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, 2001.

〔10〕 J.-H. Chang, Q.-H. Jo, D. K. Kim and N. S. Kim, "Global soft decision employing support vector machine for speech enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 57-60, Jan. 2009.

〔11〕 I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Processing Letters*, vol. 11, no. 9, pp 725-728, Sep. 2004.

금 종 모 (Jong-Mo Kum) 준회원

2008년 2월 인하대학교 전자공학과 학사

2008년 3월~현재 인하대학교 전자공학부 석사과정

<관심분야> 음성신호처리

장 준 혁 (Joon-Hyuk Chang) 종신회원

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 신호처리, 오디오 신호처리, 통신 신호처리, 휴먼/컴퓨터 인터페이스