한국음향학회지 제28권 제3호 pp. 285~289 (2009)

음향학적 반향 제거를 위한 Soft Decision 기반의 동시통화 검출

Double-Talk Detection Based on Soft Decision for Acoustic Echo Suppression

박 윤 식*, 장 준 혁*
(Yun-Sik Park, Joon-Hyuk Chang)

*인하대학교 전자공학부
(접수일자: 2009년 2월 10일; 채택일자: 2009년 3월 27일)

본 논문에서는 음향학적 반향 제거 (AES, acoustic echo suppression)를 위해 주파수영역에서 soft decision 기법에 근거한 새로운 동시통화 검출 (DTD, double-talk detection) 알고리즘을 제안한다. 제시된 방법은 효과적인 DTD를 위해 상관계수 (cross-correlation coefficient)에 기반하여 hard decision을 사용하는 기존의 알고리즘 대신 주파수 영역에서 입력 및 원단신호의 VAD (voice activity detection) 결과와 음성 통계모델에 기반한 soft decision 방법을 도입하여 전역 근단화자존재확률 (GNSPP, global near-end speech presence probability)을 DTD에 적용한다. 제안된 알고리즘은 기존의 방법과 객관적인 실험을 통해 비교 평가한 결과 다양한 배경잡음 환경에서 우수한 성능을 보였다.
핵심용어: 음향학적 반향 제거, 동시통화 검출, Soft decision
투고분야: 음성 처리 분야 (2,3)

In this paper, we propose a novel double-talk detection (DTD) technique based on soft decision in the frequency domain. In the proposed method, global near-end speech presence probability (GNSPP) considering the statistical model assumption and voice activity detection (VAD) decision of the near-end and far-end signal are applied to the DTD algorithm in the frequency domain instead of the traditional hard decision scheme using cross-correlation coefficients. The performance of the proposed algorithm is evaluated by the objective test under various environments, and yields better results compared with the conventional scheme.
Keywords: Acoustic echo suppression, Double-talk detection, Soft decision
ASK subject classification: Speech Signal Processing (2,3)

I. 서 론

이동통신 기술이 발달함에 따라 이동 전화기의 보급이 급속히 확산되면서 편리성을 고려한 핸드프리 통화 방식에서부터 휴대전화를 이용한 화상통화까지 다양하게 응용분야가 확대되고 있다. 그러나 이러한 통화 방식에서는 스피커로부터 통화음이 다양한 반향 경로를 통해 마이크로폰으로 다시 유입되는 음향학적 반향 (acoustic echo)이 발생하여 통화 음질이 저하 된다. 이러한 음향학적 반향 제거 (AES, acoustic echo suppression)를 위해 지금까지 다양한 AES 기법들이 제시되었고 [1], [2] 구체적으로, 동시통화 검출 (DTD, double-talk detection)은 AES에서 원단 (far-end) 신호로부터 다양한 반향 경로 추정을 위해 적응필터를 적용할 경우 원단신호와 무관한 근단 (near-end) 화자신호가 동시에 존재하는 동시통화 (double-talk) 시 적응필터의 오작동을 막기 위해 사용되는 필수적인 알고리즘이다 [3].

본 논문에서는 DTD를 위해 상관계수 (cross-correlation coefficient)에 기반하여 hard decision을 사용하는 기존의 방법 [3] 대신 주파수 영역에서 입력 및 원단신호의 VAD (voice activity detection) 결과와 음성 통계모델에 기반한 프레임 각각의 스펙트럼 성분들을 결합하여 프레임 단위로 전역에서 수행하는 global soft decision 방법을

책임저자: 장 준 혁 (changjh@inha.ac.kr)
인천시 남구 용현동 253 인하대학교 전자공학부
(전화: 032-860-7423; 팩스: 032-868-3654)

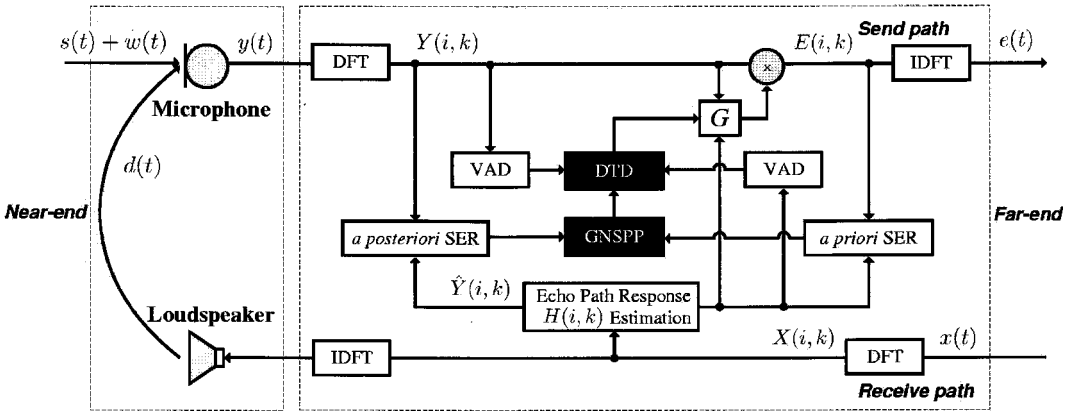


그림 1. 제안된 DTD 알고리즘의 블록 다이어그램
Fig. 1. Block diagram of the proposed DTD algorithm.

도입하여 [4], [5] 전역 근단화자존재확률 (GNSPP, global near-end speech presence probability)을 DTD에 적용하는 새로운 알고리즘을 제안한다. 제안된 방법의 성능 비교를 위해 동시통화 구간에서의 SA (speech attenuation) 테스트를 실시하였으며 제안된 기법은 다양한 잡음 환경에서 기존의 방법보다 우수한 성능을 보였다.

II. 주파수 영역에서의 임펄스응답 추정

일반적으로 입력 마이크로폰으로 전달되는 음향학적 반향신호는 다양한 반사 경로를 거쳐 입력되기 때문에 원단신호로부터 반사 경로를 고려한 임펄스응답에 대한 추정 과정이 필요하다 [6]. 음향학적 반향 제거기에서 반향신호 $d(t)$, 배경잡음 $w(t)$, 근단화자신호 $s(t)$, 원단신호와 마이크 입력신호를 각각 $x(t)$, $y(t)$라 하고 $Y(i,k)$를 $y(t)$의 i번째 프레임의 k번째 주파수 성분으로 나타낸다. 또한 반향 경로의 임펄스응답을 고려하여 d 샘플만큼 지연된 원단신호의 주파수 성분 $X_d(i,k)$부터 추정된 반향신호 $\hat{Y}(i,k)$는 다음과 같이 나타낼 수 있다 [6].

$$|\hat{Y}(i,k)| = H(i,k)|X_d(i,k)| \tag{1}$$

여기서 원단신호로부터 음향학적 반향신호를 추정하기 위한 least squares 추정 기반의 이득 필터 $H(i,k)$는 다음과 같이 계산 된다 [7].

$$H(i,k) = \left| \frac{E\{X_d^*(i,k)\,Y(i,k)\}}{E\{X_d^*(i,k)\,X_d(i,k)\}} \right| \tag{2}$$

여기서 *는 complex conjugate를 의미한다. 반향신호는 다양한 반향 경로를 거쳐 마이크로폰으로 입력되기 때문에 수시로 변화하는 반향 경로에 대한 영향을 줄이기 위해 다음식과 같이 long-term smoothing을 적용한다.

$$H(i,k) = \frac{C(i,k)}{R(i,k)} \tag{3}$$

$$C(i,k) = \zeta_C C(i-1,k) + (1-\zeta_C)|X_d^*(i,k)\,Y(i,k)| \tag{4}$$
$$R(i,k) = \zeta_R R(i-1,k) + (1-\zeta_R)X_d^*(i,k)\,X_d(i,k)$$

여기서 $\zeta_C(=0.998)$과 $\zeta_R(=0.998)$는 가중치 파라미터이다.

III. 제안된 Soft Decision 기반 Double-Talk Detection

2장에서는 주파수영역에서 least squares 추정 기반의 이득 필터를 통해 원단신호로부터 음향학적 반향신호를 추정하는 방법이 제시되었다. 3장에서는 추정된 음향학적 반향신호를 이용해 반향신호의 존재 유무에 대한 추가 정보를 확률적으로 제공하는 soft decision을 DTD에 적용하는 새로운 기법을 제안한다. 일반적으로 기존의 DTD알고리즘은 마이크 입력, 원단 및 최종적으로 추정된 근단화자신호에 대하여 서로의 신호들에 대한 상관계수 (cross-correlation coefficient)등을 추정하여 적절한 문턱 (threshold) 값에 의한 가부 판단에 따라 갱신을 결정하는 hard decision 방식을 이용하였다 [3]. 하지만 조건에 대한 확률 정보를 제공하는 파라미터를 갱신을 위한 가중치 파라미터로 적용하는 soft decision 방식이 hard decision 방법보다 향상된 성능을 보인다는 연구결과가

제안된 바 있다 [8]. 따라서 본 논문에서는 입력과 원단신호에 대한 VAD 결과와 근단 및 원단신호의 음성 통계모델에 기반하여 [9] 각각의 스펙트럼 성분들을 프레임 단위에서 전역적으로 수행하는 global soft decision 방법을 이용한 [5] 전역 근단화자존재확률 GNSPP를 DTD에 적용한다. 제안된 DTD 시스템 블록도는 그림 1과 같다.

음성의 통계 모델에 기반한 soft decision 추정을 위해 근단화자신호와 배경잡음이 상관관계가 없다는 가정 하에 근단화자신호가 존재하지 않을 때와 존재할 경우 각각의 가정 H_0, H_1에 대하여 다음과 같이 표현할 수 있다.

$$H_0\text{: near-end speech absent} : Y(i,k) = \hat{Y}(i,k) \quad (5)$$
$$H_1\text{: near-end speech present} : Y(i,k) = \hat{Y}(i,k) + S(i,k)$$

여기서 추정된 반향신호 $\hat{Y}(i,k)$는 근단화자신호인 $S(i,k)$와 통계적으로 독립이라고 가정한다. 원단신호와 근단화자신호가 complex Gaussian 분포를 따른다는 가정에서 H_0와 H_1의 확률밀도함수는 다음과 같다 [4].

$$p(Y(i,k)|H_0) = \frac{1}{\pi\lambda_r(i,k)} \exp\left[- \frac{|Y(i,k)|^2}{\lambda_r(i,k)} \right] \quad (6)$$

$$p(Y(i,k)|H_1) = \frac{1}{\pi(\lambda_s(i,k)+\lambda_r(i,k))} \cdot \exp\left[- \frac{|Y(i,k)|^2}{\lambda_s(i,k)+\lambda_r(i,k)} \right] \quad (7)$$

여기서 $\lambda_s(i,k)$, $\lambda_r(i,k)$는 각각 근단화자신호와 추정된 반향신호의 전력을 나타내며 각 주파수 채널별 near-end speech absence probability를 구하기 위해 Bayes' rule을 적용하면 다음과 같이 표현 된다 [4].

$$p(H_0|Y(i,k)) = \frac{p(Y(i,k)|H_0)p(H_0)}{p(Y(i,k)|H_0)p(H_0) + p(Y(i,k)|H_1)p(H_1)} \quad (8)$$

여기서 $p(H_0)$ $(=1 - p(H_1))$은 음성 부재에 대한 사전 확률이다. 각각의 주파수 채널별 성분이 통계적으로 독립이라는 가정으로부터, (8)식은 아래와 같이 표현 된다 [5].

$$p(H_0|Y(i)) = \frac{p(H_0)\prod_{k=1}^{N} p(Y(i,k)|H_0)}{p(H_0)\prod_{k=1}^{N} p(Y(i,k)|H_0) + p(H_1)\prod_{k=1}^{N} p(Y(i,k)|H_1)} = \frac{1}{1+qA_G(i)} \quad (9)$$

여기서 $q = p(H_1)/p(H_0)$이며, $A_G(i)$는 전역 우도비(global likelihood ratio)를 나타내며 아래와 같이 각 주파수 채널별 우도비의 곱으로 표현된다.

$$A_G(i) = \prod_{k=1}^{N} A_k(Y(i,k)) = \prod_{k=1}^{N} \frac{p(Y(i,k)|H_1)}{p(Y(i,k)|H_0)} \quad (10)$$

위의 (6)식과 (7)식으로부터 채널별 우도비 $A_k(Y(i,k))$는 다음과 같이 유도된다.

$$A_k(Y(i,k)) = \frac{p(Y(i,k)|H_1)}{p(Y(i,k)|H_0)} = \frac{1}{1+\xi(i,k)} \exp\left[\frac{\gamma(i,k)\xi(i,k)}{1+\xi(i,k)} \right] \quad (11)$$

여기서, 파라미터로 $\gamma(i,k)$, $\xi(i,k)$는 각각 *a posteriori* SER (signal to echo ratio)과 *a priori* SER로 아래와 같이 정의 된다.

$$\gamma(i,k) = \frac{|Y_S(i,k)|^2}{E\{|D(i,k)|^2\}} \quad (12)$$

$$\xi(i,k) = \frac{E\{|G(i,k)\,Y(i,k)|^2\}}{E\{|D(i,k)|^2\}} \quad (13)$$

여기서 $E[\cdot]$는 기대값 연산자, $G(i,k)$는 AES 이득을 의미하며, $|Y_S(i,k)|^2$과 $|D(i,k)|^2$는 다음과 같다.

$$|Y_S(i,k)|^2 = \zeta_Y |Y_S(i-1,k)|^2 + (1-\zeta_Y)|Y(i,k)|^2 \quad (14)$$

$$|D(i,k)|^2 = \zeta_D |D(i-1,k)|^2 + (1-\zeta_D)|\hat{Y}(i,k)|^2 \quad (15)$$

여기서 $\zeta_Y(=0.98)$과 $\zeta_D(=0.98)$는 smoothing 파라미터이다. 또한 (13)식에서 $\xi(i,k)$을 추정하기 위해 다음과 같이 Decision-Directed 추정 방법을 적용 한다 [10].

$$\hat{\xi}(i,k) = \alpha \frac{|G(i-1,k)\,Y(i-1,k)|^2}{E\{|D(i-1,k)|^2\}} + (1-\alpha)P[\gamma(i,k)-1], \quad 0 \le \alpha < 1. \quad (16)$$

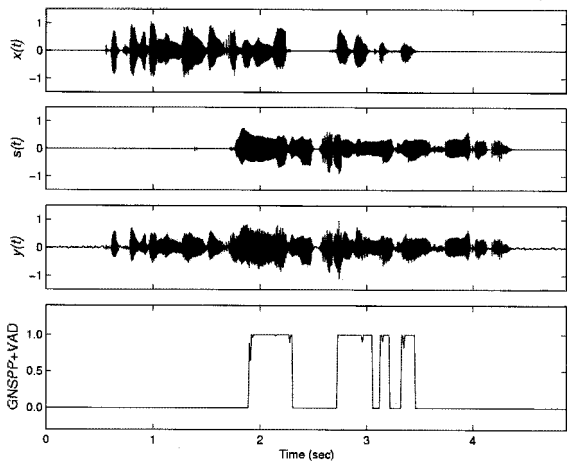


그림 2. 음향학적 반향신호에 대한 GNSPP+VAD (vehicle noise SNR=20 dB).
Fig. 2. GNSPP+VAD for acoustic echo signal (vehicle noise SNR=20 dB).

(16)식에서 α는 가중치 파라미터로 $\alpha=0.6$, $P[x]$는 $P[x]=x$ if $x\geq 0$이고, $P[x]=0$ if $x<0$을 의미하는 연산자이다. 최종적으로, 본 논문에서 제안하는 입력 및 원단신호에 대한 VAD 결과와 GNSPP $(=1-p(H_0|Y(i)))$를 적용한 DTD 알고리즘은 다음과 같다.

$$VAD(Y(i))\neq 0,\ VAD(\hat{Y}(i))\neq 0,\quad \text{and} \tag{17}$$
$$H_{DTD}(i+1,k)$$
$$=\left(1-K[p(H_0|Y(i))]\right)H_{DTD}(i,k)+K[p(H_0|Y(i))]H(i,k)$$

여기서 $VAD(Y(i))$는 IS-127 EVRC (enhanced variable rate codec)의 VAD 결과로서 GNSPP에서 원단신호 구간 결정 및 잡음신호 구간에서의 보다 정확한 결정을 위해 적용하였으며 프레임에 음성 스펙트럼이 존재할 경우 1 존재하지 않을 경우 0을 나타낸다. 따라서 $VAD(Y(i))\neq 0$과 $VAD(\hat{Y}(i))\neq 0$는 각각 입력과 원단신호에서 음성이 존재하는 경우를 나타내며 $K[x]=x$ if $x<0.4$이고, $K[x]=1$ if $x\geq 0.4$을 의미하는 연산자이다. 그림 2는 음향학적 반향신호에 대하여 vehicle 잡음이 SNR (signal-to-noise ratio) 20 dB 부가된 오염된 음성신호에 대하여 VAD decision이 적용된 GNSPP를 보여주고 있다. 따라서 제안된 DTD 알고리즘에 의해 (17)식의 반향신호 추정을 위한 이득 필터 $H_{DTD}(i,k)$는 GNSPP를 가중치 파라미터로 사용함으로서 동시통화 구간에서는 GNSPP가 1에 근접한 값을 갖기 때문에 이전 프레임의 이득 필터 값에 큰 가중치를 주게 되어 동시통화 구간에서 계산된 이득 필터의 값은 반영을 거의 하지 않고 그 이외의 구간에서는 현재 프레임에 기반한 이득 필터의 값에 큰 가중치를 주게 된다.

표 1. 동시통화 구간에서의 speech attenuation 비교.
Table 1. Speech attenuation during double-talk.

noise type	SNR (dB)	speech attenuation (dB)	
		hard decision	soft decision
white	10	1.998	1.992
	20	2.287	2.279
	30	2.326	2.305
babble	10	2.077	2.076
	20	2.296	2.266
	30	2.327	2.292
vehicle	10	1.764	1.740
	20	2.237	2.210
	30	2.320	2.284
clean speech	∞	2.331	2.295

IV. 실험 및 결과고찰

본 논문에서는 제안된 알고리즘의 성능 평가를 위해 다양한 잡음 환경에서 객관적인 실험을 수행하였다. 성능 평가는 동시통화 구간에서의 음성의 보존도를 평가하는 SA (speech attenuation) 테스트를 실시하였으며 SA의 수치는 다음과 같다.

$$SA=\frac{1}{N}\sum^{N}10\log_{10}\left[\frac{E\{s^2(t)\}}{E\{\tilde{s}^2(t)\}}\right] \tag{18}$$

(18)식에서 N은 동시통화 구간의 샘플수이고 $\tilde{s}(t)$는 출력신호의 $c(t)$에서의 근단화자신호 성분을 의미한다.

테스트 샘플을 위해 7명의 화자로 부터 얻은 8kHz로 샘플링 된 20개의 문장을 수집하고 각 문장을 원단화자와 근단화자신호로 분류하여 합성하였다. 원단화자신호로 분류된 음성은 섞기 전에 반사 경로를 고려한 실제 환경을모델링하기 위해 임펄스응답 필터를 통과 시키고 [11], [12] 입력 마이크로폰으로 들어가는 반향신호는 근단화자신호 보다 3.5 dB 작게 하였다. 모델링 환경의 장소는 $5\times 4\times 3\ m^3$ 크기로 설정 하였고 잡음 환경을 위해서 white, babble과 vehicular 잡음을 다양한 SNR로 부가하였다. 표 1은 상관계수에 기반하여 hard decision을 사용하는 기존의 DTD 방법과 soft decision에 기반한 제안된 알고리즘을 주파수 기반의 AFS 기법 [6]에 적용한 결과

파일에 대하여 동시통화 구간에서의 SA 수치를 비교한 것이다. 표 1로부터 제안된 방법이 다양한 SNR 환경에서 기존의 기법보다 SA에서 향상된 성능을 나타내고 있는 것을 볼 수 있다.

V. 결 론

본 논문에서는 주파수 영역에서 DTD 성능 향상을 위해 soft decision 기반의 새로운 알고리즘을 제안하였다. 입력 및 원단신호의 VAD 결과와 음성 통계모델에 기반하여 각각의 스펙트럼 성분들을 프레임 단위에서 전역적으로 수행하는 global soft decision 방법을 도입하여 전역 근단화자존재확률 GNSPP를 DTD에 적용하였다. 객관적 테스트 결과로부터 제안된 방법이 기존의 방법보다 동시통화 구간에서의 음성 보존도에서 개선된 결과를 나타내었다.

감사의 글

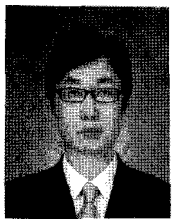
본 논문은 2007년도 정부재원 (교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었으며 (KRF-2007-313-D00734), 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2008-C1090-0902-0010).

참 고 문 헌

1. P. S. R. Diniz, *Adaptive Filtering: Algorithm and Practical Implementation*, Norwell, MA: Kluwer, 1997.
2. C. Avendano, "echo suppression in the STFT domain," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2001.
3. S. J. Park, C. G. Cho, C. Lee, and D. H. Youn, "Integrated echo and noise canceler for hands-free applications," *IEEE Trans. on Circuits and Systems II*, vol. 49, issue 3, pp. 186-195, Mar. 2002.
4. 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, no. 2, pp. 137-145, Apr. 1980.
5. 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.
6. C. Faller and C. Tournery, "Estimating the Delay and Coloration Effect of the Acoustic Echo Path for Low Complexity Echo Suppression," in *Proc. Intl. Works. on Acoust. Echo and Noise Control (IWAENC)*, 2005.
7. C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 1048-1062, Sept. 2005.
8. J. Sohn, W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 365-368, 1998.
9. 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.
10. 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, Dec. 1984.
11. S. McGovern, *A Model for Room Acoustics, 2003 [Online]*. Available: http://2pi.us/rir.html
12. S. Y. Lee and N. S. Kim, "A statistical model based residual echo suppression," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 758-761, Oct. 2007.

저자 약력

• **박 윤 식 (Yun-Sik Park)**



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

• **장 준 혁 (Joon-Hyuk Chang)**

1998년 2월 : 경북대학교 전자공학과 학사
2000년 2월 : 서울대학교 전기공학부 석사
2004년 2월 : 서울대학교 전기컴퓨터공학부 박사
2000년 3월 ~ 2005년 4월 : (주)넷더스 연구소장
2004년 5월 ~ 2005년 4월 : 캘리포니아 수립대학, 산타바바라 (UCSB) 박사후연구원
2005년 5월 ~ 2005년 8월 : 한국과학기술연구원 (KIST) 연구원
2005년 9월 ~ 현재 : 인하대학교 전자공학부 조교수