한국음향학회지 제26권 제5호 pp. 194~198 (2007)

통계적 모델 기반의 음성 검출기를 위한 변별적 가중치 학습

Discriminative Weight Training for a Statistical Model-Based Voice Activity Detection

강 상 익*, 조 규 행*, 장 준 혁*, 박 승 섭**
(Sang-Ick Kang*, Q-Haing Jo*, Joon-Hyuk Chang*, Seung Seop Park**)
*인하대학교 전자전기공학부, **서울대학교 전기컴퓨터공학부
(접수일자: 2007년 5월 2일, 수정일자: 2007년 6월 4일, 채택일자: 2007년 6월 28일)

본 논문에서는 음성의 통계적 모델에 기반한 음성검출기의 성능향상을 위해 변별적 가중치 학습 (discriminative weight training) 기반의 최적화된 우도비 테스트 (Likelihood Ratio Test, LRT)를 제안한다. 먼저, 기존의 통계모델기반의 음성검출기를 분석하고, 이를 기반으로 MCE (minimum classification error)방법을 도입하여, 각 주파수 채널별로 다른 가중치를 가지는 우도비 기반의 음성검출 결정법 (decision rule)을 제시한다. 제안된 알고리즘은 비정상 (non-stationary)잡음환경에서 기존의 동일 가중치를 가지는 기하 평균 기반의 음성검출기와 비교하였으며, 우수한 성능을 보인다.

핵심용어: 음성 검출기, Minimum classification error, 통계적 모델, 우도비

투고분야: 음성처리 분야 (2.4)

In this paper, we apply a discriminative weight training to a statistical model-based voice activity detection (VAD). In our approach, the VAD decision rule is expressed as the geometric mean of optimally weighted likelihood ratios (LRs) based on a minimum classification error (MCE) method which is different from the previous works in that different weights are assigned to each frequency bin which is considered more realistic. According to the experimental results, the proposed approach is found to be effective for the statistical model-based VAD using the LR test.

Key words: Voice activity detection, Minimum classification error, Statistical model, Likelihood ratio

ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

음성과 비음성 구간을 검출하는 음성 검출기 (voice activity detector, VAD)는 음성 부호화, 음성인식 그리고 음향학적 반향제거기 등 음성 통신 시스템에서 많이 적용된다. 특히, 음성검출기는 다중 접속 기술에서 한정된 주파수 대역을 효율적으로 사용하기 위한 가변 전송률 부호화기의 실현을 위해 필수적인 부분을 차지하고 있으며 이와 관련하여 다양한 형태의 알고리즘이 제안되고 있다. 많은 알고리즘 중에 Ephraim과 Malah의 연구에서 시작된 minimum mean square error (MMSE) 기반의 음성 향상 기법에 사용된 음성의 존재와 부재에 대한 통계적 모델을 음성 검출기에 적용한 것이 매우 우수한 성능을 가진 것으로 알려져 있는데 [1-8], 구체적으로 음성에 대한 가우시안 통계모델을 decision-directed (DD) 기법에 도입하여 최적의 음성검출 파라미터 추정에 사용하여 나온 음성의 존재와 부재에 대한 우도비 (likelihood ratio, LR)를 기하평균한 결정식으로 음성검출여부를 최종적으로 판단한다 [3].

본 논문에서는 기존의 음성의 통계적 모델 기반의 음성검출기에서 제시된 각 주파수 채널별 우도비의 단순한 기하 평균을 이용하여 문턱값을 비교하는 방법 대신, 변별적 가중치 학습 (discriminative weight training)을 위

책임저자: 장 준 혁 (changjh@inha.ac.kr)
420-751 인천시 남구 용현동 253 인하대학교 전자전기공학부
(전화: 032-860-7423, 팩스: 032-868-3654)

한 minimum classification error (MCE) 방법을 이용하여 도출된 최적화된 가중치를 각 주파수 채널별 우도비에 적용하여 기하 평균을 구성하는 새로운 방식을 제안하며, 다양한 잡음환경에서 기존의 통계적 모델 기반의 음성검출기와 성능을 비교하였다.

II. 통계모델 기반의 음성 검출기의 이해

시간축 상에서 원래의 음성신호 $x(t)$에 잡음신호 $n(t)$이 인가된 입력신호 $y(t)$을 discrete Fourier transform (DFT)을 통해 주파수 축으로 변환되어 아래와 같이 표현된다.

$$Y(t) = X(t) + N(t) \tag{1}$$

여기서 $Y(t) = [Y_1, Y_2, ..., Y_M]$, $X(t) = [X_1, X_2, ..., X_M]$, 그리고 $N(t) = [N_1, N_2, ..., N_M]$는 각각 잡음에 오염된 음성신호, 원래의 음성신호, 잡음신호의 DFT 계수 벡터를 나타낸다. 주어진 가설 H_0, H_1이 각각 음성의 부재와 존재를 표현한다고 하면 각 주파수 채널별로 다음과 같이 기술된다.

$$H_0: \text{speech absent} \quad : \quad Y_k(t) = N_k(t) \tag{2}$$

$$H_1: \text{speech present} : \quad Y_k(t) = X_k(t) + N_k(t). \tag{3}$$

음성과 잡음신호의 스펙트럼이 복소 가우시안 분포를 따른다는 가정으로부터 가설 H_0와 H_1을 조건으로 한 확률밀도함수는 아래와 같이 주어진다 [3].

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

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

여기서 $\lambda_{x,k}$와 $\lambda_{d,k}$는 각각 채널별 음성과 잡음의 분산이며, 이 때 k번째 주파수 밴드에 대한 우도비는 아래와 같이 구한다.

$$\Lambda_k \equiv \frac{p(Y_k|H_1)}{p(Y_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \tag{6}$$

여기서 $\xi_k = \lambda_{x,k}/\lambda_{d,k}$와 $\gamma_k = Y_k/\lambda_{d,k}$는 각각 a priori signal-to-noise ratio (SNR)와 a posteriori SNR이다 [3]. 음성 부재 구간에서 갱신되는 잡음 신호로부터 구한 잡음 분산 $\lambda_{d,k}$를 이용하여 a posteriori SNR γ_k를 추정하며, 또한 a priori SNR ξ_k는 decision-directed (DD) 방식을 이용하여 아래와 같이 추정한다 [1].

$$\hat{\xi}_k(t) = \alpha\frac{|\hat{X}_k(t-1)|^2}{\lambda_{d,k}(t-1)} + (1-\alpha)P[\gamma_k(t)-1] \tag{7}$$

여기서 $|\hat{X}_k(t-1)|^2$은 이전 프레임에서 추정된 음성 신호의 k번째 스펙트럼 성분의 크기에 대한 추정치이며, MMSE에 기반하여 구한다 [3]. 또한 α는 가중치 값이며, 연산자 $P[\cdot]$은 아래와 같이 정의된다.

$$P[x] = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

기존의 일반적인 통계적 모델 기반의 음성 검출기에 대한 결정식은 각각의 주파수 채널에서 구해진 우도비를 기하 평균하여 아래와 같이 음성 검출 여부를 판단한다 [2-8].

$$\log \Lambda(t) = \frac{1}{M}\sum_{k=1}^{M} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

여기서 M은 전체 주파수 대역의 개수이며, η는 음성 검출 문턱값이다.

III. MCE 훈련을 이용한 최적화된 가중치의 도출

기존의 음성의 통계모델에 기반한 음성검출기는 식 (9)에서 보듯이 각 주파수별 성분이 독립이라는 가정에 기반하여 간단한 우도비의 기하 평균식을 이용한 점을 살펴볼 수 있다. 그러나, 각 주파수 채널별 우도비가 음성검출 성능에 균일한 기여를 한다는 것은 음성신호의 주파수특성의 분포 등을 고려하면 실제적이지 않다. 따라서, 본 논문에서는 각 우도비 $\log \Lambda_k$에 최적화된 가중치를 인가함으로서 보다 효과적인 음성검출기를 제안하고 새로운 결정식을 다음과 같이 정의한다.

$$\Lambda_w = \frac{1}{M}\sum_{k=1}^{M} w_k \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{10}$$

먼저, 입력 신호로부터 구한 각 주파수 채널별 우도비에 각각 다른 가중치 w_k를 적용하여 새로운 우도비 $w_k \log \Lambda_k$를 구하며 각 가중치는 다음의 조건을 만족한다.

$$\sum_{k=1}^{M} w_k = 1, \quad w_k \geq 0 \text{ for } k = 0, 1, \cdots, N-1. \tag{11}$$

최적화된 가중치를 적용한 우도비 벡터를 $\Lambda_w = \{w_1 \log \Lambda_1, w_2 \log \Lambda_2, \ldots, w_M \log \Lambda_M\}$라 하고 Λ_w를 $\frac{1}{M}\sum_{k=1}^{M} w_k \log \Lambda_k$와 같이 정의한다. 그리고 훈련할 데이터의 각각의 프레임에서 음성 $g_s(\cdot)$와 비음성 $g_n(\cdot)$을 구분하는 두 개의 함수를 다음과 같이 정의한다.

$$g_s(\Lambda_w) = \Lambda_w - \theta \tag{12}$$

$$g_n(\Lambda_w) = \theta - \Lambda_w, \tag{13}$$

여기서 θ는 음성과 비음성을 구분하는 문턱값이다. 이때 문턱값은 음성과 비음성 훈련 데이터의 분포에서 겹치는 경계값 (θ=0)을 사용하였다. 제안된 연구에서는 최적화 알고리즘에 기반한 가중치를 구하기 위해 generalized probabilistic descent (GPD) 기반의 MCE 훈련을 적용하며 [9], 실제로 훈련 데이터 프레임 $\Lambda_w(t)$의 분류 오류 D를 다음과 같이 정의한다.

$$D(\Lambda_w(t)) = \begin{cases} -g_s(\Lambda_w(t)) + g_n(\Lambda_w(t)) \text{ if } g_s \text{ is ture class} \\ -g_n(\Lambda_w(t)) + g_s(\Lambda_w(t)) \text{ if } g_n \text{ is true class} \end{cases} \tag{14}$$

여기서 식 (14)이 음수인 값을 가질 때 올바른 분류가 되며 이를 기반으로 하는 손실함수 (loss function) L은 다음과 같이 sigmoid 함수 형태로 정의된다.

$$L = \frac{1}{1+\exp(-\beta D(\Lambda_w))}, \ \beta > 0 \tag{15}$$

여기서 β는 sigmoid 함수의 기울기를 나타낸다. 최적화된 가중치를 구하기 위해선 손실함수가 최소가 되어야 한다. MCE 훈련과정을 통해 가중치를 조정하는 과정에서 식 (11)과 같은 제약조건 때문에 가중치 w를 $\tilde{w}$로 변환한다.

$$\tilde{w} = \{\tilde{w_1}, \tilde{w_2}, \cdots, \tilde{w_k}\} \tag{16}$$

$$\tilde{w_k} = \log w_k. \tag{17}$$

가중치 $\tilde{w_k}$는 매 프레임마다 연속적으로 존재하는데, 각 주파수 가중치는 다음과 같은 식으로 갱신된다 [10].

$$\tilde{w_k}(n+1) = \tilde{w_k}(n) - \epsilon \frac{\partial L}{\partial w_k}\Big|_{\tilde{w_k} = \tilde{w_k}(n)} \tag{18}$$

여기서 ϵ (> 0)는 단조롭게 감소하는 구간의 크기이다. $\tilde{w_k}$를 갱신한 후에 아래의 식과 같이 w_k로 복원된다.

$$w_k = \frac{\exp(\tilde{w_k})}{\sum_{i=1}^{M} \exp(\tilde{w_i})} \tag{19}$$

식 (19)에서 정규화 된 가중치를 사용했을 때 식 (11)를 만족한다.

기존의 통계적 모델 기반의 음성 검출기의 결정식인 식 (9)와 비교하여, 본 논문에서는 위에서 제시된 MCE 훈련 방법을 이용해 구한 식 (19)의 가중치를 각각의 주파수 채널에 곱해서 구해진 우도비를 기하 평균하여 (10)과 같이 최종적으로 음성 검출 여부를 판단한다. 결론적으로 각 주파수가 음성검출의 성능에 다르게 미치는 영향을 고려한 최적화된 가중치가 적용된 새로운 우도비를 이용하여 기하 평균 기반의 결정식을 도출하는 점이 주목할 만하다.

IV. 실험 결과 비교 및 분석

본 논문에서 제안된 MCE 기법을 이용한 최적화된 가중치 기반의 음성 검출기의 성능을 평가하기 위해 기존의 우도비 테스트를 이용한 통계적 모델 기반의 음성 검출기의 성능과 Receiver Operating Characteristic (ROC) 곡선을 이용하여 비교하였다 [3]. 최적화된 가중치를 도출하기 위해 손실함수 L에서 정의된 가파기 파라미터 $\beta = 1$로 결정하였다. 실험에 사용된 데이터는 성능 평가를 위해 총 230초의 깨끗한 음성 데이터에 음성과 비음성 부분을 10 ms마다 수동으로 표시하였다. 분류된 음성 데이터의 음성 구간은 총 57.1%로 유성음 44.0%, 무성음 13.1%로 구성되었으며 잡음 환경은 음성 데이터에

street, car 잡음이 5, 15 dB SNR로 부과되었다. 그림 1, 2, 3, 4의 (a)는 각각 street와 car 잡음 환경에 따라 음성 검출기의 문턱값을 변경하면서 실제 음성을 음성이라고 판단한 음성 검출 확률 (Pd)과 비음성에 대해 음성이라고 판단한 오경보 확률 (Pf)을 5, 15 dB SNR에서 측정

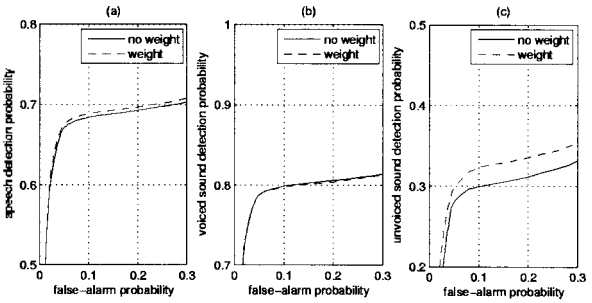

그림 1. Street 잡음 5 dB에서의 ROC 곡선
Fig. 1. ROC curves for street noise (SNR = 5 dB)
(a) speech (b) voiced sound (c) unvoiced sound.

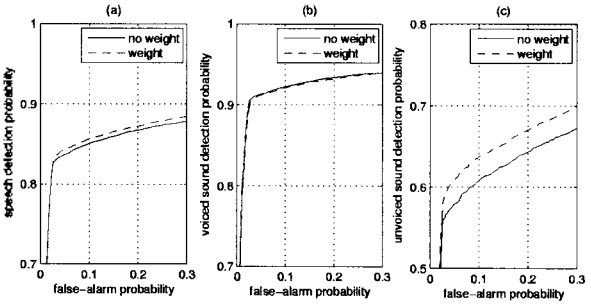

그림 2. Car 잡음 5 dB SNR에서의 ROC 곡선
Fig. 2. ROC curves for car noise (SNR = 5 dB)
(a) speech (b) voiced sound (c) unvoiced sound.

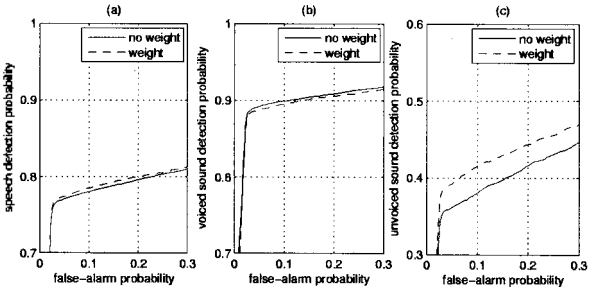

그림 3. Street 잡음 15 dB SNR에서의 ROC 곡선
Fig. 3. ROC curves for street noise (SNR = 15 dB)
(a) speech (b) voiced sound (c) unvoiced sound.

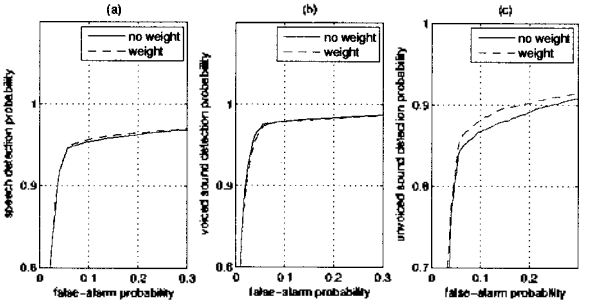

그림 4. Car 잡음 15 dB SNR에서의 ROC 곡선
Fig. 4. ROC curves for car noise (SNR = 15 dB)
(a) speech (b) voiced sound (c) unvoiced sound.

한 ROC 곡선이며, (b)와 (c)는 (a)와 동일한 SNR과 동일한 잡음 환경에서 각각 유성음과 무성음에서의 ROC 곡선이다. 실험 분석 결과 동일한 SNR의 주어진 잡음 조건에서 각 채널별 가중치를 적용한 음성 검출기의 경우 전체적으로 기존의 음성 검출기보다 향상된 성능을 보여주며, 특히 그림 (b)와 (c)를 비교해보면 유성음 구간에서는 거의 성능이 동일하였으나 무성음 구간에서 뚜렷한 성능개선이 이루어졌다. 이를 통해 MCE 훈련으로부터 도출된 가중치가 주로 무성음구간에서 음성검출성능향상에 기여하는 것을 알 수 있다. 결론적으로 기존의 우도비의 기하 평균으로 음성을 검출하는 것 보다 각 채널별 최적화된 가중치를 이용한 음성 검출 방법이 우수한 것을 확인할 수 있다.

V. 결 론

본 논문에서는 음성의 존재와 부재에 대한 통계적 모델에 기반한 각 주파수 채널별 우도비를 단순히 기하 평균을 취하여 문턱값과 비교하는 기존의 방법 대신, MCE를 이용하여 도출한 최적화된 가중치를 각 주파수 채널별 우도비에 적용하는 새로운 방법으로 기존의 방식보다 향상된 통계모델기반의 음성 검출기를 제시하였으며, 객관적인 실험 결과로부터 제안된 음성 검출기의 성능이 우수함을 알 수 있었다.

감사의 글

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

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. Acoustics, Speech, Sig. Process., ASSP-32, (6) 1190-1121, Dec. 1984.
2. J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," Proc. Int. Conf. Acoustics, Speech, and Sig. Process., 1, 365-368, May

1998.
3. J. Sohn, N. S. Kim, and W. Sung, "A statistical model- based voice activity detection," IEEE Sig. Process. Lett., 6 (1) 1-3, Jan. 1999.
4. Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," IEEE Sig. Process. Lett., 8 (10) 276-278, Oct. 2001.
5. J. -H. Chang, J. W. Shin, and N. S. Kim, "Voice activity detector employing generalised gaussian distribution," Electron. Lett., 40 (24) 1561-1563, Nov. 2004.
6. J. -H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," IEEE Trans. Sig. Process., 54 (6) 1965-1976, June 2006.
7. Y. C. Lee and S. S. Ahn, "Statistical model-based VAD algorithm with wavelet Transform," IEICE Trans. Fundamentals., E89-A, (6) 1594-1600, June 2006.
8. J. Ramirez, J. M. Gorriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," IEEE Sig. Process. Lett., 13 (8) 497-500, Aug. 2006.
9. B. -H. Juang, W. Chou, and C. -H. Lee, "Minum classification error rate methods for speech recognition," IEEE Trans. Speech Audio Processing, 5 (3) 257-265, May 1997.
10. Y. Kida, T. Kawahara, "Voice activity detection based on optimally weighted combination of muliple feature," Interspeech, 2621-2624, Sep. 2005.

저자 약력

• 강 상 익 (Sang-Ick Kang)

2007년 2월 : 인하대학교 전자공학과 학사
2007년 3월~현재 : 인하대학교 전자공학과 석사과정

• 조 규 행 (Q-Haing Jo)

2004년 2월 : 인하대학교 전자공학과 학사
2006년 9월~현재 : 인하대학교 전자공학과 석사과정

• 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 : 퓨넷더스 연구소장
2004년 5월~2005년 4월 : 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원
2005년 5월~2005년 8월 : 한국과학기술연구원 (KIST) 연구원
2005년 9월~현재 : 인하대학교 전자전기공학부 조교수

• 박 승 섭 (Seung Seop Park)

2000년 2월 : 서울대학교 전기공학과 학사
2002년 2월 : 서울대학교 전기공학부 석사
2007년 2월 : 서울대학교 전기컴퓨터공학부 박사
2000년 3월~2003년 4월 : 퓨넷더스 연구소 연구원
2007년 3월~현재 : 서울대학교 박사후연구원