우도비를 이용한 DBN 기반의 음성 검출기

Voice Activity Detection based on DBN using the Likelihood Ratio

김상균, 이상민*
S. K. Kim, S. M. Lee

요 약

본 논문에서는 입력된 신호에 의해 결정되는 각 주파수 밴드별 우도비(likelihood ratio, LR)를 deep belief networks(DBN)의 입력층으로 이용하는 새로운 음성 검출기(voice activity detection, VAD) 알고리즘을 제안한다. 기존의 통계적 모델 기반의 음성 검출기는 음성 구간을 판단하기 위해 우도비를 기하 평균을 이용한 결정식을 사용한다. 제안된 음성 검출기는 이 결정식을 대신해 DBN을 이용하여, 오검출 확률을 최소화하도록 학습을 한다. 제안된 DBN 기반의 음성 검출 알고리즘은 통계적 모델 기반의 음성 검출기의 성능을 개선한 support vector machine(SVM) 기반의 음성 검출기와 정상 및 비정상 잡음 환경에서 다양한 조건을 부과하여 비교하였다. 제안된 알고리즘이 기존의 SVM 기반의 알고리즘보다 전체 오분류 확률 [0.7, 2.7]의 향상 폭을 보였다.

ABSTRACT

In this paper, we propose a novel scheme to improve the performance of a voice activity detection(VAD) which is based on the deep belief networks(DBN) with the likelihood ratio(LR). The proposed algorithm applies the DBN learning method which is trained in order to minimize the probability of detection error instead of the conventional decision rule using geometric mean. Experimental results show that the proposed algorithm yields better results compared to the conventional VAD algorithm in various noise environments.

Keyword : Voice Activity Detection, Likelihood Ratio, Deep Belief Networks

1. 서론

잡음 환경에서 음성이 활동하는 구간을 분별하는 음성 검출기(voice activity detector, VAD)는 음성 향상, 음성 부호화기, 음성 인식 그리고 피드백 제거 등 음성신호처리 기술에 있어서 매우 중요한 기술이며 다양한 분야에 적용된다[1-3]. 예를 들어 보청기에서는 리시버 단에서 증폭된 출력 신호가 다양한 경로를 통해 다시 마이크를 통해 입력되는 피드백 신호가 발생한다. 이는 보청기 사용자에게 매우 불쾌한 소리를 생성하는 원인이 되므로 반드시 제거돼야만 한다. 이때 이용되는 피드백 제거 알고리즘 중 대표적인 인접투사(affine projection, AP) 알고리즘의 수렴속도와 정상상태오차에 영향을 주는 스텝사이즈(step size)의 크기를 음성의 존재 확률에 따라 가변적으로 부여하는 연구가 최근 발표되었으며 우수한 성능을 보였다[4]. 이 가변 스텝사이즈를 이용한 AP 기반의 피드백 제거 알고리즘에서는 음성 구간을 정확히 판단하는 것이 피드백 제거에 큰 영향을 미친다. 또 다른 적용 분야인 음성 향상에서는 잡음을 제거하기 위해 정확한 잡음 전력을 추정하는 것이 핵심 기술이다. 이때 대부분의 음성 향상 기법에서 음성이 없는 구간에서 잡음 전력을 업데이트한다[5-6]. 음성이 있는 구간에서 업

접 수 일 : 2014.07.25
심사완료일 : 2014.08.18
게재확정일 : 2014.08.19.
* 김상균 : 인하대학교 전자공학부 박사과정
greenwhity@nate.com (주저자)
이상민 : 인하대학교 전자공학부 교수
sanglee@inha.ac.kr
※ 본 연구는 서울시 산학연 협력사업(SS100022) 지원에 의하여 이루어진 연구로서, 관계부처에 감사드립니다.

데이트가 이루어진다면 잡음뿐만 아니라 음성 성분도 손실을 입기 때문에 음질이 매우 떨어지게 된다. 이러한 이유로 음성 검출기의 성능을 개선시키기 위해 현재까지 다양한 알고리즘들이 연구되고 있다.

그중 우도비(likelihood ratio, LR)를 기하 평균하여 결정식으로 이용한 통계적 모델 기반 음성 검출 방법이 제안되었으며 발표 당시 성능이 우수한 것으로 알려져 있다[7]. 통계적 모델에 기반한 음성 검출기에서 주목할 점은 Ephraim과 Malah의 연구[8]에서 제안한 minimum mean square error(MMSE) 기반의 음성 향상 기법에 사용된 음성의 존재와 부재에 대한 통계적 모델을 가우시안 분포로 가정하여 우도비 테스트(likelihood ratio test, LRT)에 적용한 것이다. 또한 직접 구할 수 없는 음성 파라미터인 사전 신호 대 잡음비(*a priori* signal-to-noise ratio, *a priori* SNR)를 decision-directed 기법을 이용하여 추정을 한다.

이러한 통계적 모델 기반의 음성 검출 알고리즘은 각 주파수 밴드에서 구한 우도비를 단순한 기하평균을 하기 때문에 멀티 특징 벡터로서의 이점을 반영하는데 제한적이다. 단점을 보완하고자 멀티 특징 벡터를 입력으로 이용하여 추상화된 특징벡터를 추출하는 방법인 support vector machine(SVM) 기반의 음성 검출기가 발표되었다[9]. 여기서 사용된 SVM은 선형 분리가 어려운 패턴을 kernel function을 이용하여 고차원으로 사상시킴으로써 선형 분리가 가능하게 만들며 이진 분류에 뛰어난 성능을 보인다. 하지만 SVM은 1개의 층을 가지는 구조로 성능 개선에 한계가 있다[10].

최근 다층 회로망 기반의 deep belief networks(DBN)이 멀티 특징 벡터의 추상화된 특성을 잘 도출한다고 알려져 있다[11-12]. Hinton은 기존의 학습과정에서 지역 극소로 수렵하는 문제를 해결하고자 역전파 알고리즘을 수해하기 전에 다층 신경망의 초기 값을 설정하는 pre-training을 해주는 방법을 제안하였다[13]. 따라서 본 논문에서는 음성 검출기의 성능을 향상시키기 위해 각 주파수 밴드에서 구한 우도비를 DBN의 멀티 특징 벡터로 사용하여 음성 구간을 검출하는 음성 검출 알고리즘을 제안한다. 제안된 음성 검출 방법은 다양한 잡음 환경에서 기존의 음성 검출 알고리즘들과 비교하였으며 향상된 성능을 보였다.

본 논문의 2 장에서는 제안한 알고리즘의 입력층에 사용될 멀티 특징 벡터인 우도비를 유도하는 과정을 소개하고 3 장에서는 우도비를 이용한 DBN 기반의 음성 검출 알고리즘에 대해 논한다. 4 장에서는 제안된 방법의 우수성을 기존의 음성 검출 방법과의 성능 비교를 통해 보여주며, 마지막으로 5 장에서 결론을 기술하였다.

2. 통계적 모델 기반 음성검출기

시간영역에서 입력 신호 $y(t)$는 깨끗한 음성 신호 $x(t)$에 잡음 신호 $d(t)$를 합한 신호로 가정하고 여기서 t는 이산 시간을 나타낸다. 입력 신호를 이산 퓨리에 변환(discrete Fourier transform, DFT)을 취해 주파수 영역으로 변환하면 다음과 같다.

$$Y(k,n) = X(k,n) + D(k,n) \tag{1}$$

여기서 $Y(k,n)$, $X(k,n)$ 그리고 $D(k,n)$은 n번째 프레임 k번째 주파수 밴드에서 입력 신호, 깨끗한 음성 신호 그리고 잡음 신호의 DFT 계수이다. 음성의 부재와 존재를 가설 $H_0(k,n)$와 $H_1(k,n)$로 표현 하면 각 주파수 밴드별로 아래와 같이 표현할 수 있다.

$$H_0(k,n) \ : \ Y(k,n) = D(k,n) \tag{2}$$

$$H_1(k,n) \ : \ Y(k,n) = X(k,n) + D(k,n) \tag{3}$$

음성과 잡음 신호의 스펙트럼이 복소 가우시안 분포를 따른다고 가정을 하면, 가설 $H_0(k,n)$, $H_1(k,n)$을 조건부 확률로 적용한 확률밀도함수는 아래와 같다[7].

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

$$p(Y(k,n)|H_1(k,n)) = \frac{1}{\pi[\lambda_x(k,n)+\lambda_d(k,n)]} \exp\left\{-\frac{|Y(k,n)|^2}{\lambda_x(k,n)+\lambda_d(k,n)}\right\} \tag{5}$$

여기서 $\lambda_x(k,n)$와 $\lambda_d(k,n)$는 각 프레임에서 주파수 밴드별 음성과 잡음의 분산이며, 이때 k번째 주파수 밴드에 대한 우도비는 다음과 같이 구한다.

$$\Lambda(k,n) \equiv \frac{p(Y(k,n)|H_1)}{p(Y(k,n)|H_0)} = \frac{1}{1+\xi(k,n)} \exp\left\{\frac{\gamma(k,n)\xi(k,n)}{1+\xi(k,n)}\right\} \tag{6}$$

여기서 $\xi(k,n)$은 사전 신호 대 잡음비이고 $\gamma(k,n)$은 사후 신호 대 잡음비이며 아래와 같이 얻을 수 있다[8].

$$\xi(k,n)=\frac{\lambda_x(k,n)}{\lambda_d(k,n)} \tag{7}$$

$$\gamma(k,n)=\frac{|Y(k,n)|^2}{\lambda_d(k,n)} \tag{8}$$

여기서 사후 신호 대 잡음비 $\gamma(k,n)$은 음성 부재 구간에서 갱신되는 신호로부터 얻은 잡음 분산 $\lambda_d(k,n)$을 이용하여 추정하며, 사전 신호 대 잡음비 $\xi(k,n)$은 decision-directed 기법을 이용하여 다음과 같이 추정한다[8].

$$\hat{\xi}(k,n)=\alpha\frac{|\hat{X}(k,n-1)|^2}{\lambda_d(k,n-1)}+(1-\alpha)P[\gamma(k,n)-1] \tag{9}$$

여기서 $|\hat{X}(k,n-1)|$은 이전 프레임의 k번째 주파수 밴드에서 추정된 음성 신호의 스펙트럼 성분의 크기이며, MMSE를 기반으로 구한다[8]. 또한 α는 가중치 값이며, $P[\bullet]$ 연산자는 다음과 같이 정의된다.

$$P[x]=\begin{cases}x & \text{if} \quad x\geq 0\\ 0 & otherwise.\end{cases} \tag{10}$$

기존의 통계적 모델 기반의 음성검출기에 대한 결정식은 우도비를 기하 평균하여 아래와 같이 문턱값 η와 비교하여 음성 활동 여부를 판단 한다[7].

$$\log\Lambda(n)=\frac{1}{M}\sum_{k=1}^{M}\log\Lambda(k,n)\underset{H_0}{\overset{H_1}{\gtrless}}\eta \tag{11}$$

여기서 M은 전체 주파수 대역의 개수이다.

3. 제안된 DBN 기반 음성검출기

3.1 DBN 모델의 학습

기존의 다층 신경망의 학습은 일반적으로 학습 데이터와 이에 대응하는 레이블을 이용하여 역전파 (back-propagation, BP) 알고리즘을 통해서 이루어졌다. 하지만 일반적으로 학습 다층 신경망에서 레이어의 개수가 증가할수록 역전파 알고리즘의 학습 시간이 증가하고, 학습 결과가 지역 극소로 수렴하

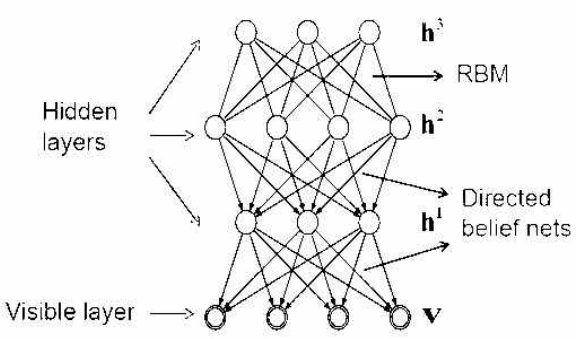

그림 1. Deep belief network 구조

여 성능이 저하되는 경우가 발생한다. 이러한 문제점을 해결하기 위해서 Hinton 교수는 BP을 수행하기 전에 다층 신경망의 초기 값을 설정해 주기 위해서 RBM (restricted Boltzmann machine) 기반의 pre-trainig 방법을 제안하였다[13].

그림 1은 3개의 은닉층과 1개의 입력층을 갖는 DBN 구조를 나타낸다. 은닉층의 각 유닛들을 서로 조건부 독립이라 가정고 다음과 같은 절차로 학습을 수행한다.

step 1. 트레이인 샘플로부터 $v^{(0)}=[v_1,v_2]$를 입력 벡터로 받음

step 2. 각 입력 벡터를 이용하여 은닉층의 유닛이 활성화 확률을 계산하며 여기서 $sigm(\bullet)$은 시그모이드 함수를 나타낸다.

$$\boldsymbol{h}^{(0)}=\begin{cases}p(h_1=1|\boldsymbol{v}^{(0)})=sigm(b_1+\sum_{i=1}^{2}w_{i,1}v_i^{(0)})\\ p(h_2=1|\boldsymbol{v}^{(0)})=sigm(b_2+\sum_{i=1}^{2}w_{i,2}v_i^{(0)})\\ p(h_3=1|\boldsymbol{v}^{(0)})=sigm(b_3+\sum_{i=1}^{2}w_{i,3}v_i^{(0)})\end{cases} \tag{12}$$

step 3. 구해진 은닉층 유닛의 확률을 입력벡터로 사용하여 입력층 유닛이 활성화 확률을 계산한다.

$$\boldsymbol{v}^{(1)}=\begin{cases}p(v_1^{(1)}|\boldsymbol{h}^{(0)})=sigm(c_1+\sum_{j}^{3}w_{1,j}h_j)\\ p(v_2^{(1)}|\boldsymbol{h}^{(0)})=sigm(c_2+\sum_{j}^{3}w_{2,j}h_j)\end{cases} \tag{13}$$

step 4. 위에서 구한 $v^{(1)}$를 이용하여 은닉층 유닉의 활성화 확률을 구한다.

$$\boldsymbol{h}^{(1)} = \begin{cases} p(h_1 = 1|\boldsymbol{v}^{(1)}) = sigm(b_1 + \sum_{i=1}^{2} w_{i,1} v_i^{(1)}) \\ p(h_2 = 1|\boldsymbol{v}^{(1)}) = sigm(b_2 + \sum_{i=1}^{2} w_{i,2} v_i^{(1)}) \\ p(h_3 = 1|\boldsymbol{v}^{(1)}) = sigm(b_3 + \sum_{i=1}^{2} w_{i,3} v_i^{(1)}) \end{cases} \quad (14)$$

step 5. step 1~4에서 구한 값을 이용하여 가중치 w, 아래 방향 바이어스 b, 위 방향 바이어스 c를 업데이트 한다. $< \bullet >$ 연산자는 minibatch에 의해 구해지는 평균값을 나타낸다.

$$\begin{aligned} \triangle w_{ij} &= \epsilon\left(< v_i^{(0)} h_j^{(0)} > - < v_i^{(1)} h_j^{(1)} >\right) \\ \triangle b_i &= \epsilon\left(< v_i^{(0)} > - < v_i^{(1)} >\right) \\ \triangle c_j &= \epsilon\left(< h_j^{(0)} > - < h_j^{(1)} >\right) \end{aligned} \quad (15)$$

3.2 DBN 모델의 음성 검출

DBN기반의 음성 검출은 2장에서 구한 우도비를 입력 벡터로 이용하여 학습을 통해 파라미터를 구한 후 피드포워드(feedforward) 방식으로 이루어진다[13].

$$\begin{aligned} d_k &= sigm(b_k^{(L+1)} \sum_i w_{k,i}^{(L+1,L)} h_i^L) \\ h_i^L &= sigm(b_i^{(L)} \sum_j w_{k,i}^{(L,L-1)} h_i^{L-1}) \\ &\vdots \\ h_m^1 &= sigm(b_m^{(1)} \sum_r w_{m,r}^{(1,0)} v_r) \end{aligned} \quad (16)$$

여기서 $w_{i,j}^{(L,L-1)}$와 $b_i^{(L)}$는 인접한 두 층의 가중치와 바이어스를 나타낸다. d_k는 최종 출력층의 값으로 잡음 모델에 대한 출력값은 d_1, 음성 모델에 대한 출력값은 d_2로 나타낸다. 제안한 음성 검출기의 최종 결정식은 아래와 같다.

$$d_2 - d_1 \underset{noise}{\overset{speech}{\gtrless}} \eta \quad (17)$$

위 결정식은 단순히 d_2와 d_1의 차이를 문턱값과 비교하여 음성 유/무를 판단한다. 이후 연구에서는 최적의 음성 검출을 위해 이 두 개의 출력값을 매개변수로 사용한 함수를 구하도록 할 것이다.

4. 실험 결과 및 고찰

표 1. SVM 기반의 음성 검출기와 제안된 음성검출기의 성능 비교

		SVM-based			Proposed		
Noise	SNR (dB)	P_e	P_m	P_{fa}	P_e	P_m	P_{fa}
Car	5	6.1	3.4	9.9	5.4	6.5	3.8
	10	5.7	2.8	9.7	4.5	4.0	5.3
	15	5.4	2.5	9.4	4.4	3.8	5.1
Office	5	16.5	10.2	25.3	15.3	4.3	30.6
	10	13.1	9.4	18.6	12.1	3.0	24.7
	15	10.1	6.9	14.6	8.0	4.9	12.4
Street	5	10.8	7.3	15.7	8.2	4.9	12.8
	10	9.3	6.7	12.9	7.6	4.9	11.4
	15	8.9	6.5	12.2	6.2	3.4	10.2

본 논문에서 제안한 새로운 음성 검출 방법의 성능을 평가하기 위해 P_e(probability of total error), P_m(probability of miss) 그리고 P_{fa}(probability of false alarm)를 측정하였다. 제안된 알고리즘은 기존의 통계적 모델 기반의 음성 검출기의 성능을 향상시킨 SVM기반의 음성 검출 알고리즘과 성능을 비교하였다. SVM은 확률 분포가 주어지지 않은 데이터에 대해 오분류 확률을 최소화 하는 구조적 위험 최소화 방법에 기초하고 있다. 실험에 사용된 데이터는 총 456초의 깨끗한 음성 데이터에 음성과 묵음 부분을 10ms 마다 수동으로 표시하였으며 8 kHz로 샘플링하였다. 음성 데이터의 음성 구간은 총 57.1%로 유성음 44.0%, 무성음 13.1%로 구성되었으며 잡음환경을 만들기 위해 Car, Office, Street 잡음을 5, 10 그리고 15 dB SNR로 깨끗한 음성 데이터에 더하여 사용하였다.

표 1은 기존의 음성 검출 알고리즘과 제안된 음성 검출 알고리즘의 P_e, P_m, P_{fa}를 보여준다. Car 잡음은 저주파 대역에 잡음이 집중되어있는 비교적 정상 잡음이며 Office와 Street 잡음은 시간에 따라 통계적 특성이 변하는 비정상 잡음이다. 두 음성 검

출기는 공통적으로 비정상 잡음보다 정상 잡음에서 좋은 음성 검출 성능을 보인다. 표 1을 보면 전체적으로 제안된 알고리즘이 기존의 SVM 기반의 음성 검출 알고리즘보다 우수한 성능을 확인할 수가 있다. 일반적으로 음성 신호처리에서는 음성 성분을 놓치는 P_m을 최소화하려는 경향이 있다. 하지만 Car 잡음에서는 제안된 알고리즘의 P_m이 높고 P_f이 낮은 경향을 보이는데 이것은 전체 성능을 보여주는 P_e가 최소가 되는 조건으로 학습을 시킨 결과로 이러한 결과가 나온 것이다. 이는 DBN 학습을 하는 과정에서 최소가 되는 값을 어떤 것으로 할 것인가에 따라서 다양한 목적에 최적화된 결과를 얻을 수 있다는 것을 의미한다.

5. 결론

본 논문에서는 음성 검출의 정확도를 향상시키기 위해 우도비를 특징벡터로 사용한 DBN 기반의 음성 검출 알고리즘을 제안하였다. 기존의 통계적 모델 기반의 음성 검출 방법에서 각 주파수 밴드별 우도비를 구하고 DBN의 학습 방법과 음성 결정식을 유도하였다.

제안된 알고리즘의 우수성을 보이기 위해 기존의 SVM 기반의 음성 검출 알고리즘과 표 1에서 P_e, P_m, P_{fa}를 비교하였다. 다양한 잡음 조건에서 제안된 알고리즘이 기존의 알고리즘보다 전체 오분류 확률 [0.7, 2.7]의 향상 폭을 보였다. 표 1에서 확인한 것처럼 기존의 알고리즘보다 제안된 음성 검출기의 성능이 향상된 것을 알 수 있었다.

참 고 문 헌

[1] Y. S. Park and S. Lee, "Voice activity detection using global speech absence probability based on teager energy for speech enhancement," IEICE TRANSACTIONS on Information and Systems, vol. E95-D, no. 10, pp. 2568-2571, Oct. 2012.

[2] B. F. Wu and K. C. Wang, "Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 13, no. 5, pp. 762-775, Sep. 2005.

[3] 김상균, 권장우, 이상민, "일반화된 정규-라플라스 분포를 이용한 음성검출기," 멀티미디어학회 논문지, 제17권, 제3호, pp. 294-299, 2014.

[4] Y. S. Kim, J. H. Song, S. K. Kim, S. Lee, "Variable step-size affine projection algorithm based on global speech absence probability for adaptive feedback cancellation." Journal of Central South University, vol. 21, pp. 646-650, 2014.

[5] R. Martin, "Spectral subtraction based on minimum statistics," in Proc. Eur. Signal Processing Conf., pp. 1182 - 1185, Sep. 1994.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controller recursive averaging," IEEE Trans. Speech Audio Processing, vol. 11, no. 5, pp. 466 - 475, Sep. 2003.

[7] 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.

[8] 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. 1190-1121, Dec. 1984.

[9] Q. H. Jo, Y. S. Park, K. H. Lee, and J. H. Chang. "A support vector machine-based voice activity detection employing effective feature vectors." IEICE Transactions on Communications, vol. E91-B, no. 6, pp. 2090-2093, 2008.

[10] V. N Vapnik, "An overview of statistical learning theory," IEEE Trans. Neural Networks, vol. 10, no. 5, pp. 988-999, Sep. 1999.

[11] Y. Bengio, "Learning deep architectures for AI," Foundat. Trends in Mach. Learn., vol. 2, no. 1, pp. 1-127, 2009.

[12] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layerwise training of deep networks," Proc. Adv. Neural Inf. Process. Syst., vol. 19, pp. 153-161, 2007.

[13] G. Hinton, "A practical guide to training restricted Boltzmann machines," Momentum, vol 9, pp. 1-9, 2010.

김 상 균

2008년 2월 인하대학교 전자공학과 학사
2010년 10월 인하대학교 전자공학부 석사
2013년 3월 - 현재 인하대학교 전자공학부 박사과정

관심분야 : Speech Signal Processing, Acoustic Signal Processing

이 상 민

1987년 2월 인하대학교 전자공학과 학사
1989년 2월 인하대학교 전자공학과 석사
2000년 인하대학교 전자공학과 박사
2006년 6월 - 현재 인하대학교 전자공학과 부교수

관심분야 : Brain-Machine interface, Bio-Siganl Processing, Psyco-Acoustic