2000年度 第13回 信號處理合同學術大會 論文集 第13卷 1號

음성스펙트럼의 클러스터링을 이용한 음성검출기법 개선 [1]

김 태 영, 김 남 수, 김 태 정

서울대학교 전기공학부

Speech Detection using Speech Spectrum Clustering

Tae Young Kim, Nam Soo Kim and Taejeong Kim

School of Electrical Engineering, Seoul National University

e-mail: {kty@infolab, nkim@, tkim@}.snu.ac.kr

요약문

본 연구에서는 기존의 통계 이론에 근거한 음성 검출 기법을 제안하는 음성 스펙트럼 모형화기법을 통해 개선시키고자 한다. 기존의 방법과는 달리 음성을 하나의 단일 모형이 아닌 여러 클래스(class) 모형의 결합체로 간주한다. 각 클래스 모형의 추정을 위해 신호원 부호화(source coding)의 클러스터링(clustering)과 유사한 기법을 제안하고, 이를 이용한 두 가지의 검출 기법을 제안한다. 하나는 각각의 클래스에 대해 LRT(likelihood ratio test)를 수행하고, 이를 최종적으로 통합하는 기법이고 다른 하나는 각 클래스의 모형으로부터 혼합모형(mixture model)을 구하여 이를 이용하여 LRT를 수행하는 방법이다. 제안한 두 가지 방법 모두 비교적 적은 연산량 증가에도 불구하고 실험 결과 기존 방법에 비해 매우 우수한 성능을 보였다.

1 서론

음성 검출(speech detection)은 음성의 분석 또는 음성 인식의 전단계로서 많이 연구되어 오고 있다. 특히 최근에는 이동 통신 및 인터넷 전화의 발달과 함께 이에 이용되는 가변 전송율 음성 부호화기에의 응용 목적으로 많은 관심을 받고 있다. 음성 검출 방법은 크게 경험적인 방법과 통계 이론적인 방법의 두 가지로 나눌 수 있다. 전자의 방법은 현재 널리 쓰이는 방법으로 입력 신호의 에너지, 영교차율(zero crossing rate), 저주파 에너지 등의 몇 개의 특징벡터를 이용하여 문턱치(threshold)와 비교하는 방법이다. 이 방법은 비교적 간단하나 문턱치 설정에 있어서 대개 경험적인 방법에 의존하고 다양한 잡음 환경에서 견실한 성능을 보이지 못한다는 단점이 있다. 반면, 후자의 방법은 최근에 활발히 연구되는 방법으로서 신호에 대한 통계모델을 기반으로 검출이론에 입각하여 음성존재 여부를 결정한다. 현재 많이 사용되는 통계모델은 DFT(Discrete Fourier Transform) 계수에 대한 복소수 가우시안 분포 모델(complex gaussian model)이다[1][2]. 본 연구에서는 기존의 통계 이론적인 방법을 새로운 음성 스펙트럼 모형화를 통해 개선시키고자 한다. 기존 방법에서는 음성 및 잡음 스펙트럼을 각각 하나의 단일 모형으로 추정하였다. 하지만, 실제 음성은 유성음 무성음으로 구별 가능하다는 사실에서 알 수 있듯이 여러 클래스의 모형이 결합된 형태라고 보는 것이 더 타당하다.

본 연구에서는 음성의 각 클래스의 모형을 추정하고, 이를 이용하여 새로운 음성 검출 기법을 제안하고자 한다. 각 클래스 모형의 추정을 위해 신호원 부호화(source coding)의 클러스터링(clustering)과 유사한 방법을 제안하고, 가정하는 음성 신호원 모형에 따라 다음의 두 가지 검출방법을 제안한다. 하나는 각각의 클래스에 대해 LRT(likelihood ratio test)를 수행하고, 이를 최종적으로 통합하는 기법이고 다른 하나는 각 클래스의 모형으로부터 혼합모형(mixture model)을 구하여 이를 이용하여 LRT를 수행하는 방법이다. 실험 결과 제안하는 두 가지 방법 모두 다양한 잡음 환경에서 기존 방법에 비해 더

[1]본 연구는 한국과학재단 특정기초연구(과제번호 98-0101-03-01-3)의 지원으로 수행되었음.

낮은 성능을 보였다.

2 음성 검출 모형

우리는 음성 검출 문제를 DFT 영역에서 다루려고 한다. 각 입력 프레임에 대하여 $\mathbf{S}$, $\mathbf{N}$, $\mathbf{X}$는 음성, 잡음, 입력신호의 L-차원 DFT 계수 벡터이다. 각 벡터의 k번째 원소는 S_k, N_k, X_k로 각각 주어진다. 음성 클래스의 개수를 M개라 하면, 음성 검출 문제의 가설은 다음과 같이 주어진다.

H_0: 음성 부재: $\mathbf{X}=\mathbf{N}$
S: 음성 존재: $\mathbf{X}=\mathbf{S}+\mathbf{N}$
H_1: 음성이 클래스1 에 속함
H_2: 음성이 클래스2 에 속함
...
H_M: 음성이 클래스M 에 속함

음성 신호와 잡음 신호가 독립이라 가정하고, 두 신호의 DFT계수들이 각각 서로 독립인 가우시안 확률 변수라고 가정한다면[3], 각 가설에서의 확률 밀도 함수는 다음과 같이 주어진다.

$$p(X_k|H_0)=\frac{1}{\pi\lambda_N(k)}\exp\left\{-\frac{|X_k|^2}{\lambda_N(k)}\right\} \quad (1)$$

$$p(X_k|H_i)=\frac{1}{\pi[\lambda_S^{(i)}(k)+\lambda_N(k)]}\exp\left\{-\frac{|X_k|^2}{\lambda_S^{(i)}(k)+\lambda_N(k)}\right\} \quad (i=1,2,\cdots,M), \quad (2)$$

여기서 $\lambda_N(k)$는 잡음 신호 N_k의 분산을, $\lambda_S^{(i)}(k)$는 클래스i 에 속하는 음성 신호 S_k의 분산을 나타낸다.

우선, 음성 신호가 매 프레임에서 각 클래스중 하나의 클래스에 의해서만 생성된다고 가정하자. 이때, 음성의 M개의 클래스는 음성을 분할(partition)한다 가정한다. 이는 $S=\cup_{i=1}^{M}H_i$와 $H_i\cap H_j=\emptyset(i\neq j)$임을 의미한다. 이 때, 음성 유무에 대한 LRT는 다음과 같이 주어진다.

$$\Lambda=\frac{p(\mathbf{X}|S)}{p(\mathbf{X}|H_0)}=\frac{1}{p(\mathbf{X}|H_0)}\cdot\frac{\cup_{i=1}^{M}p(H_i)p(\mathbf{X}|H_i)}{p(S)}$$

$$=\sum_{i=1}^{M}p(H_i|S)\Lambda_{i0}. \quad (3)$$

단, $\Lambda_{i0}=\frac{p(\mathbf{X}|H_i)}{p(\mathbf{X}|H_0)}$이며, 각 클래스의 사전 확률이 동일하다 가정하면 $p(H_i|S)=\frac{1}{M}\cdot(1-p(H_0))$로 주어진다.

식(1)과 식(2)를 식(3)에 대입하면 최종적인 Λ는

$$\Lambda=\sum_{i=1}^{M}p(H_i|S)\prod_{k=0}^{L-1}\left(\frac{1}{1+\xi_{ki}}\exp\left\{\frac{\gamma_k\xi_{ki}}{1+\xi_{ki}}\right\}\right) \quad (4)$$

로 주어지게 된다[2][3]. 여기서 $\xi_k=\frac{\lambda_S^{(i)}(k)}{\lambda_N(k)}$는 음성 클래스$i$ 에서의 *a priori* SNR이며, $\gamma_k=\frac{|X_k|^2}{\lambda_N(k)}$는 *a posteriori* SNR이다.

이제, 음성 신호가 매 프레임에서 클래스의 혼합상태(mixture)에서 생성된다고 가정하자. 이 경우에는 $p(X_k|S)$는

$$p(X_k|S)=\sum_{i=1}^{M}p(H_i|S)p(X_k|H_i) \quad (5)$$

이므로, 최종적인 공산비 Λ는

$$\Lambda=\frac{p(\mathbf{X}|S)}{p(\mathbf{X}|H_0)}=\prod_{k=0}^{L-1}\sum_{i=1}^{M}\left(\frac{p(H_i|S)}{1+\xi_{ki}}\exp\left\{\frac{\gamma_k\xi_{ki}}{1+\xi_{ki}}\right\}\right) \quad (6)$$

로 주어지게 된다.

3 파워 스펙트럼 추정

좋은 검출 성능을 위해서는 견실한 파라미터의 추정이 필요하다. 현재 널리 사용되는 잡음 모형의 추정은 음성 신호가 존재하지 않는 경우에는 잡음 모형을 갱신하지만, 존재하는 경우에는 잡음 모형을 갱신하지 않는 방법이다. 본 논문에서는 위의 아이디어를 음성 유무에 대한 soft-decision에 기반하여 아래와 같은 잡음 스펙트럼의 추정 기법을 제안한다.

$$\lambda_N^{(m)}(k)=p(H_0|\mathbf{X}^{(m)})\cdot[\alpha\lambda_N^{(m-1)}(k)+(1-\alpha)|X_k^{(m)}|^2]$$
$$+(1-p(H_0|\mathbf{X}^{(m)}))\cdot\lambda_N^{(m-1)}(k). \quad (7)$$

여기서 m은 프레임 번호, α는 기억상수(forgetting factor)이며, $p(H_0|\mathbf{X}^{(m)})$은 다음과 같이 주어진다.

$$p(H_0|\mathbf{X}^{(m)})=\frac{p(H_0|\mathbf{X}^{(m)})}{\sum_{i=0}^{M}p(H_i|\mathbf{X}^{(m)})}=\frac{p(H_0)p(\mathbf{X}^{(m)}|H_0)}{\sum_{i=0}^{M}p(H_i)p(\mathbf{X}^{(m)}|H_i)}$$
$$=\frac{1}{1+\sum_{i=1}^{M}\frac{p(H_i)}{p(H_0)}\Lambda_{i0}^{(m)}}. \quad (8)$$

음성 모형의 추정은 잡음 모형 추정과는 반대로 음성 신호가 존재하지 않는 경우에는 음성 모형을 갱신하지 않고, 존재하는 경우에는 음성 모형을 갱신하는 것이 바람직하다. 이때, 음성 모형의 갱신은 매 프레임에서 가장 확률이 높은 하나의 클래스에 대해서만 실시한다. 위 아이디어를 soft-decision에 근거하여 다음과 같은 음성 스펙트럼 추정기법을 제안한다.

$I=\arg\max_{i=1}^{M}\Lambda_{i0}$

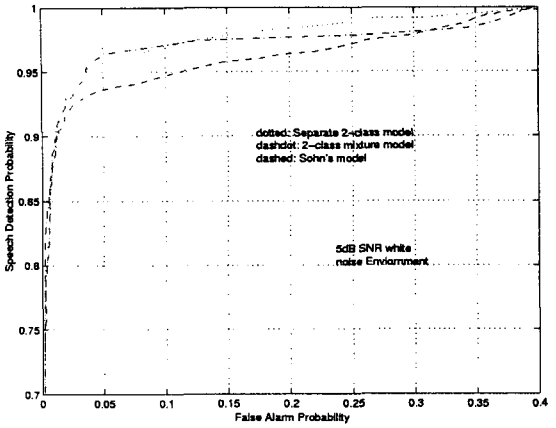

그림 1: 기존 방법과 제안 방법의 ROC 특성 비교: 5dB SNR 백색 잡음

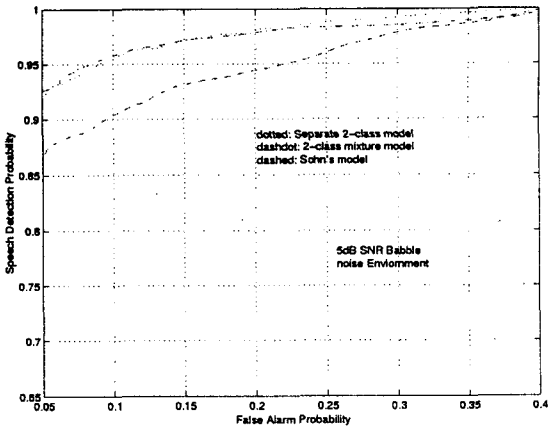

그림 2: 기존 방법과 제안 방법의 ROC 특성 비교: 5dB SNR 배블 잡음

$$\lambda_{S(I)}^{(m)}(k) = p(H_0|\mathbf{X}^{(m)})\lambda_{S(I)}^{(m-1)}(k)+(1-p(H_0|\mathbf{X}^{(m)}))\times$$
$$\{\alpha\lambda_{S(I)}^{(m-1)}(k) + (1-\alpha)U[|X_k^{(m)}|^2 - \lambda_{S(I)}^{(m-1)}(k)]\}$$

$$\lambda_{S(i)}^{(m)}(k) = \lambda_{S(i)}^{(m-1)}(k), \qquad i \neq I$$

단, $U[x]$는 $x \geq 0$이면 $U[x] = x$, 그렇지 않으면 $U[x] = 0$이다. 제안하는 방법은 신호원(source)의 클러스터링에 많이 사용되는 Lloyd 방법의 첫 부분과 유사하다[4].

4 실험 결과

컴퓨터 모의실험을 통해 제안하는 검출기법과 이전에 제안된 [2]의 방법을 비교하였다. 실험을 위해 여성 화자에 의한 60초 길이의 음성샘플을 이용하였다. 오염되지 않은 음성샘플을 20ms 단위의 프레임으로 나눈 후 수작업을 통해 음성유무를 미리 판별하여 이를 기준으로 삼았다. 음성 검출은 보통의 경우 매 프레임에서의 판정결과를 바로 최종결과로 삼지 않고, 과거의 판정결과를 이용하여 다시 수정한다. 이를 행오버(Hang-over)라 하는데, 이를 위해 [2]에서 제안된 HMM기반의 방법을 사용하였다. 음성에 첨가하는 배경잡음으로는 NOISEX-92 데이터 베이스의 백색잡음과 배블(babble)잡음을 이용하였다. 본 실험에서는 2가지의 음성 스펙트럼 클래스를 이용하였다. 성능 비교를 위해 ROC(Receiver Operating Curve)를 구하였다. 그림 1과 그림 2에서 제안하는 방법들이 기존 방법에 비해 더 낳은 성능을 보임을 확인 할 수 있다. 제안하는 두 가지 음성 모형중 클래스의 혼합모형 보다는 분할모형이 약간 더 낳은 검출 성능을 보임을 알 수 있다.

음성 스펙트럼 추정에 있어서 각 클래스의 초기조건이 어떤 영향을 주는 지에 대한 실험의 결과를 그림3에 나타내었다. 각 클래스는 초기조건과 상관없이 시간이 지남에 따라 동일한 형태의 클래스모형으로 수렴해 감을 확인할 수 있다. 이는 실제 음성 신호원이 뚜렷이 구별되는 두 가지 클래스의 모형으로 이루어져 있음을 의미한다. 이 실험을 통해 제안하는 검출 모형은 실제 음성 모형을 더 잘 표현하고, 따라서 더 낳은 검출성능을 보임을 확인할 수 있다.

5 결론

본 논문에서는 스펙트럼의 통계 모형에 기반한 음성 검출기를 새로운 스펙트럼 모형 및 추정 방법을 통해 개선시켰다. 음성 스펙트럼을 하나의 모형이 아닌 여러 클래스의 합으로 모형화하였고, 이에 사용되는 클래스 모형의 추정방법 및 음성검출기법을 제안하였다. 실험을 통해 제안하는 방법이 음성을 기존 방법에 비해 더 잘 모형화하고, 따라서 음성 검출 성능을 향상시킴을 확인 할 수 있었다. 제안하는 방법은 특별한 훈련과정이 필요하지 않고, 기존 방법에 비해 연산량을 클래스의 개수 배 만큼 증가시킨다. 실제 실험결과 2개의 클래스만으로도 성능향상이 확연하기에 제안하는 방법은 실제 이용가치가 매우 높음을 알 수 있다.

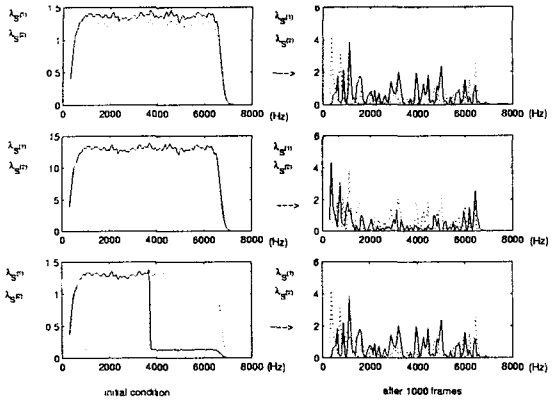

그림 3: 초기조건에 따른 음성 스펙트럼의 클래스 모형 추정 결과

참고문헌

[1] J. Sohn and W. Sung, "A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Estimat ion," *Proc. of Int. Conf. Acoust., Speech, and Signal Processing,* 1998, pp.365-368.

[2] J. Sohn, N. Kim and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1-3, Jan. 1999.

[3] Robert J. McAulay and Marilyn L. Malpass, Martin Cohn, and Roger Khazan, "Speech Enhancement Using Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, and Signal Processing,* vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[4] Allen Gersho and Robert M. Gray, Martin Cohn, and Roger Khazan, *Vector Quantization and Signal Compression,* Kluwer Academic Publishers, 1992.