제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

통계적 모델에 근거한 음성 검출기의 설계

손 종 서, 김 남 수, 성 원 용

서울대학교 전기공학부

Design of a Statistical Model Based Voice Activity Detector

Jongseo Sohn, Nam Soo Kim and Wonyong Sung

School of Electrical Engineering, Seoul National University

email : {sohn@lpc, nkim@plaza, wysung@dsp}.snu.ac.kr

요약문

본 연구에서는 가변 전송율 음성 부호화기를 위한 음성 검출기를 통계적 모델을 적용하여 설계한다. 제안된 음성 검출기는 음성 파라미터를 decision-directed 방식으로 추정함으로써 LRT(likelihood ratio test)를 이용하여 동작 특성이 우수한 판정 규칙을 유도한다. 또한 음성 발생 사건들을 1차의 Markov process로 모델링함으로써 과거의 관찰들을 현재 프레임의 음성 검출 과정에서 고려할 수 있는 행오버 (hang-over) 알고리즘을 개발한다. 개발된 음성 검출기는 고려된 실험 환경에서 ITU-T 표준인 G.729 Annex B 음성 검출기보다 매우 우수한 성능을 나타내었다.

1 서론

가변 전송율 음성 부호화기(variable rate speech coder)에서는 음성 신호의 유무를 판별하여 음성 신호가 존재하는 구간에 보다 많은 비트를 할당함으로써 음질의 저하없이 평균 전송율을 낮출 수 있다. 최근 CDMA(Code Division Multiple Access)를 이용한 이동 통신이나 디지탈 음성 저장 등과 같은 가변 전송율 음성 부호화 응용에 대한 수요가 증가됨에 따라 음성 검출기(VAD : voice activity detector)의 역할이 점점 중요해지고 있다. 기존의 음성 검출기 알고리즘들은 배경 잡음의 통계적 특성이 음성 신호보다 비교적 오랜 기간동안 변하지 않는다고 가정함으로써 음성 신호의 간섭에도 불구하고 시간에 따라 변하는 배경 잡음의 통계량을 추정한다. 각 프레임에서 음성 신호의 유무를 판정하기 위하여 현재 프레임에서 관찰된 신호의 통계적 특성은 추정된 배경 잡음의 통계량과 각각의 판정 규칙(decision rule)에 따라 비교된다. 이러한 초기 판정 결과는 약한 음성 신호의 misdetection을 방지하기 위해 행오버 (hang-over) 방식에 의해 보정된다.

지금까지의 음성 검출기들은 대부분 경험적으로 설계되었기 때문에 관련된 파라미터 값들을 최적화하기가 어렵고 다양한 배경 잡음 환경에 대한 견실성이 떨어지는 문제점이 있다. 이러한 문제점을 해결하기 위하여 최근 음성과 배경 잡음 신호의 DFT (discrete Fourier transform) 계수들에 대한 가우시안 모델을 바탕으로 음성 검출기를 최적화하고자 하는 연구가 행해졌다 [1]. 그 연구에서는 음성 파라미터를 ML (maximum likelihood) 방식으로 추정함으로써 LRT(likelihood ratio test)를 적용하여 판정 규칙을 유도하였다. 본 연구에서는 음성 파라미터를 decision-directed (DD) 방식으로 추정함으로써 판정 규칙을 개선시키고, 또한 HMM(hidden Markov model)에 근거한 행오버 방식을 제안하여 음성 검출기의 동작 특성을 크게 향상시킨다.

2 LRT를 이용한 판정 규칙

음성 신호가 상관도가 없는 가산적 잡음에 (uncorrelated additive noise) 의해 왜곡되었을 때, 음성 검출기가 매

프레임마다 선택하여야 할 두 가설은 다음과 같다.

$$H_0 : \text{speech absent} \ : \mathbf{X} = \mathbf{N}$$
$$H_1 : \text{speech present} : \mathbf{X} = \mathbf{N} + \mathbf{S}$$

위에서 $\mathbf{S}, \mathbf{N}, \mathbf{X}$는 각각 그들의 k번째 원소들이 S_k, N_k, X_k로 표현되는 L 차원 DFT 계수 벡터들이다. 본 연구에서는 각 프로세스들의 DFT 계수들이 대략적으로 독립인 가우시안 확률 변수들이라는 통계적 모델을 가정한다 [2]. H_0와 H_1이 발생하였을때의 조건부 확률은 다음과 같이 주어진다.

$$p(\mathbf{X}|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\{-\frac{|X_k|^2}{\lambda_N(k)}\} \quad (1)$$

$$p(\mathbf{X}|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k)+\lambda_S(k)]} \cdot \exp\{-\frac{|X_k|^2}{\lambda_N(k)+\lambda_S(k)}\} \quad (2)$$

단 $\lambda_N(k)$와 $\lambda_S(k)$는 각각 N_k와 S_k의 분산들로서 power spectral density의 각 주파수에서의 샘플 값들에 해당한다. k번째 주파수 밴드에서의 likelihood ratio는 아래와 같다.

$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\{\frac{\gamma_k\xi_k}{1+\xi_k}\} \quad (3)$$

위에서 $\xi_k = \lambda_S(k)/\lambda_N(k)$이고 $\gamma_k = |X_k|^2/\lambda_N(k)$이며, 이들은 각각 *a priori*와 *a posteriori* SNR (signal to noise ratios)로 불린다 [2]. 판정 규칙은 각 주파수 밴드들에서의 likelihood ratio들의 기하 평균으로부터 다음과 같이 유도된다.

$$\log\Lambda = \frac{1}{L}\sum_{k=0}^{L-1} \log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \quad (4)$$

배경 잡음 통계량 추정 과정에서 $\lambda_N(k)$들은 이미 주어졌다고 가정하면, 미지의 파라미터 ξ_k들이 각 프레임마다 추정되어야 한다.

ξ_k를 위한 ML estimator는 다음과 같이 유도된다.

$$\hat{\xi}_k^{(ML)} = \gamma_k - 1. \quad (5)$$

식 (5)를 식 (4)에 대입하고 LRT를 적용하면 다음과 같은 ISD (Itakura-Saito distortion) 기반 판정 규칙이 얻어진다 [1].

$$\log\hat{\Lambda}^{(ML)} = \frac{1}{L}\sum_{k=0}^{L-1}\{\gamma_k - \log\gamma_k - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \quad (6)$$

식 (6)의 좌변은 ISD의 특성상 항상 0이상의 값을 나타내는데 이는 추정된 likelihood ratio가 H_1으로 바이어스(bias)되어 있음을 의미한다.

이 바이어스를 줄이기 위해 본 연구에서는 다음과 같은 decision-directed (DD) *a priori* SNR 추정 방법을 고려하였다[2].

$$\hat{\xi}_k(n)^{(DD)} = \alpha\frac{\hat{A}_k^2(n-1)}{\lambda_N(k,n-1)} + (1-\alpha)P[\gamma_k(n)-1] \quad (7)$$

위에서 n은 프레임 인덱스이고, $x \geq 0$이면 $P[x] = x$, 아니면 $P[x] = 0$로 정의 되며, $\hat{A}_k(n-1)$들은 이전 프레임의 음성 신호의 크기 추정치(signal amplitude estimates)들이다. 크기 추정치를 구하기 위해 최소 자승 오차 추정기를 (minimum mean-square error estimator) 사용하였다[2]. 식 (7)의 DD 방식은 ML 방식보다 보다 부드럽고 완만한 *a priori* SNR 추정값들을 제공하기 때문에[3], 음성이 없는 구간에서 추정된 likelihood ratio들의 불규칙한 변동이 줄어들게 된다.

3 HMM 기반 행오버

실제 음성 검출기에서는 음성의 꼬리 부분등과 같이 약한 음성 신호를 배경 잡음으로 오판하는 경우를 방지하기 위해 과거의 관찰들 혹은 판정 결과들을 고려하여 초기 판정 결과를 보정하는 행오버를 행한다. 기존의 행오버 알고리즘들은 음성의 misdetection을 줄이기 위해 음성 프레임으로부터 잡음 프레임으로의 전이를 지연시키는 등의 방법을 택함으로서 false-alarm 비율을 증가시키는 단점이 있다.

일반적으로 행오버는 인접한 음성 프레임의 발생간에는 강한 상관도가 있다는 사실에 기반을 둔다. 이러한 특성을 외연적이고 정량적으로 표현하기 위해 본 연구에서는 프레임 상태들의 수열을 1차의 Markov process로 모델링하였다. 현재 상태는 오직 바로 직전 상태에만 의존한다는 Markov process의 특성때문에 음성 프레임 발생간의 상관적인 특성은 조건부 확률 $P(q_n = H_1|q_{n-1} = H_1)$로 표현될 수 있다. 이때 다음과 같은 제약이 필요하다.

$$P(q_n = H_1|q_{n-1} = H_1) > P(q_n = H_1) \quad (8)$$

위에서 q_n은 n번째 프레임의 상태로서 H_0와 H_1 중 하나이다.

위의 Markov process가 시불변(time invariant)이라고 가정하면 $a_{ij} = P(q_n = H_j|q_{n-1} = H_i)$과 같은 표기가 가능하다. 또한 process가 stationary하다고

가정함으로써 $P(q_n = H_i) = P(H_i)$을 얻을 수 있는데, $P(H_0)$와 $P(H_1)$는 Markov process의 정상 상태 분포(steady state distribution)로서 $P(H_0) + P(H_1) = 1$의 조건과 다음과 같은 stationarity 방정식으로 부터 얻어진다.

$$a_{01}P(H_0) = a_{10}P(H_1) \tag{9}$$

그러므로 전체 프로세스는 오직 두개의 파라미터로 (예를 들면 a_{01}와 a_{10}) 특징지을 수 있다.

이 Markov 프레임 상태 모델에서는 현재 상태가 현재의 관찰 뿐아니라 과거의 관찰들에도 의존하기 때문에, 본 연구에서는 과거의 관찰들을 다음과 같은 방식으로 판정 규칙에 반영시킨다.

$$\begin{aligned}\mathcal{L}(n) &= \frac{p(\mathcal{X}_n|q_n = H_1)}{p(\mathcal{X}_n|q_n = H_0)} \\ &= \frac{P(H_0)}{P(H_1)}\frac{P(q_n = H_1|\mathcal{X}_n)}{P(q_n = H_0|\mathcal{X}_n)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta\end{aligned} \tag{10}$$

이때 $\mathcal{X}_n = \{\mathbf{X}(n), \mathbf{X}(n-1), \ldots, \mathbf{X}(1)\}$는 현재 프레임 n까지 관찰들의 집합을 나타낸다. 식 (10)에서의 *a posteriori* 확률 비를, 즉 $\Gamma(n) = P(q_n = H_1|\mathcal{X}_n)/P(q_n = H_0|\mathcal{X}_n)$를, 계산하기 위하여 다음과 같은 전향 변수(forward variable)를 정의한다.

$$\alpha_n(i) = p(q_n = H_i, \mathcal{X}_n) \tag{11}$$

$\alpha_n(i)$은 잘 알려진 forward procedure [4]를 이용하여 아래와 같이 풀수 있다.

$$\alpha_n(i) = \begin{cases} P(H_i)p(\mathbf{X}(1)|q_1 = H_i), & \text{if } n = 1 \\ \left(\alpha_{n-1}(0)\,a_{0j} + \alpha_{n-1}(1)\,a_{1j}\right) \cdot & \\ \quad p(\mathbf{X}(n)|q_n = H_i), & \text{if } n \geq 2. \end{cases} \tag{12}$$

위의 결과를 이용하여 $\Gamma(n)$을 구하기 위한 회귀적 식은 다음과 같이 얻어진다.

$$\Gamma(n) = \frac{\alpha_n(1)}{\alpha_n(0)} = \frac{a_{01} + a_{11}\Gamma(n-1)}{a_{00} + a_{10}\Gamma(n-1)}\Lambda(n) \tag{13}$$

위에서 $\Lambda(n)$은 식 (4)에서 주어지는 n번째 프레임의 likelihood ratio를 의미한다. 따라서 최종 판정 통계량은 $\mathcal{L}(n) = [P(H_0)/P(H_1)]\Gamma(n)$로 구한다.

4 실험 결과

제안된 알고리즘들의 유효성을 검증하기 위하여 본 연구에서는 행오버를 적용했을 때와 적용하지 않았을때 ML 및 DD 방식의 판정 규칙들의 음성 검출 (detection) 및 오검출 (false-alarm) 확률들을 (P_d와 P_f) 비교하였다. 프레임 상태 모델을 위해 $a_{01} = 0.2$와 $a_{10} = 0.1$를 사용하였다. P_d와 P_f를 계산하기 위해여 46초 동안의 깨끗한 음성 샘플들에 대해 매 10 msec 프레임마다 수동으로 기준 판정을 만들었다. 기준 판정에서 음성 프레임의 비율은 34.27 %(유성음과 무성음 프레임이 각각 27.76 %, 6.51 %)이었다. P_d는 기준 음성 프레임들 중 음성 검출기가 음성 프레임으로 판별한 비율, P_f는 기준 잡음 프레임들 중 음성 검출기가 음성으로 오판한 비율로 정의한다.

전술한 네가지의 판정 규칙들의 P_d와 P_f간의 trade-off를 나타내는 동작 특성(ROC : receiver operating characteristic)이 그림 1에 나타나 있다. 음성 검출기 알고리즘들은 각각 NOISEX-92 데이터 베이스에서 얻어진 차량(vehicular) 및 백색(white) 잡음에 의해 5 dB SNR로 더럽혀진 음성들에 적용되었다. 그림 1에서 보여진 바와 같이 DD 방식에 근거한 판정 규칙은 ML 기반 판정 규칙보다 P_f가 의미있는 영역에서 더 높은 P_d를 나타낸다. 또한 제안된 행오버 방식은 ML 및 DD 판정 규칙에서 모두 P_d를 증가시킴을 볼 수 있다.

최종적으로 제안된 행오버 알고리즘을 DD 판정 규칙에 적용시키고, [1]에서 제안된 연판정 정보를 이용한 배경 잡음 스펙트럼 추정 알고리즘을 결합하여 음성 검출기를 개발하였다. 제안된 음성 검출기의 성능을 평가하기 위하여 다양한 환경에서 전술한 음성 샘플과 기준 판정을 이용하여 P_d와 P_f를 측정하였다. 제안된 음성 검출기는 ITU 표준인 G.729 Annex B [5]에 규정된 음성 검출기와 비교되었다. 그 결과가 표 1에 나타나 있다. 표 1에서 음성 구간은 유성음과 무성음 구간으로 세분화되어 각각의 P_d를 별도로 나타내었다. 모든 테스트 조건에서 제안된 음성 검출기는 대부분 G.729B 음성 검출기보다 매우 우수하거나 필적한 만한 성능을 나타내었다.

5 결론

본 연구에서는 decision-directed 파라미터 추정 방식이 likelihood ratio test에 적용되었는데, 이는 보다 매끄러운 *a priori* SNR의 추정치를 제공함으로써 음성 검출기의 성능을 향상시킨다. 제안된 HMM 기반 행오버 방식 또한 주어진 P_f에서 P_d를 증가시키는 것을 모의 실험을 통해 확인하였다. 개발된 VAD는 G.729B 음성 검출기와 비교하였을 때 매우 적은 갯수의 최적화할 파라미터만을 갖으면서도 다양한 환경에서 더 우수한 성능을 나타낸다.

표 1: 다양한 환경에서 제안된 음성 검출기 및 G.729B 음성 검출기의 P_d와 P_f.

		Proposed VAD				G.729B VAD			
Environments		P_d (%)			P_f (%)	P_d (%)			P_f (%)
Noise	SNR	Voiced	UV	Speech	Noise	Voiced	UV	Speech	Noise
Vehicle	5 dB	97.29	97.36	97.30	4.84	97.83	79.21	94.29	46.52
	15 dB	99.77	99.01	99.62	7.19	99.46	93.07	98.24	43.80
	25 dB	100.00	99.34	99.87	7.78	100.00	99.01	99.81	42.79
White	5 dB	87.46	72.28	84.58	1.34	75.62	15.18	64.14	1.01
	15 dB	97.83	93.07	96.93	3.27	93.42	50.83	85.33	1.63
	25 dB	99.69	99.01	99.87	5.17	99.07	85.15	96.43	3.56
Babble	5 dB	92.96	93.40	93.04	23.18	86.38	50.49	79.56	32.63
	15 dB	98.45	98.35	98.43	23.80	94.89	64.36	89.09	25.47
	25 dB	99.77	99.67	99.75	24.75	99.38	88.78	97.37	23.90

참고문헌

[1] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1998, pp. 365–368.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[4] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[5] ITU-T Rec. G.729, Annex B, A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70.

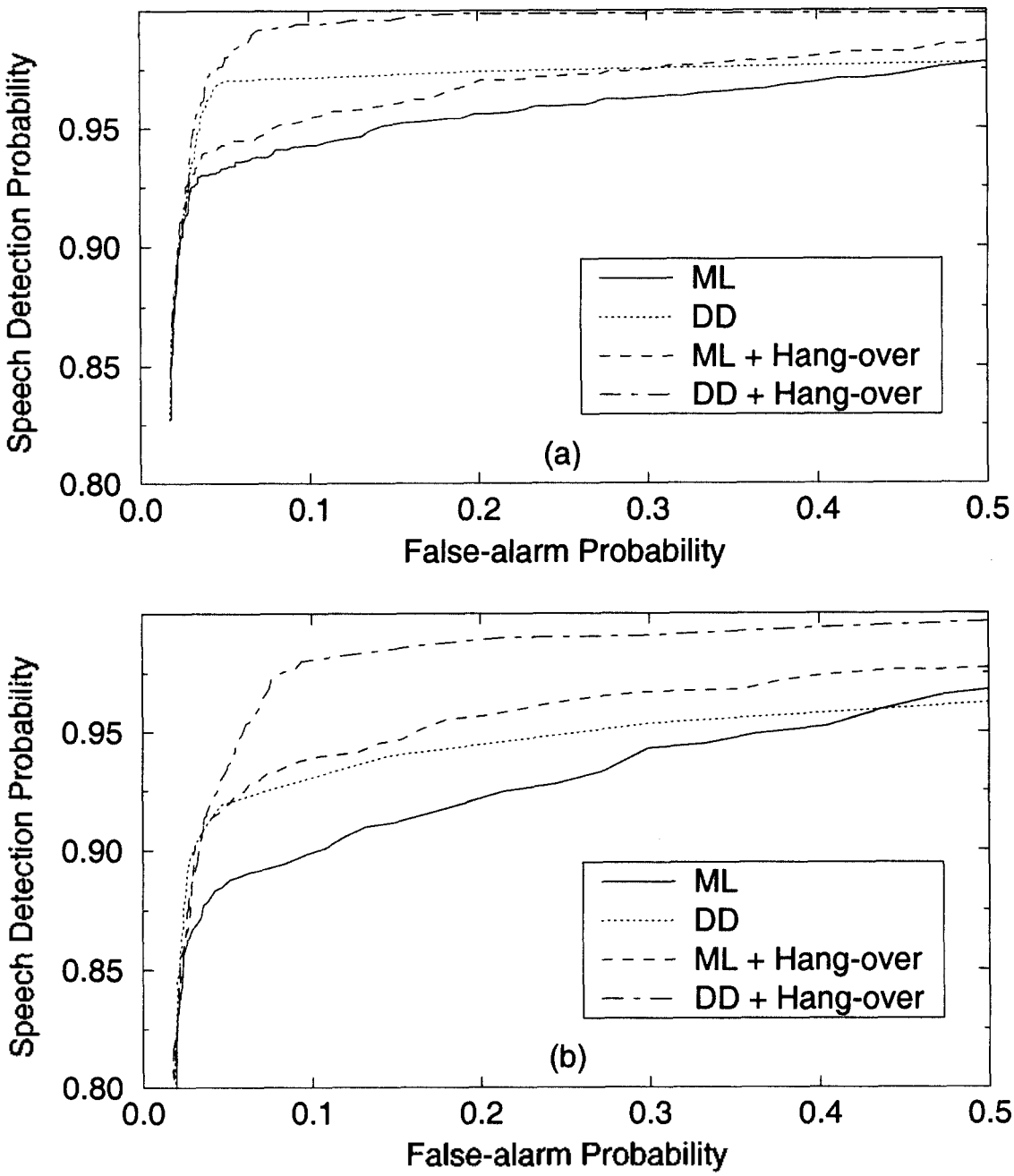

그림 1: 행오버를 적용하거나 적용하지 않았을 때 ML 및 DD 방식에 근거한 판정 규칙들의 동작 특성. (a) 차량 잡음. (b) 백색 잡음.