주파수 차지확률을 이용한 음성검출기 제안 [1]

홍 성 봉, 김 태 영, 김 남 수, 김 태 정

서울대학교 전기공학부

Speech detection using the probability of spectral occupancy

Seong Bong Hong, Tae Young Kim, Nam Soo Kim and Taejeong Kim

School of Electrical Engineering, Seoul National University

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

Abstract

In this paper, we improve statistical-model-based speech detector using the probability that a speech occupies a frequency bin. While the previous method assumes speech energy occupies all the frequency components and use them with equal weights in the likelihood ratio test for speech detection, the proposed method assumes speech energy occupies just some frequency componets and use them with different weights in accordance with the probabilities of spectral occupancy in the test. The probability is iteratively updated for speech frames to contribute to the likelihood ratio test. The proposed method well reflects the characteristic distribution of speech spectrum, and yields better detection performance.

1 Introduction

Speech detection is needed for low bit-rate speech coding applications such as mobile communications and internet telephony services. It is traditionally based on some heuristics like comparsion of the specified feature parameters with given thresholds. Recently, statistical-model-based detectors have been proposed under the assumption that all the discrete Fourier transform (DFT) coefficients of speech and noise are independent, where the decision is made based upon the likelihood ratio test (LRT) using all the frequency bins with equal weights[1][2]. However, the spectrum of a typical speech signal is very locally distributed, and it is usual in voiced speech that the energy is observed dominantly in only several frequency bins. Thus, if we know that some frequency bins have no discernible speech components, it may be advantageous not to use these bins to calculate the LRT since they will just increase the variance of the LRT. This viewpoint leads us to modify the typical LRT using the probability that a speech occupies a frequency bin, which we will define in the following section.

[1]This research is supported by Korea Science and Engineering Foundation under Grant No. 98-0101-03-01-3.

2 Improving the LRT

Two hypotheses in the speech detection problem for an input frame are as follows:

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

H_0: speech absent: $\mathbf{X} = \mathbf{N}$

where $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{X}$ are L-dimensional DFT coefficient vectors of speech, noise, and noisy speech with their kth elements S_k, N_k, and X_k, respectively. The

events of speech absence and presence in the kth frequency bin are denoted respectively by H_0^k and H_1^k. We define β_k to be the probability that the speech occupies the kth frequency bin, *i.e.*, $p(H_1^k|H_1)$. If we assume that the speech signal is independent of the noise and that the DFT coefficients of both the speech and noise are independent Gaussian random variables[3], the probability density functions conditioned on H_0^k and H_1^k are given by

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

$$p(X_k|H_1^k) = \frac{1}{\pi[\lambda_S(k)+\lambda_N(k)]} \exp\left\{-\frac{|X_k|^2}{\lambda_S(k)+\lambda_N(k)}\right\}, \quad (2)$$

where $\lambda_N(k)$ and $\lambda_S(k)$ respectively denote the variances of N_k and S_k. Given (1) and (2) we now have the likelihood ratio for the speech presence in the kth frequency bin as

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

where $\xi_k = \frac{\lambda_S(k)}{\lambda_N(k)}$ and $\gamma_k = \frac{|X_k|^2}{\lambda_N(k)}$ are called the *a priori* and the *a posteriori* SNR's, respectively. With the assumption of statistical independence among separate frequency components, the likelihood ratio for speech detection is given by

$$\Lambda = \frac{p(\mathbf{X}|H_1)}{p(\mathbf{X}|H_0)} = \prod_{k=0}^{L-1} \frac{p(X_k|H_1)}{p(X_k|H_0)}. \quad (4)$$

Let $\Lambda'_k = \frac{p(X_k|H_1)}{p(X_k|H_0)}$(Note the difference between Λ_k and Λ'_k.) Then, it is not difficult to derive

$$\Lambda'_k = \frac{p(X_k|H_1^k)p(H_1^k|H_1) + p(X_k|H_0^k)p(H_0^k|H_1)}{p(X_k|H_0^k)}$$

$$= \beta_k\Lambda_k + (1-\beta_k), \quad (5)$$

where $\beta_k = p(H_1^k|H_1)$.(Note $p(X_k|H_0) = p(X_k|H_0^k)$.) Now, the decision rule is given by

$$\Lambda = \prod_{k=0}^{L-1} \Lambda'_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \quad (6)$$

We see from (5) and (6) that β_k plays the weighting factor in Λ_k's affecting Λ[4]. For example, if $\beta_k = 0$, then $\Lambda'_k = 1$ has no effect on Λ.

3 Parameter adaptation

We now consider how to update β_k's from a speech frame to the next. Since β_k's need to be updated only when the present frame contains a speech signal, we propose an updating scheme as follows:

$$\beta_k(n) = \begin{cases} \alpha\beta_k(n-1) + (1-\alpha)I_k(n) & if\ H_1 \\ \beta_k(n-1) & if\ H_0\ , \end{cases} \quad (7)$$

where α is a forgetting factor, n is the frame index, and $I_k(n)$ is the binary indicator for the speech presence test in the kth frequency bin with a threshold, η_k:

$$I_k(n) = \begin{cases} 1 & if\ \Lambda_k(n) > \eta_k(n-1) \\ 0 & otherwise \end{cases} \quad (8)$$

where $\eta_k(n)$ is a threshold. $\eta_k(n)$ may be chosen as a constant or be given by $\eta_k(n) = \frac{1-p_n(H_1^k)}{p_n(H_1^k)}$ under minimum probability of error criterion, where $p_n(H_1^k) = p(H_1)\beta_k(n)$.

The posteriori probability of H_1 in the nth frame can be estimated by

$$p(H_1|\mathbf{X}(n)) = \frac{p(H_1|\mathbf{X}(n))}{p(H_0|\mathbf{X}(n)) + p(H_1|\mathbf{X}(n))}$$

$$= \frac{\frac{p(H_1)}{p(H_0)}\Lambda(n)}{1+\frac{p(H_1)}{p(H_0)}\Lambda(n)}. \quad (9)$$

Using (7) and (9), we propose a soft-decision-based recursive equation for updating β_k:

$$\hat{\beta}_k(n) = p(H_1|\mathbf{X}(n)) \cdot \{\alpha\hat{\beta}_k(n-1) + (1-\alpha)I_k(n)\}$$

$$+[1-p(H_1|\mathbf{X}(n))] \cdot \hat{\beta}_k(n-1). \quad (10)$$

For adaptation of $\lambda_N(k)$ and $\lambda_S(k)$, we use the methods described in [1] and [3], respectively. $P(H_1)$ and $P(H_0)$ are set to fixed values as in [2].

4 Experimental Results

We compare the proposed detection method with that in [2] by experiments. We made reference decision for clean speech material(5 speakers) of 60 s long by labelling manually at every 20 ms frame. We used HMM-based hang-over technique in [2] to refine the decesion on frame. We applied the two algorithms to

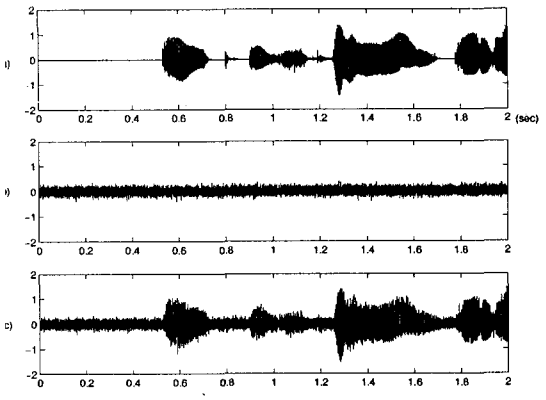

Figure 1: (a) Clean speech samples (b) additive white noise (c) speech sample corrupted by additive white noise(10dB SNR)

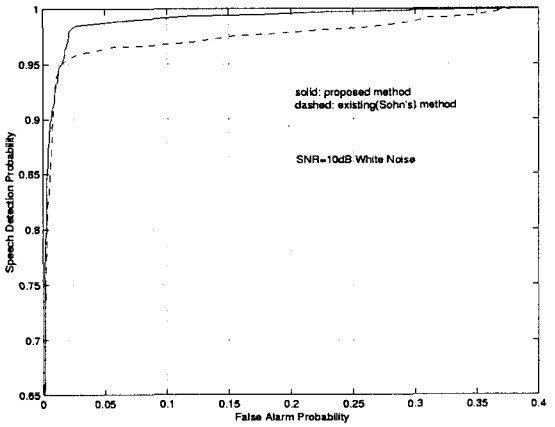

Figure 2: The ROC's of the proposed(solid) and the existing(dashed) schemes: 10dB SNR white noise case

speech signal samples with additive white noise from NOISEX-92 data base at 10dB SNR shown in Figure 1. The receiver operating characteristics (ROC's) in Figure 2 obtained by simulation prove the merit of the proposed model-based detector. The merit of the proposed method is verified in various noise environment such as pink noise and f16 noise environment as shown in Figure 3 and Figure 4.

5 Conclusions

A new speech spectrum model is proposed for statistical speech detection by taking into account the localization characteristic of its spectrum distribution. Such spectral characteristic is represented by a set of probabilities that a speech occupies frequency bins. The probabilities are updated sequentially for each speech frame by a recursive rule. Using these parameters, the proposed detector makes a decision through local soft decision for each frequency bin and then for each frame. Experiments show that the proposed method improves the detection performance in various environmental conditions.

References

[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] Hong, S. B., "Statistical Model Based VAD with Speech Presence Tracking," M. S. Thesis, Seoul National University, 2000.

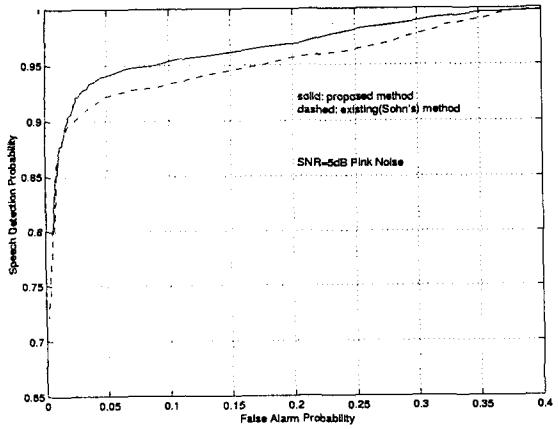

Figure 3: The ROC's of the proposed(solid) and the existing(dashed) schemes: 5dB SNR pink noise case

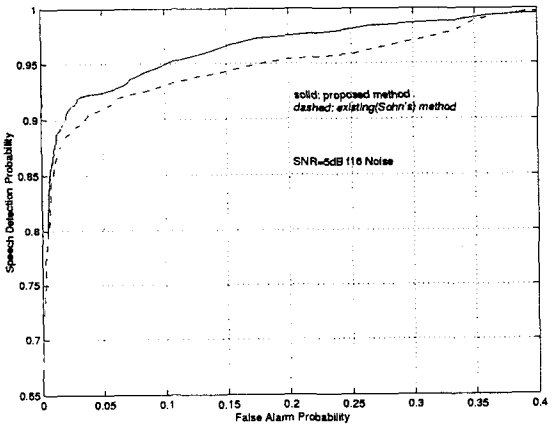

Figure 4: The ROC's of the proposed(solid) and the existing(dashed) schemes: 5dB SNR f16 noise case