A Uniformly Most Powerful Test for Statistical Model-Based Voice Activity Detection

Keun Won Jang[1], *Dong Kook Kim*[1], *Joon-Hyuk Chang*[2]

[1]Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea
[2]Department of Electronic Engineering, Inha University, Incheon, Korea
jsource@hanmail.net, dkim@chonnam.ac.kr, changjh@inha.ac.kr

10.21437/Interspeech.2007-727

Abstract

This paper presents a new voice activity detection (VAD) method using the Gaussian distribution and a uniformly most powerful (UMP) test. The UMP test is employed to derive the new decision rule based on likelihood ratio test (LRT). This method requires the Rayleigh distribution for the magnitude of the noisy spectral component and the adaptive threshold estimated from the noise statistics. Experimental results show that the proposed VAD algorithm based on UMP test outperforms the conventional scheme.

Index Terms: speech enhancement, voice activity detection, speech coding

1. Introduction

Voice activity detection (VAD) refers to the classical problem of distinguishing active speech from nonspeech. It has been used for various applications such as speech coding, speech enhancement, and echo cancelation. Recently, many statistical model-based VAD algorithms in which the likelihood ratio test (LRT) is applied to a set of hypotheses have been developed [1]-[5]. These methods adopt the statistical models which operate in the discrete Fourier transform (DFT) domain. Usually the distributions of noisy speech and noise spectra are assumed to be complex Gaussians [1]. Chang et. al., utilized the complex Laplacian and Gamma probability density functions (pdfs) to model the distributions of noisy speech and noise spectra, and showed that VAD based on these distributions was better than that based on the complex Gaussian distribution [2], [3]. Furthermore, Shin et. al., applied the generalized gamma distribution to model the distribution of the clean speech spectrum and showed that it outperformed the algorithms based on other statistical models [4]. Also Davis et. al., proposed the statistical VAD method that makes no assumption about the distributions of the speech spectra [5]. This scheme incorporates a low-variance spectrum estimate and a statistical detection mechanism and removes the need to estimate the undesired *a priori* signal-to-noise ratio (SNR).

In this paper, we propose a novel VAD algorithm based on the Gaussian distribution and uniformly most powerful (UMP) test. We apply the UMP test to derive the new decision rule which requires the Rayleigh distribution for the magnitudes of noisy speech and noise spectra. The proposed method depends only on the magnitude of the noisy spectral component and the threshold based on the estimated noise statistics, which is computationally efficient in its implementation. Experimental results indicate that the proposed VAD algorithm based on UMP test shows better results compared with the conventional algorithm based on the Gaussian distribution.

2. Likelihood Ratio Test for VAD

In this section, we briefly review the statistical model-based VAD using the Gaussian distribution [1]. We assume that a noise n is added to a speech signal s, with their sum being denoted by x. Given two hypotheses H_0 and H_1, which, respectively, indicate speech absence and presence, it is assumed that

$$\begin{aligned} H_0\text{: speech absent} \;:\; \mathbf{X}(t) &= \mathbf{N}(t) \\ H_1\text{: speech present} \;:\; \mathbf{X}(t) &= \mathbf{N}(t)+\mathbf{S}(t) \end{aligned} \tag{1}$$

in which, $X(t) = [X_0(t), X_1(t), \ldots, X_{M-1}(t)]^T$, $N(t) = [N_0(t),\ N_1(t),\ \ldots,\ N_{M-1}(t)]^T$ and $S(t) = [S_0(t), S_1(t), \ldots, S_{M-1}(t)]^T$ are the DFT coefficients at frame t of the noisy speech, noise, and clean speech, respectively. Also, M is total number of frequency bins, and T is transposed matrix. Assuming that each spectral component of speech and noise signals has complex Gaussian distribution, the distributions of the noisy spectral components conditioned on both hypotheses are obtained:

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

$$p(X_k|H_1) = \frac{1}{\pi[\lambda_{s,k}+\lambda_{n,k}]} \exp\left\{-\frac{|X_k|^2}{\lambda_{s,k}+\lambda_{n,k}}\right\} \tag{3}$$

where $\lambda_{s,k}$ and $\lambda_{n,k}$denote the variances of S_k and N_k, respectively. The likelihood ratio (LR) for the kth frequency bin is given by

$$\Lambda_k \equiv \frac{p\left(X_k|H_1\right)}{p\left(X_k|H_0\right)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \tag{4}$$

where $\xi_k = \lambda_{s,k}/\lambda_{n,k}$ and $\gamma_k = |X_k|^2/\lambda_{n,k}$, and they are called the *a priori* SNR and *a posteriori* SNR, respectively [6]. The decision rule is constructed as the geometric mean of the LRs computed for the individual frequency bins such that

$$\log\Lambda = \frac{1}{M}\sum_{k=0}^{M-1}\log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5}$$

with η denoting the threshold of detection.

3. Likelihood ratio test based on UMP test

3.1. VAD Based on UMP Test

In this section, we present a new statistical model-based VAD method using the UMP test. To find the new decision rule, the

decision statistic in (4) can be rewritten as

$$\Lambda_k = \frac{\lambda_{n,k}}{\lambda_{s,k}+\lambda_{n,k}} \exp\left\{ \frac{\lambda_{s,k}|X_k|^2}{\lambda_{n,k}[\lambda_{s,k}+\lambda_{n,k}]} \right\} \quad (6)$$

where the value of $\lambda_{s,k}$ is unknown parameter, although a priori we know that $\lambda_{s,k} > 0$. We assume that the noise variance, $\lambda_{n,k}$ is known, which can be estimated during periods of non-speech activity [6]. Then, the LRT is to decide H_1 if

$$\frac{\lambda_{n,k}}{\lambda_{s,k}+\lambda_{n,k}} \exp\left\{ \frac{\lambda_{s,k}|X_k|^2}{\lambda_{n,k}[\lambda_{s,k}+\lambda_{n,k}]} \right\} > \eta \,. \quad (7)$$

Since it is known that $\lambda_{s,k} > 0$ and $\lambda_{n,k} > 0$, after taking the logarithm and some manipulation, we have finally the new decision rule as follows:

$$\begin{aligned} |X_k|^2 &> \frac{\lambda_{s,k}+\lambda_{n,k}}{\lambda_{s,k}/\lambda_{n,k}} \left(\ln\eta - \ln\left[\frac{\lambda_{s,k}+\lambda_{n,k}}{\lambda_{n,k}} \right] \right) \quad (8) \\ &= \eta_k^{'} \end{aligned}$$

which may be also rewritten in the simple form as

$$|X_k| > \sqrt{\eta_k^{'}} \quad (9)$$

where $\sqrt{\eta_k^{'}}$ is the new threshold in the kth spectral bin which depends on the variances, $\lambda_{s,k}$ and $\lambda_{n,k}$.

Clearly, rather than comparing a test statistic like the LR to a threshold, the presented VAD method compares the magnitude of the noisy spectral component to a threshold $\sqrt{\eta_k^{'}}$. But the key question is whether we can implement this detector without the estimate of the value of $\lambda_{s,k}$. The $|X_k|$ does not depend on $\lambda_{s,k}$ but it appears that the threshold $\sqrt{\eta_k^{'}}$ does. If $|X_k|$ is large, then the speech signal is probably present. But to prevent the noise from causing the large $|X_k|$, we have to adjust the threshold $\sqrt{\eta_k^{'}}$ to control the probability of false alarm (FA), P_{FA}, with larger threshold values reducing P_{FA}. Over all possible detectors that have a given P_{FA}, this detector is optimal in that it yields the highest detection probability, P_D, for any value of $\lambda_{s,k}$, as long as $\lambda_{s,k} > 0$, This type of test, when it exists, is called a UMP test [7]. But the UMP test is not adopted in the case of $-\infty < \lambda_{s,k} < \infty$, which means that two-sided problems never produce the UMP tests. For the UMP test to exist, the parameter test must be one-sided test [7]. After finding the threshold based on the P_{FA}, the decision rule is established from the arithmetic mean of the magnitude and threshold of kth frequency bins, respectively, which is given by

$$\frac{1}{M}\sum_{k=1}^{M} |X_k| \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{M}\sum_{k=1}^{M} \sqrt{\eta_k^{'}} \,. \quad (10)$$

In order to determine the detection performance, the distribution for the magnitude, $|X_k|$, of the noisy spectral component should be found. It is well known that the magnitude of kth frequency bin of the speech signal, $\bar{X}_k = |X_k|$, has a Rayleigh distribution as follows [7], [8]:

$$P(\bar{X}_k) = \frac{2\bar{X}_k}{\lambda_{x,k}} \exp\left(-\frac{\bar{X}_k^2}{\lambda_{x,k}} \right) U(\bar{X}_k) \quad (11)$$

where $U(\bar{X}_k)$ denotes the unit-step function. With the Rayleigh pdf, the distributions of the noisy spectral components conditioned on H_1 and H_0 are given by

$$p(\bar{X}_k|H_0) = \frac{2\bar{X}_k}{\lambda_{n,k}} \exp\left(-\frac{\bar{X}_k^2}{\lambda_{n,k}} \right) U(\bar{X}_k) \quad (12)$$

$$p(\bar{X}_k|H_1) = \frac{2\bar{X}_k}{\lambda_{s,k}+\lambda_{n,k}} \exp\left(-\frac{\bar{X}_k^2}{\lambda_{s,k}+\lambda_{n,k}} \right) U(\bar{X}_k) \quad (13)$$

The key problem now becomes one of determining the optimal threshold in some best manner. Specifically, we describe how to find the threshold below.

3.2. Decision of Threshold

Since the threshold in (9) and (10) depends on the variances $\lambda_{s,k}$ and $\lambda_{n,k}$, it is very important to control the threshold according to the noise statistics to improve the detection performance. To find the threshold, we can use the false-alarm probability, P_{FA}, [5], [7]. A false alarm is realized when the magnitude of the spectral component is larger than the threshold, given the null hypothesis is present. By considering the assumed Rayleigh pdf under the null hypothesis, the false-alarm probability is given by

$$\begin{aligned} P_{FA} &= Pr\left\{ |X_k| > \sqrt{\eta_k}^{'} : H_0 \right\} \quad (14) \\ &= Pr\left\{ \frac{|X_k|}{\sqrt{\lambda_{n,k}}} > \frac{\sqrt{\eta_k}^{'}}{\sqrt{\lambda_{n,k}}} : H_0 \right\} \\ &= Pr\left\{ \frac{|X_k|^2}{\lambda_{n,k}} > \frac{\eta_k^{'}}{\lambda_{n,k}} : H_0 \right\} \\ &= Q_{\mathcal{X}_2^2}\left(\frac{\eta_k^{'}}{\lambda_{n,k}} \right) \end{aligned}$$

where $Q_{\mathcal{X}_2^2}(\frac{\eta_k^{'}}{\lambda_{n,k}})$ is the right-tail probability of a chi-squared pdf with 2-degrees of freedom [7], which is defined as

$$Q_{\mathcal{X}_2^2}\left(\frac{\eta_k^{'}}{\lambda_{n,k}} \right) = \exp\left(-\frac{\eta_k^{'}}{2\lambda_{n,k}} \right) U(\eta_k^{'}) \,. \quad (15)$$

It is clear that the threshold $\sqrt{\eta_k}^{'}$ can be chosen from the false-alarm probability, P_{FA}. Assuming $\sqrt{\eta_k}^{'} = \eta_k$, and solving for η_k, the threshold is given as follows:

$$P_{FA} = \exp\left(-\frac{\eta_k^2}{2\lambda_{n,k}} \right) U(\eta_k) \quad (16)$$

so that

$$\eta_k = \sqrt{-2\lambda_{n,k}\ln(P_{FA})} \quad (17)$$

which is independent of $\lambda_{s,k}$. According to (17), we can adjust the threshold by the noise statistics and the false-alarm probability.

3.3. Discussion

The proposed UMP-based VAD approach is similar to that in [5]. [5] uses a SNR measure for the statistical VAD method without assumption about the distribution of the speech while our scheme uses the magnitude derived from the LRT and the UMP test. Both methods requires the adaptive threshold which

depends only on the noise statistics and the false-alarm probability.

In order to implement the proposed VAD algorithm for various noise environments, several factors such as the magnitude, threshold, and the noise power need to be considered [5]. It is well known that the smoothing parameter using a forgetting factor and a hang-over scheme improves the detection performance [1], [5]. Therefore, the magnitude and threshold in (10) are modified to incorporate the forgetting factor schemes such that

$$|\hat{X}_k(t)| = (1-\zeta_{|X_k|}) \cdot |X_k(t)| + \zeta_{|X_k|} \cdot |\hat{X}_k(t-1)| \quad (18)$$

$$\hat{\eta}_k(t) = (1-\zeta_{\eta_k}) \cdot \eta_k(t) + \zeta_{\eta_k} \cdot \hat{\eta}_k(t-1) \quad (19)$$

where $\zeta_{|X_k|}$ and ζ_{η_k} are the forgetting factors for the magnitude and threshold, respectively.

4. Experimental results

In order to evaluate the performance of the proposed algorithms, we investigated the speech detection and false-alarm probabilities (P_D and P_{FA}) for each VAD algorithm. We define P_D as the ratio of correct speech decisions to the hand-marked speech frames, while P_{FA} as that of false speech decisions to the hand-marked noise frames. To obtain P_D and P_{FA}, we made reference decisions for a clean speech material of 456 s long by labeling manually at every 10 ms frame. The percentage of the speech sounds is 58.2%, which consists of 44.8% voiced sounds and 13.4% unvoiced sounds. To make a noisy signal, we added the babble and white noises from NOISEX-92 database [9] to the clean speech waveform with varying SNR. The VAD test was performed for each 10 ms frame.

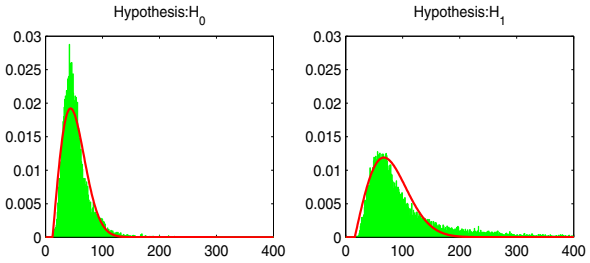

Figure 1: Normalized histograms of the magnitude, $|X_k|$, for both hypotheses in babble noise environment for 8th frequency bin (SNR = 5 dB).

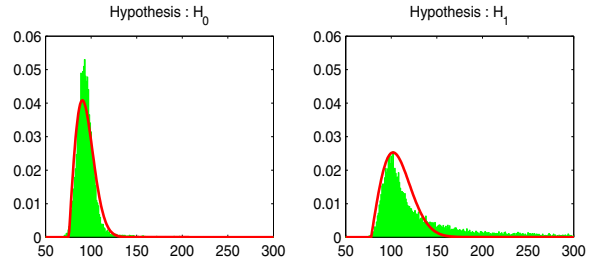

Figure 2: Normalized histograms of the magnitude, $|X_k|$, for both hypotheses in white noise environment for 8th frequency bin (SNR = 5 dB).

To validate the Rayleigh distribution assumption, the magnitude of $|X_k|$ for 8th frequency bin was experimentally calculated in the babble and white noise environments at a variety of SNRs. The magnitude was found to closely follow a Rayleigh distribution in a range of noise environments. The normalized histograms with a Rayleigh fit for both noise environments at SNR = 5 dB are displayed in Figs. 1 and 2, respectively. The left side denotes the histogram of H_0 while the right side denotes H_1. From Figs. 1 and 2, it is noted that the distribution of H_1 has large value and more spread than H_0. Based on these distributions, We observe that the distributions conditioned on both hypotheses overlapped each other such that it degrades the detection performance.

An example of the magnitude, $|X_k|$, and threshold for a 14 s speech waveform under the babble noise condition at SNR = 5 dB is displayed in Fig. 3. The final VAD detection is also shown in conjunction with the noisy speech samples. The arithmetic mean over the frequency of the magnitude was found and compared to the threshold, resulting in the final VAD results.

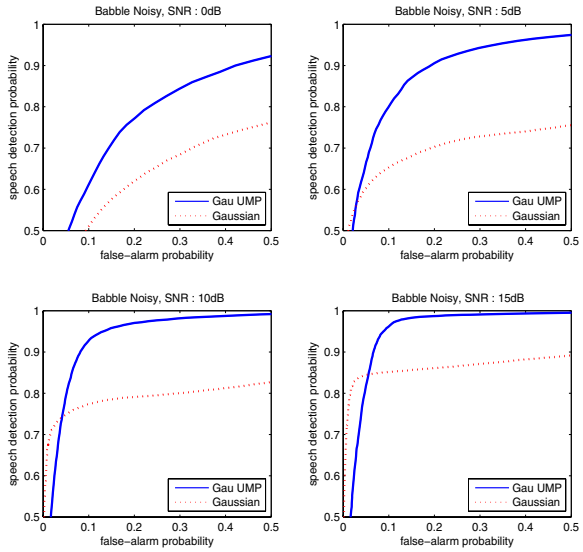

Figure 4: ROC curves for the Gaussian and UMP test approaches with the babble noise at 0 dB, 5 dB, 10 dB, 15 dB SNR, respectively.

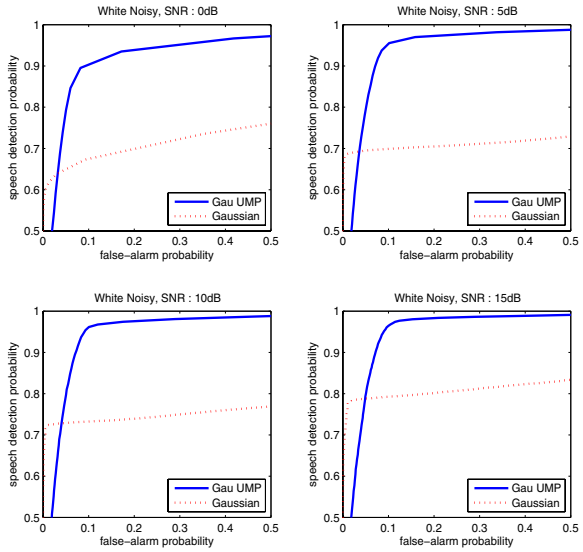

Figure 5: ROC curves for the Gaussian and UMP test approaches with the white noise at 0 dB, 5 dB, 10 dB, 15 dB SNR, respectively.

The ROC curves of the proposed method are shown in Figs.

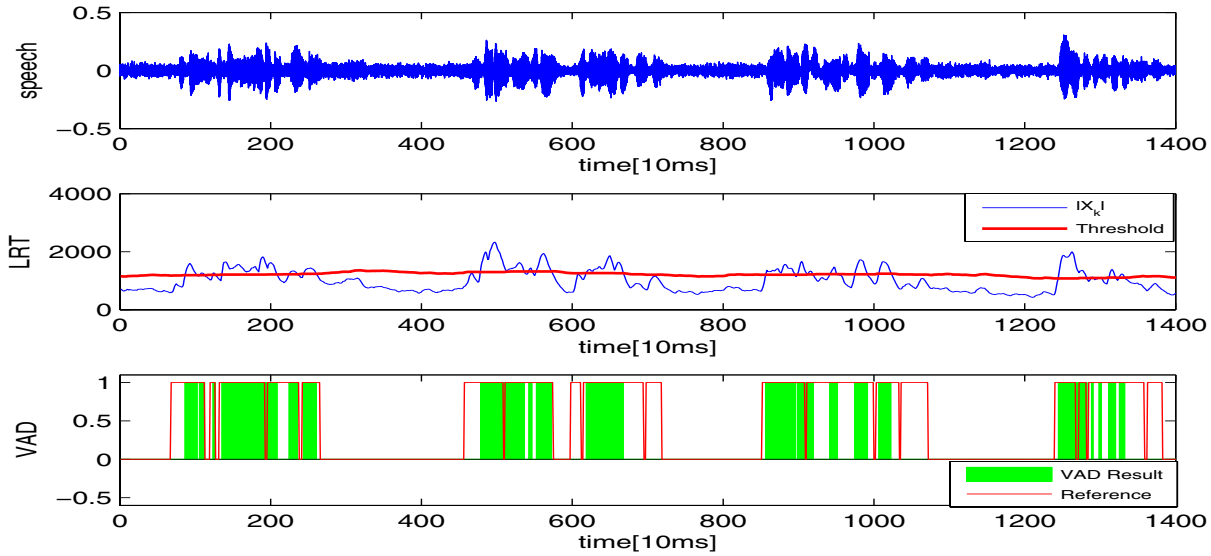

Figure 3: VAD results for the proposed method with the babble noise (SNR = 5 dB). The magnitude $|X_k|$, threshold and corresponding noisy speech waveform are plotted for easy comparison.

4 and 5 and where the VAD algorithms were applied to the speech data corrupted by the babble and white noises at a variety of SNRs (0, 5, 10, and 15 dB), respectively. VAD algorithms were implemented with the forgetting factor scheme while keeping $\zeta_{|X_k|} = 0.9$ and $\zeta_{\eta_k} = 0.2$. In the case of the babble noise, which is shown in Fig. 4, the UMP-based VAD algorithm outperformed the Gaussian approach. It was also observed that the difference in detection accuracy became larger as the SNR gets higher. On the other hand, the Gaussian-based VAD algorithm yielded a slightly better performance than the UMP-based method only when $P_{FA} < 0.06$ at SNR $= 15$ dB. From Fig. 5, which shows the results for the white noise, the UMP-based algorithm yielded a performance superior to the Gaussian method when $P_{FA} > 0.06$ at all SNRs, Also the performance difference became larger as the SNR increased. From the experimental results, it is evident that the proposed method based on the UMP yields higher performance than the Gaussian-based algorithm in the most noisy conditions.

5. Conclusion

We have presented a novel VAD algorithm based on the Gaussian distribution and UMP test to detect the speech or nonspeech from the input noisy signal. We have applied the UMP test to find the new decision rule based on LRT as well as the Rayleigh distribution of the magnitude of the noisy spectral components. The efficient decision rule is derived to compare the arithmetic mean of the magnitude with the threshold which depends only on the noise statistics and the false-alarm probability which is very effective in its computation. It has been found that the VAD algorithm based UMP test outperformed those based on widely used Gaussian distribution in a number of experiments.

6. References

[1] Sohn, J. S., Kim, N. S. and Sung, W. Y., "A Statistical Model-Based Voice Activity Detection," IEEE Signal Processing Lett., Vol. 6(1):1-3, 1999.

[2] Chang, J. -H. and Kim, N. S., "Voice Activity Detection Based on Complex Laplacian Model," IEE Electronics Lett., Vol. 39(7):632 - 634, 2003.

[3] Chang, J. -H., Kim, N. S. and Mitra, S. K., "Voice Activity Detection Based on Multiple Statistical Models," IEEE Trans. Signal Processing, Vol. 54(6):1965 - 1976, 2006.

[4] Shin, J. W., Chang, J. -H., Yun, H. S. and Kim, N. S., "Voice Activity Detection Based on Generalized Gamma Distribution," Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc., Vol. 1:I-781-I-784, 2005.

[5] Davis, A., Nordholm, S. and Togneri, R., "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," IEEE Trans. Audio, Speech, and Lang. Proc., Vol. 14(2):412-424, 2006.

[6] Ephraim, Y. and Malah, D., "Speech Enhancement Using A Minimum Mean-square Error Short-time Spectral Amplitude Estimator," IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-32:1109-1121, 1984.

[7] Kay, S. M., Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory, Prentice Hall, 1998.

[8] Vary, P. and Martin, R., Digital Speech Transmission: Enhancement, Coding and Error Concealment, John Wiley & Sons Inc, 2006.

[9] Varga, A. and Steeneken, H.J.M., "Assessment for Automatic Speech Recognition: II.NOISEX-92: A Database and An Experiment to Study The Effect of Additive Noise on Speech Recognition Systems," Speech Communication, Vol. 12(3):247-251, 1993.