Likelihood Ratio Test with Complex Laplacian Model for Voice Activity Detection

Joon-Hyuk Chang, Jong-Won Shin and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University Seoul, Korea.
Kwanak P.O.Box 34, Seoul 151-742, Korea,
changjh@snu.ac.kr, jwshin@hi.snu.ac.kr nkim@snu.ac.kr

Abstract

This paper proposes a voice activity detector (VAD) based on the complex Laplacian model. With the use of a goodness-of-fit (GOF) test, it is discovered that the Laplacian model is more suitable to describe noisy speech distribution than the conventional Gaussian model. The likelihood ratio (LR) based on the Laplacian model is computed and then applied to the VAD operation. According to the experimental results, we can find that the Laplacian statistical model is more suitable for the VAD algorithm compared to the Gaussian model.

1. Introduction

Voice activity detector (VAD) has become an indispensable part of the variable-rate speech coding and noisy speech enhancement techniques. Recently, a number of efforts to enhance the performance of the VAD by employing a statistical model have been made with the decision rule being derived from the likelihood ratio test (LRT) applied to a set of hypotheses [1]. Furthermore, these statistical model-based approaches have shown improved performance with the incorporation of soft decision schemes [2], [3]. In most of the conventional VAD algorithms which mainly operate in the discrete Fourier transform (DFT) domain, it is assumed that the distributions of the clean speech and noise spectra are characterized by the complex Gaussian densities. Recently, it has been reported that the DFT coefficients of clean speech and noise are better modeled by the Gamma and Laplacian distributions, respectively [4].

In this paper, we introduce the complex Laplacian model for the purpose of voice activity detection with the DFT coefficients of noisy speech under various noise conditions. In contrast to [4], instead of modeling the clean speech distribution we attempt to directly characterize the noisy speech distribution. We compare the Laplacian model with the conventional Gaussian model by applying the goodness-of-fit (GOF) test under various noisy signal conditions [5], [6]. The GOF test reports that the empirical noisy speech distribution is more closely approximated by the Laplacian model than the Gaussian model. From a number of experiments, the VAD algorithm based on the complex Laplacian model is found to provide better performance compared to that with the Gaussian model.

This research was supported by the Brain Korea 21 project.

2. Statistical Models

2.1. Basic principle

We assume that a noise signal 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

$$H_0 : \text{ speech absent} : \mathbf{X}(t) = \mathbf{N}(t) \quad (1)$$

$$H_1 : \text{ speech present} : \mathbf{X}(t) = \mathbf{N}(t) + \mathbf{S}(t) \quad (2)$$

in which $\mathbf{X}(t) = [X_0(t), X_1(t), ..., X_{M-1}(t)]^T, \mathbf{N}(t) = [N_0(t), N_1(t), ..., N_{M-1}(t)]^T$, and $\mathbf{S}(t) = [S_0(t), S_1(t), ..., S_{M-1}(t)]^T$ are the discrete Fourier transform (DFT) coefficients of the noisy speech, noise and clean speech, respectively.

The above statistical model is completed with a suitable specification of the DFT coefficients' distribution. In this paper, we consider two different probabilistic density functions (pdf's) for the candidate distribution. The first one is the complex Gaussian pdf which is most widely applied to characterize the DFT coefficients distribution in speech analysis [1], [7]. With the Gaussian pdf assumption, the distributions of the noisy spectral components conditioned on both hypotheses are given by

$$p_G(X_k|H_0) = \frac{1}{\pi\lambda_{n,k}} \exp\{-\frac{|X_k|^2}{\lambda_{n,k}}\} \quad (3)$$

$$p_G(X_k|H_1) = \frac{1}{\pi[\lambda_{n,k}+\lambda_{s,k}]} \exp\{-\frac{|X_k|^2}{\lambda_{n,k}+\lambda_{s,k}}\} \quad (4)$$

where $\lambda_{n,k}$ and $\lambda_{s,k}$ denote the variances of N_k and S_k, respectively. The second distribution is the complex Laplacian pdf. Let $X_{k(R)}$ and $X_{k(I)}$ denote the real and

imaginary parts of the DFT coefficient X_k, respectively. Then, according to the complex Laplacian pdf, $X_{k(R)}$ and $X_{k(I)}$ are assumed to be distributed as follows:

$$p(X_{k(R)}) = \frac{1}{\sigma_x}\exp\Big\{-\frac{2|X_{k(R)}|}{\sigma_x}\Big\} \quad (5)$$

$$p(X_{k(I)}) = \frac{1}{\sigma_x}\exp\Big\{-\frac{2|X_{k(I)}|}{\sigma_x}\Big\} \quad (6)$$

where σ_x^2 denotes the variance of X_k. If the real and imaginary parts of X_k are assumed to be independent [8], we can obtain $p(X_k)$ such that

$$p(X_k) = p(X_{k(R)})\cdot p(X_{k(I)}) \quad (7)$$

$$= \frac{1}{\sigma_x^2}\exp\Big\{-\frac{2(|X_{k(R)}|+|X_{k(I)}|)}{\sigma_x}\Big\}. \quad (8)$$

Using (8), the distributions of the noisy DFT coefficients are described as follows:

$$p_L(X_k|H_0) = \frac{1}{\lambda_{n,k}}\exp\{-\frac{2(|X_{k(R)}|+|X_{k(I)}|)}{\sqrt{\lambda_{n,k}}}\} \quad (9)$$

$$p_L(X_k|H_1) = \frac{1}{\lambda_{n,k}+\lambda_{s,k}} \cdot\exp\{-\frac{2(|X_{k(R)}|+|X_{k(I)}|)}{\sqrt{\lambda_{n,k}+\lambda_{s,k}}}\}. \quad (10)$$

2.2. Goodness-of-Fit Test

For a successful VAD operation, we must select a model that fits better to the given noisy speech spectra. For this, we carry out a statistical fitting test for the noisy spectral components conditioned on both H_0 and H_1 under various noise conditions. This is distinguished from the approach proposed by Martin [4] in which the minimum mean-square error (MMSE) estimator is obtained based on a set of clean speech distributions. For the pdf selection, we apply the *Kolomogorov-Sriminov* (*KS*) test which is well-known as the goodness-of-fit (GOF) test. Due to the incorporation of *KS* test, a reliable survey of each statistical assumption is guaranteed.

Let $\mathbf{X} = [X_0, X_1, \cdots, X_{N-1}]^T$ be a vector which represents the DFT coefficients of the noisy speech. The *KS* test compares the empirical cumulative distribution function (CDF) F_X to a given distribution function F. The empirical CDF is defined by [6]

$$F_X(z) = \begin{cases} 0, & z < X_{(1)} \\ \frac{n}{N}, & X_{(n)} \le z < X_{(n+1)}, \quad n = 0, 1, \cdots, N-1 \\ 1, & z \ge X_{(N)} \end{cases} \quad (11)$$

where $X_{(n)}, n = 0, \cdots, N-1$ are the order statistics of the data $\mathbf{X}$. In order to compute there order statistics, we sort and order the elements of $\mathbf{X}$ so that $X_{(0)}$ is the smallest element of $\mathbf{X}$ and $X_{(N-1)}$ is the largest [5].

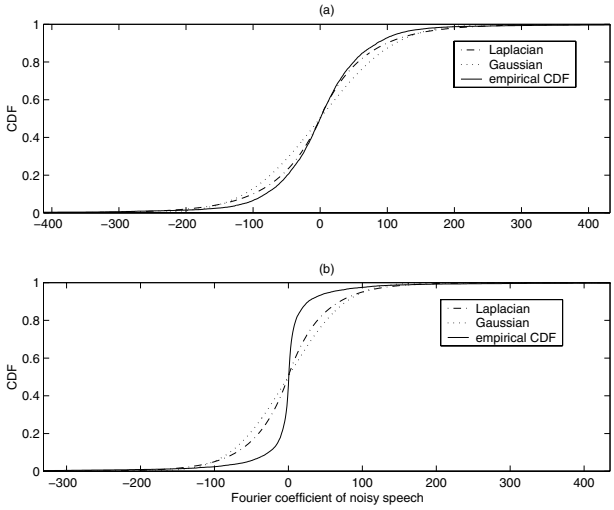

Figure 1: Comparison of Laplacian and Gaussian CDF of noisy speech spectra (real part) at H_1 (a) white noise (SNR=10dB) (b) vehicular noise (SNR=10dB).

Speech material of 64 sec duration was collected from different 4 male and 4 female speakers. For simulating noisy environments, the white and vehicular noises from the NOISEX-92 database were added to the clean speech signal at SNR = 10 dB. Using the collected data, sample mean and variance were calculated and embedded into the given Laplacian and Gaussian distributions, respectively. Fig. 1 shows the comparison of empirical CDF and given functions, respectively. From the comparison, we can see that the curve of the Laplacian CDF is closer to the empirical CDF than that of the Gaussian CDF under both the white and vehicular noise conditions.

For defining the distance measure between the empirical CDF and the given distribution, we use the *KS* statistic [6]. The *KS* test statistic T is defined by

$$T = \max_i |F_X(X_i) - F(X_i)| \quad (12)$$

where the distance is given by the maximum difference between $F_X(\cdot)$ and $F(\cdot)$ evaluated at the sample points $\{X_i\}$. When testing the data against several distributions, the distribution that yields the smallest *KS* statistic is considered the one that fits best to the given the data. Table 1 shows the *KS* statistics evaluated under various conditions. From Table 1, we can see that the *KS* statistic T of the Laplacian model is smaller than that of the Gaussian model in all the tested noisy conditions. Therefore, we can conclude that the Laplacian model is more adequate for characterizing the DFT coefficients of noisy speech than the Gaussian model.

Table 1: Results of *Kolomogorov-Smirnov* test for the DFT coefficients of noisy speech conditioned on various noise environments. *G* and *L* denotes the Gaussian and Laplaician distribution, respectively.

noise		white			vehicular			babble		
SNR(dB)		5	10	15	5	10	15	5	10	15
H_1	$G; X_{k(R)}$	0.043	0.078	0.129	0.211	0.223	0.231	0.129	0.165	0.198
	$L; X_{k(R)}$	0.031	0.025	0.068	0.164	0.177	0.186	0.071	0.107	0.145
	$G; X_{k(I)}$	0.044	0.081	0.134	0.214	0.225	0.232	0.142	0.173	0.203
	$L; X_{k(I)}$	0.028	0.026	0.073	0.164	0.178	0.187	0.080	0.116	0.149
H_0	$G; X_{k(R)}$	0.045	0.052	0.063	0.238	0.270	0.311	0.149	0.127	0.136
	$L; X_{k(R)}$	0.024	0.024	0.023	0.189	0.237	0.277	0.088	0.167	0.078
	$G; X_{k(I)}$	0.051	0.059	0.071	0.243	0.275	0.325	0.153	0.127	0.134
	$L; X_{k(I)}$	0.019	0.016	0.021	0.243	0.237	0.278	0.093	0.067	0.075

3. Decision Rule based on Likelihood Ratio Test

Based on the assumed statistical models, we compute the likelihood ratio for the kth frequency bin such that [1]

$$\Lambda_k \equiv \frac{p(X_k|H_1)}{p(X_k|H_0)}. \tag{13}$$

The decision rule for the VAD is established from the geometric mean of the likelihood ratios for the individual frequency bins, which is given by

$$\log \Lambda = \frac{1}{M} \sum_{k=0}^{M-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{14}$$

where η is a threshold.

The likelihood ratio with the conventional Gaussian distribution for H_0 and H_1 is given by

$$\Lambda_k^{(G)} = \frac{p_G(X_k|H_1)}{p_G(X_k|H_0)} \tag{15}$$

$$= \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\} \tag{16}$$

where $\xi_k = \lambda_{s,k}/\lambda_{n,k}$ and $\gamma_k = |X_k|^2/\lambda_n$, which are called the *a priori* and *a posteriori* SNR's, respectively [1]. On the other hand, the likelihood ratio computed based on the Laplacian model is described as follows:

$$\Lambda_k^{(L)} = \frac{p_L(X_k|H_1)}{p_L(X_k|H_0)} \tag{17}$$

$$= \frac{1}{1+\xi_k} \exp\left\{2\big(|X_{k(R)}|+|X_{k(I)}|\big)\Big(\frac{|X_k| - \sqrt{\lambda_{n,k}}}{\sqrt{|X_k|\lambda_{n,k}}}\Big)\right\}. \tag{18}$$

Here, the success or failure of the VAD depends on not only the statistical model but also the reliable estimation of noise power $\{\lambda_{n,k}(t)\}$ and speech power $\{\lambda_{s,k}(t)\}$. In this paper, we follow the noise and speech power estimation procedure proposed in [1].

4. Experimental Results

To compare the performance of Laplacian model with that of Gaussian model, we investigate the speech detection and false-alarm probabilities (P_d and P_f) for each statistical model. To obtain P_d and P_f, we made reference decisions on a clean speech material of 32 sec long by labeling manually at every 10 ms frame. The percentage of the hand-marked speech frames is 36.32% which consists of 19.87% voiced sounds and 16.45% unvoiced sounds frames.

The receiver operating characteristics (ROC's), showing the trade-off between P_d and P_f, of the two statistical models are shown in Fig. 2 where the VAD algorithm was applied to the noisy speech samples corrupted by the white and vehicular noise sources from the NOISEX-92 database at 5 dB SNR. As shown in Fig. 2, the complex Laplacian model-based decision rule performs better than the complex Gaussian-based one only if P_d is allowed to lie in the normal range (upper than 90 %). In order to further evaluate the performance of the presented VAD algorithm, we measured P_d and P_f in various noisy conditions using the aforementioned speech materials and referred decision rules. The results are summarized in Table 2 where the hand-marked speech materials are subdivided into the voiced and unvoiced sounds. In most of the conditions except for the vehicular noise at high SNR and the babble noise at low SNR, the proposed Laplacian model outperformed the Gaussian model.

5. Conclusions

We have presented an approach to incorporate the complex Laplacian distribution to the VAD. We analyze the statistical properties of the Laplacian distribution compared with the Gaussian distribution by the use of GOF test. The proposed complex Laplacian model based-VAD shows better performance than the Gaussian model-based one in various environments.

Table 2: P_d's and P_f's of the Proposed Laplacian and Gaussian model VAD's For Various Environmental Conditions

		Laplacian				Gaussian			
environments		P_d			P_f	P_d			P_f
noise	SNR (dB)	Voiced	UV	speech	noise	Voiced	UV	speech	noise
white	5	91.23	67.43	84.87	13.20	89.07	70.58	84.12	13.06
	10	94.46	77.94	90.03	15.99	93.76	81.72	90.54	16.09
	15	98.23	90.54	96.17	18.66	96.15	86.97	93.69	24.83
	20	99.92	98.94	99.66	19.65	99.84	99.15	99.66	28.42
vehicle	5	97.92	93.90	96.84	5.40	97.76	96.42	97.40	8.48
	10	99.30	97.50	98.70	9.93	99.30	90.30	96.90	8.88
	15	99.92	99.36	99.77	10.12	99.92	92.85	98.02	9.32
	20	99.96	99.57	99.98	10.50	100.00	98.31	99.54	9.87
babble	5	94.46	84.24	91.72	15.95	94.00	82.77	90.99	12.48
	10	96.38	87.39	93.97	15.94	95.53	83.19	92.22	11.40
	15	97.61	89.71	95.49	14.64	95.92	84.66	92.90	15.83
	20	99.61	95.61	98.42	15.96	98.92	98.52	98.81	19.21

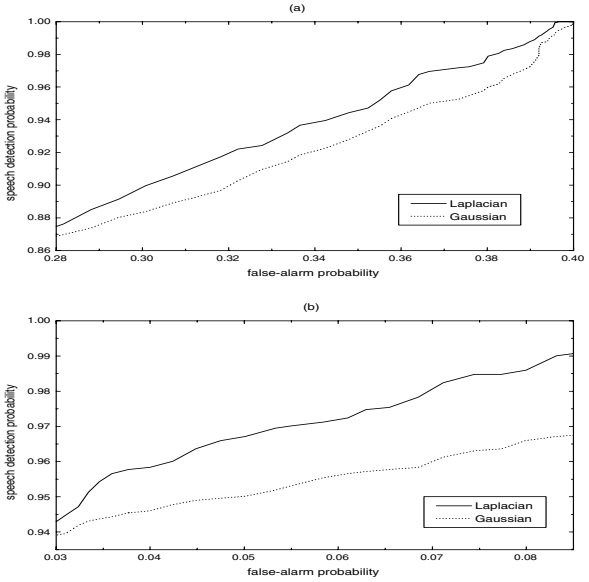

Figure 2: Receiver operating characteristics of VAD with the Laplacian and Gaussian model at 5 dB SNR (a) white noise (b) vehicular noise

6. References

[1] 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.

[2] N. S. Kim and J. -H. Chang, " Spectral enhancement based on global soft decision ," *IEEE Signal Processing Letters*, Vol. 7, No. 5, pp. 108-110, May 2000.

[3] J. -H. Chang and N. S. Kim, "Speech enhancement : new approaches to soft decision," *IEICE Trans. Inf. and Syst.*, Vol. 27, E84-D, pp. 1231-1240, Sep. 2001.

[4] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL. , May 2002.

[5] A. G. Glen, L. M. Leemis, and D. R. Barr, "Order statistics in goodness-of-fit testing," *IEEE Trans. Reliability.*, Vol. 50, No. 2, June 2001.

[6] R. C. Reininger and J. D. Gibson, "Distributions of the two dimensional DCT coefficients for images," *IEEE Trans. Commnuications.*, Vol. Com-31, No. 6, June 1983.

[7] 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. 32, No. 6, pp. 1109-1121, Dec. 1984.

[8] D. R. Brillinger, *Time Series: Data Analysis and Theory.* New York: Holden-Day, 1981.