LETTER

Voice Activity Detection Based on Generalized Normal-Laplace Distribution Incorporating Conditional MAP

Ji-Hyun SONG[†a)], ***Nonmember and*** **Sangmin LEE**[†], ***Member***

SUMMARY In this paper, we propose a novel voice activity detection (VAD) algorithm based on the generalized normal-Laplace (GNL) distribution to provide enhanced performance in adverse noise environments. Specifically, the probability density function (PDF) of a noisy speech signal is represented by the GNL distribution; the variance of the speech and noise of the GNL distribution are estimated using higher-order moments. After in-depth analysis of estimated variances, a feature that is useful for discrimination between speech and noise at low SNRs is derived and compared to a threshold to detect speech activity. To consider the inter-frame correlation of speech activity, the result from the previous frame is employed in the decision rule of the proposed VAD algorithm. The performance of our proposed VAD algorithm is evaluated in terms of receiver operating characteristics (ROC) and detection accuracy. Results show that the proposed method yields better results than conventional VAD algorithms.

key words: *voice activity detection, generalized normal-Laplace distribution, higher order moments, conditional maximum a posteriori*

1. Introduction

In many speech processing procedures such as speech enhancement, speech recognition, and speech coding, a voice activity detection (VAD) algorithm has become an essential component, because the performance of these speech processing procedures relies on the accurate detection of a speech signal from within a noisy signal. For instance, a VAD module is a key component of variable-rate speech coding, because it provides an effective way of enhancing the capacity and coverage of the communication bandwidth. For this reason, a variety of VAD algorithms has been proposed. Early VAD algorithms were based on heuristic rules on several features such as spectral energy, zero crossing rate (ZCR), linear prediction coding, pitch, spectral deviation, and higher-order energy [1], [2]. More recently, VAD-algorithm-based pattern recognition such as support vector machines (SVMs) and Gaussian mixture models (GMMs), which use a mixture of conventional features, have been proposed [3]. However, these VAD algorithms have shown poor performance in adverse noise environments because conventional features cannot distinguish between speech and noise at low SNRs. Therefore, features that can accurately specify the characteristics of speech in adverse noise environments are needed.

Recently, the features of higher-order moments based on the generalized normal-Laplace (GNL) distribution (which is used to represent the probability density function (PDF) of the noisy speech signal) have been proposed for estimating the SNR; they have shown good performance in adverse noise environments [4]. These features can be used to detect the speech activity at low SNR because the instantaneous SNR in the time-domain is closely related to speech activity.

In this paper, we propose a novel VAD algorithm based on the GNL distribution to improve the performance of VAD in various noisy environments. First, we analyze features (the same as those used in [4]), namely, the estimated variance of speech and noise in the GNL distribution. On the basis of these two features, a robust feature that is accurately able to distinguish speech from a noisy signal is derived; this new metric is compared to a threshold to detect speech activity. In addition, to further enhance the performance of VAD, a conditional maximum a posteriori (CMAP) criterion, which considers the inter-frame correlation of voice activity, is adopted in the decision rule.

2. The Variance Estimation for the Speech and Noise Model Based on GNL

In this section, we briefly review the method for estimating the variance of the speech and noise in a GNL distribution, which arises as the convolution of the independent normal and generalized Laplace distribution [4], [5].

Let $s(t)$ and $n(t)$ denote a clean speech and an uncorrelated additive noise signal, respectively. The noisy speech signal is the sum of a clean speech signal and a noise signal. Assuming that the clean speech signal and the noise signal are statistically independent and the that PDF of noise and speech is characterized by a zero-mean Gaussian and Laplace distribution, the PDF of the noisy speech signal is obtained by the convolution of the Gaussian and Laplace component.

$$f_T(t) = \frac{e^{\frac{-t^2}{2\sigma_n{}^2}}}{\sqrt{2\pi}\sigma_n} * \frac{1}{\sqrt{2}\sigma_s} e^{-\sqrt{2}\sigma_s{}^{-1}|t|} \tag{1}$$

where $f_T(t)$, $\sigma_s{}^2$, and $\sigma_n{}^2$ are the PDF of the GNL and the variances of the speech and noise, respectively. This convolution can be expressed as the product of the characteristic function of the normal and Laplace distributions because the characteristic function is the inverse Fourier transform of the PDF.

Manuscript received May 14, 2013.
Manuscript revised August 9, 2013.

[†]The authors are with the Department of Electronic Engineering, Inha University, Incheon, Korea.
a) E-mail: neverjin0109@naver.com

DOI: 10.1587/transinf.E96.D.2888

$$\Phi_{GNL}(t) = \Phi_{ND}(t) \cdot \Phi_{LD}(t) = \left(\frac{e^{-\sigma_n{}^2 t^2/2}}{1+\sigma_s{}^2 t^2/2}\right)^{\gamma} \quad (2)$$

where Φ_{ND}, Φ_{LD}, and Φ_{GNL} are the characteristic functions of the normal, Laplace, and GNL distributions, respectively. γ is the shape parameter; it measures the peakedness of the distribution.

In (2), the unknown parameters ($\sigma_s{}^2$, $\sigma_n{}^2$, and γ) can be estimated on the basis of the moments of the distribution. The moment of the distribution is defined in terms of its characteristic function as follows:

$$M_p = i^{-p}\left[\frac{d^p}{dt^p}\Phi_{GNL}(t)\right]_{t=0} \quad (3)$$

where p is the order of the moment. From (2) and (3), the higher order moments of the GNL distribution can be computed as [4]

$$M_2 = \gamma(\sigma_n{}^2 + \sigma_s{}^2) \quad (4)$$

$$M_4 = 3(\gamma^2(\sigma_n{}^2 + \sigma_s{}^2)^2 + 3\sigma_s{}^4\gamma) \quad (5)$$

$$M_6 = 15(\gamma^3(\sigma_n{}^2+\sigma_s{}^2)^3+3\gamma^2\sigma_s^4(\sigma_n{}^2+\sigma_s{}^2)+2\gamma\sigma_s{}^6) \quad (6)$$

Here, the higher-order moments are approximated by the sample moments. Using the approximated higher-order moments, the variance of the speech and noise are as follows:

$$\hat{\sigma}_s^2 = \frac{-45c\hat{M}_2 + \sqrt{2025c^2\hat{M}_2{}^2 - 900c\hat{M}_2{}^3 + 60c\hat{M}_6{}^2}}{60c} \quad (7)$$

$$\hat{\sigma}_n^2 = \hat{\sigma}_s^2\left(\frac{\hat{\sigma}_s^2\hat{M}_2}{4c} - 1\right), \quad c = \frac{\hat{M}_4 - 3\hat{M}_2}{12} \quad (8)$$

where $\hat{M}_p$ is the sample moment of the pth order.

3. Proposed VAD Based on the GNL Distribution Incorporating Conditional MAP

In the previous section, it was noted that the variance of speech and noise in the PDF model could be obtained using high-order moments and the characteristic function of the GNL. Based on these feature vectors, we propose a novel VAD algorithm that has an improved probability of detecting speech activity under adverse noise conditions. To derive the robust feature in the VAD algorithm, we first analyze the speech to noise variance ratio (SNVR).

$$SNVR = \hat{\sigma}_s^2/\hat{\sigma}_n^2 \quad (9)$$

As we assume that the speech signal is characterized by the Laplace distribution, if the speech in the input signal is more salient, the variance of the Laplace component will have more low values than that of the Gaussian component. Therefore, the SNVR during speech periods tends to be lower than that during noise-only periods. Further, because the speech signal has a higher variation than the noise signal over time, the SNVR during speech periods has a larger dynamic range than that during noise-only periods.

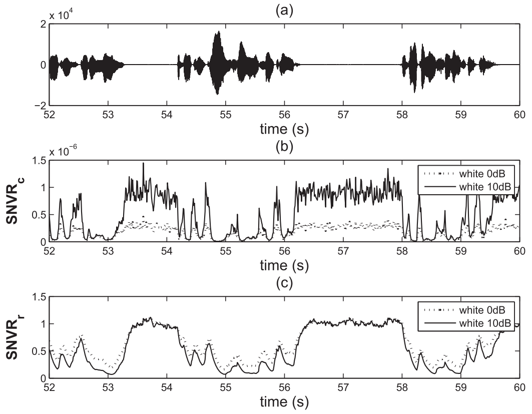

Fig. 1 Comparison of $SNVR_c$ and $SNVR_r$ under white noise conditions (SNR = 0, 10 dB).

Figure 1 (b) shows SNVR values for two different SNR values for white noise (0 dB = dotted line, 10 dB = bold line). This figure shows the aforementioned tendency. Additionally, this figure indicates that the direct application of the SNVR to detect voice activity is difficult because the ranges of the SNVR values during both speech and noise are different for different SNR values. On the basis of this observation, we define a new feature ($SNVR_r$), which is the ratio of the SNVR of the current frame ($SNVR_c$) to the smoothed SNVR of the noise-only frame ($SNVR_n$).

$$SNVR_r(t) = SNVR_c(t)/SNVR_n(t) \quad (10)$$

where the initial value of the $SNVR_n$ is calculated by averaging the initial four frames. $SNVR_n$ is updated during the speech absence as follows:

$$SNVR_n(t) = SNVR_n(t-1) \times (1-\alpha_s) + SNVR_c(t) \times \alpha_s \quad (11)$$

where α_s is a smoothing parameter. Figure 1 (c) shows the proposed feature ($SNVR_r$) for white noise (0 and 10 dB SNR). This figure also shows that $SNVR_r$ during both speech and noise periods has a similar value for different SNR values. In other words, the same threshold can be used to detect speech activity for different SNR values. To verify whether $SNVR_r$ can serve as a useful feature for VAD, we obtain histograms to determine the statistical distribution of $SNVR_r$ for speech and noise. Figure 2 shows the histogram of $SNVR_r$ and compares it to that of $SNVR_c$. Figure 2 shows that the proposed feature provides superior discrimination between speech and noise than $SNVR_c$ in the various noise environments. Furthermore, this feature is unimodal, and therefore, could be successfully represented by a Gaussian basis function. On the basis of this observation, the speech activity is detected by a decision rule in the proposed algorithm based on the likelihood ratio test (LRT) as follows:

$$\frac{P(SNVR_r(t)|H_1(t))}{P(SNVR_r(t)|H_0(t))} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha\frac{P(H(t) = H_0)}{P(H(t) = H_1)} \quad (12)$$

where H_0, H_1, $P(H(t) = H_i)$, and α indicate speech absence,

Table 1 Comparison of total error rate (P_E), false rejection rate (P_R), and false acceptance rate (P_A) of the proposed and conventional algorithms.

Environments		J.Shon			GSAP			G.729 Annex B			G.729 Appendix III			Proposed		
Noise	SNR	P_E	P_R	P_A	P_E	P_R	P_A	P_E	P_R	P_A	P_E	P_R	P_A	P_E	P_R	P_A
White	0	16.2	19.7	11.3	12.6	12.7	12.3	34.5	41.7	24.4	22.5	31.8	9.6	9.0	7.6	11.6
	5	12.8	15.7	9.1	11.4	11.7	11.0	27.8	30.0	24.7	11.8	12.2	11.2	7.7	5.1	10.6
	10	10.3	12.5	7.1	9.1	9.1	9.0	22.7	22.2	23.2	9.7	6.7	13.8	5.9	2.9	9.6
	15	9.1	10.5	6.5	7.3	6.9	7.5	18.6	13.9	25.2	7.4	3.8	12.3	5.4	1.4	8.8
Babble	0	34.0	34.8	32.9	33.2	32.7	34.1	37.2	63.5	0.5	27.4	26.4	30.3	28.2	27.1	30.4
	5	25.7	34.0	15.4	24.6	29.7	18.3	25.2	42.8	0.7	21.3	13.0	32.8	21.5	25.6	17.5
	10	22.3	21.0	23.9	18.1	18.1	18.1	17.4	29.1	0.9	15.4	5.2	28.9	16.1	15.2	16.5
	15	17.9	16.9	19.2	12.5	12.5	12.5	12.9	20.0	3.1	12.7	1.7	27.5	12.1	13.0	10.5
Office	0	23.9	26.7	20.0	20.6	21.5	19.3	28.5	38.2	15.1	18.3	17.3	19.7	18.2	18.8	18.2
	5	18.1	16.7	19.9	16.6	16.5	16.9	26.5	28.7	23.4	14.8	9.7	22.9	14.5	15.0	14.8
	10	15.6	13.4	17.8	14.4	13.9	14.9	22.7	22.3	23.4	12.8	5.0	23.7	12.2	13.0	12.6
	15	13.8	11.4	16.0	12.4	11.8	13.1	19.3	17.3	22.0	12.1	2.6	25.3	10.2	11.1	10.5

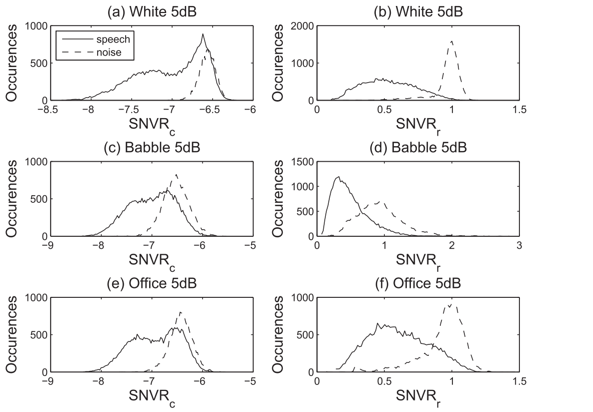

Fig. 2 Histogram of $SNVR_c$ and $SNVR_r$ for white, babble and office noise condition (SNR = 5 dB).

speech presence, the *a priori* probability of H_i, and a compensation factor, respectively. In addition, as the speech signal has a strong correlation between the consecutive frames, we consider the result of VAD in the previous frame to create the decision rule in the proposed VAD as follows [6]:

$$\frac{P(SNVR_r|H(t) = H_1, H(t-1) = H_i)}{P(SNVR_r|H(t) = H_0, H(t-1) = H_i)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha'_i \tag{13}$$

where

$$\alpha'_i = \alpha \frac{P(H(t) = H_0|H(t-1) = H_i)}{P(H(t) = H_1|H(t-1) = H_i)}, \quad i = 0, 1. \tag{14}$$

As the voice activity of the current frame is predominantly affected by the $SNVR_r$ value in the current frame, the final decision rule of the proposed VAD algorithm can be simplified as follows:

$$\frac{P(SNVR_r|H(t) = H_1)}{P(SNVR_r|H(t) = H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha'_i, \quad i = 0, 1. \tag{15}$$

From (15), we can see that the final decision rule has two separate thresholds, specified according to the speech activity in the previous frame.

4. Experimental Results

To evaluate the performance of the proposed VAD algorithm, we performed speech detection under various noise conditions for each VAD algorithm. In our experiments, speech material spoken by four male and four female speakers was sampled at 8 kHz. To evaluate the detection accuracy, we made reference decisions on the clean speech materials of 456 s long by manually labeling each frame at 10 ms intervals. The proportion of voiced, unvoiced, and silent frames were 44.8%, 13.4% and 41.8%, respectively. To verify the performance of the proposed method in terms of noise characteristics, we selected white noise, babble noise (from the NOISEX-92 database), and office noise (from the Dynastat database). Whereas the white noise was completely stationary, the babble and office noises were typical non-stationary noises. In order to simulate noisy condition, we added these noises to clean speech signal at SNRs of 0, 5, 10, and 15 dB. Table 1 summarizes the detection accuracy of the VAD algorithms based on the proposed method and statistical models [7], [8]. Here, the threshold of the aforementioned VAD algorithm was experimentally determined to minimize the total error rate under a large number of noisy speech data samples containing a variety of noises and SNR conditions. In this table, P_R (false rejection rate) is the probability that noise is (mistakenly) identified when speech is present and P_A (false acceptance rate) is the probability that speech is (mistakenly) detected when no speech is present. P_E (total error rate) is the false detection probability of all frames. To show that the performance of the proposed method is acceptable in practice, the results for the well-known standard VAD algorithms, ITU-T G.729 Annex B and ITU-T G.729 Appendix III are also included [9], [10]. Our experimental results show that not only does the proposed VAD algorithm based on GNL outperform other statistical approaches, but it also exhibits better (or comparable) performance than the standard VAD algorithms in most environmental conditions.

The receiver operating characteristics (ROCs) that illustrate a trade-off between the speech detection rate (100-P_R) and P_A for babble, white, and office noise environments, are shown in Figs. 3–5. These figures show the overall performance differences among the aforementioned algorithms. The working points of the standard VAD algorithms are also included. These figures indicate that the proposed VAD algorithm yields better performance than statistical-based

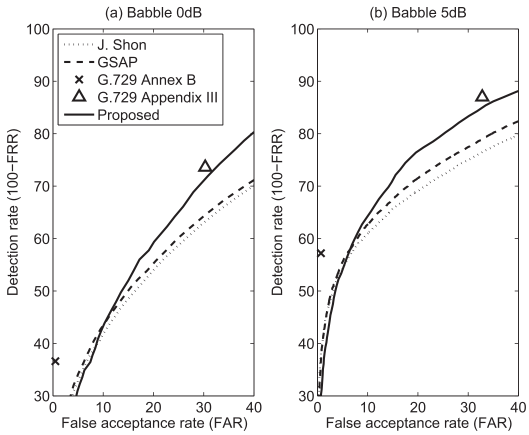

Fig. 3 ROC curves for babble noise conditions (SNR = 0, 5 dB).

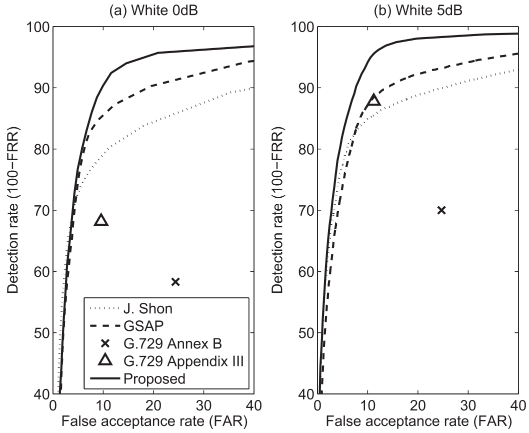

Fig. 4 ROC curves for white noise conditions (SNR = 0, 5 dB).

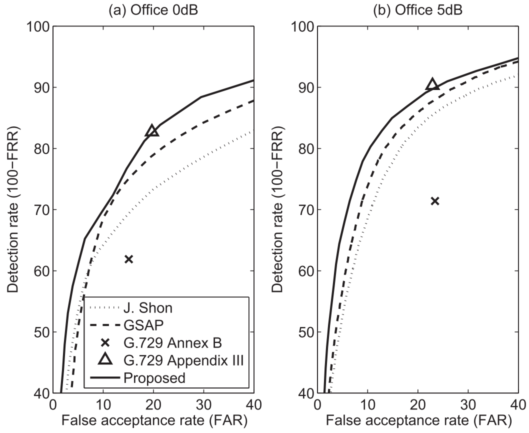

Fig. 5 ROC curves for office noise conditions (SNR = 0, 5 dB).

algorithms and G.729 Annex B under the noisiest conditions. Although the performance of the proposed method is slightly lower than that decribed in G.729 Appendix III for babble and office noises, determining the performance of the proposed method is still a useful result because it uses only one feature vector (SNVR) whereas the methods in G.729 Appendix III are obtained using a combination of multiple features such as line spectral frequency (LSF), low-band energy, and zero crossing rate (ZCR). Additionally, as the feature vector (SNVR) of the proposed method is directly calculated from the microphone input signal and has low computational complexity, our approach has the advantage of being easy to incorporate with a real-time speech signal processing applications.

5. Conclusion

In this paper, we have proposed a novel VAD algorithm based on GNL distribution for use in the time domain. The feature vector was obtained by estimating the variance of speech and noise in the GNL distribution using high-order sample moments. Our experimental results (i.e., ROCs and detection accuracy) indicate that the performance of the proposed VAD algorithm was superior to that of conventional statistical-based VAD algorithms (e.g., the LR-based method and GSAP). Additionally, a comparison between the proposed method and the standard VAD algorithm (G.729 Annex B, G.729 Appendix III) indicated that the performance of the proposed method is acceptable in practical applications. Further improvement is expected if we can incorporate frequency-related information such as spectral deviation and channel SNR.

Acknowledgement

This work was supported by grant No.SS100022 by Seoul R&BD Program and was supported by Inha University Research Grant.

References

[1] R. Tucker, "Voice activity detection using a periodicity measure," Proc. Inst. Electr. Eng., vol.139, pp.377–380, Aug. 1992.

[2] Y.S. Park and S. Lee, "Voice activity detection using global speech absence probability based on teager energy for speech ehancement," IEICE Trans. Inf. & Syst., vol.E95-D, no.10, pp.2568–2571, Oct. 2012.

[3] J. Wu and X.L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," IEEE Signal Process. Lett., vol.18, no.8, pp.466–499, June 2011.

[4] T. Moazzeni, A. Amei, J. Ma, and Y. Jiang, "Statistic model based SNR estimation method for speech signals," Electron. Lett., vol.48, no.12, pp.727–728, June 2012.

[5] J. Reed, "The Normal-Laplace distribution and its relatives," Advances in Distribution Theory, Order Statistics and Inference, N. Balakrishna and Birkhauser, pp.61–74, 2006.

[6] J.W. Shin, H.J. Kwon, S.H. Jin, and N.S. Kim, "Voice activity detection based on conditional MAP criterion," IEEE Signal Process. Lett., vol.15, pp.257–260, Feb. 2008.

[7] N.S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," IEEE Signal Process. Lett., vol.7, no.5, pp.108–110, May 2000.

[8] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Process. Lett., vol.6, no.1, pp.1–3, Jan. 1999.

[9] ITU-T, A silence compression scheme for G.729 optimised for terminals conforming to recommendation V.70, ITU-T Rec. G.729, Annex B, 1996.

[10] ITU-T, Appendix III: G.729 Annex B enhancement in voice-over-IP applications-Option 2, 2005.