VOICE ACTIVITY DETECTION BASED ON GENERALIZED GAMMA DISTRIBUTION

Jong Won Shin[1], *Joon-Hyuk Chang*[2], *Hwan Sik Yun*[1] *and Nam Soo Kim*[1]

School of Electrical Engineering and INMC[1]
Seoul National University, Seoul 151-742, Korea
Department of Electrical and Computer Science[2]
University of California, Santa Barbara, CA 93106-9560, USA
E-mail: jwshin@hi.snu.ac.kr, jhchang@ece.ucsb.edu, hsyun@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

We propose a voice activity detection (VAD) algorithm based on the generalized gamma distribution (GΓD). The distributions of noise spectra and noisy speech spectra including speech-inactive intervals are modeled by a set of GΓD's and applied to the likelihood ratio test (LRT) for VAD. The parameters of GΓD are estimated through an on-line maximum likelihood (ML) estimation procedure where the global speech absence probability (GSAP) is incorporated under a forgetting scheme. Experimental results show that the proposed VAD algorithm based on GΓD outperformed the algorithms based on other statistical models.

1. INTRODUCTION

As the need of bandwidth efficiency in speech communication system increases, voice activity detection (VAD) has become an indispensable part of the variable rate speech coders. Recently, VAD algorithms based on likelihood ratio test (LRT) employing statistical models have been proposed and shown good performances [1], [2]. In most of the conventional VAD algorithms adopting statistical models which operate in the discrete Fourier transform (DFT) domain, the distributions of noisy speech spectra and noise spectra are assumed to be complex Gaussians [1]. Chang et. al. [2] utilized the Laplacian probability density function (pdf) to model the distributions of noisy speech spectra and noise spectra, which was shown to be a better model for the distribution of clean speech [3], [4], and showed that VAD based on this complex Laplacian model was better than that based on the complex Gaussian model. Recently, it was also reported that the generalized gamma distribution (GΓD) provides a better model of the distribution of clean speech spectra than the Gaussian, Laplacian or Gamma pdf [5].

In this paper, we propose a novel VAD algorithm in which the generalized gamma distribution (GΓD) is employed for the LRT. The on-line maximum likelihood (ML) parameter estimation algorithm proposed in [5] is modified such that it can be applied to VAD by incorporating the global speech absence probability (GSAP). Experimental results show that VAD based on GΓD outperforms those which employ other pdf's and its performance is even better than that of a number of standardized VAD algorithms including ETSI AMR VAD option 2 and ITU-T G.729 annex B VAD.

2. ON-LINE ML ESTIMATION OF THE PARAMETERS OF GΓD

In this section, we briefly review the on-line ML procedure for the estimation of the GΓD parameters [5]. GΓD is defined by

$$f_{\mathbf{x}}(x) = \frac{\gamma\beta^{\eta}}{2\Gamma(\eta)}|x|^{\eta\gamma-1}\exp(-\beta|x|^{\gamma}) \qquad (1)$$

where $\Gamma(z)$ denotes the gamma function, and η, β and γ are positive real valued parameters. This covers a fairly flexible family of distributions which includes most of the commonly used speech distributions. It is observed that if $\gamma = 2$ and $\eta = 0.5$, it becomes the Gaussian pdf, and if $\gamma = 1$ and $\eta = 1$, it represents the Laplacian pdf. The pdf commonly referred to as just the 'Gamma pdf' is a special case of the gamma pdf with $\gamma = 1$ and $\eta = 0.5$.

The parameters η, β, and γ should be estimated to take advantage of the assumed pdf for various applications. Here, we apply the ML criterion to estimate the parameters of GΓD. Given N data $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$, with the assumption that the data are mutually independent, the log-likelihood function is given as follows:

$$\begin{aligned}\log f_{\mathbf{x}}(\mathbf{x};\eta,\beta,\gamma) = \; & N\log\tfrac{\gamma\beta^{\eta}}{2\Gamma(\eta)} \\ & +(\eta\gamma-1)\textstyle\sum_{i=1}^{N}\log|x_i| \; -\beta\sum_{i=1}^{N}|x_i|^{\gamma}. \qquad (2)\end{aligned}$$

By differentiating the log-likelihood function with respect

to η, β and γ and setting them to zero, we obtain the following three equations:

$$\psi_0(\eta) = \log\beta + \frac{1}{N}\sum_{i=1}^{N}\log|x_i|^\gamma \quad (3)$$

$$\beta = \eta\frac{1}{\frac{1}{N}\sum_{i=1}^{N}|x_i|^\gamma} \quad (4)$$

$$\frac{1}{\eta} + \psi_0(\eta) - \log\beta - \frac{\beta}{\eta}\frac{1}{N}\sum_{i=1}^{N}|x_i|^\gamma\log|x_i|^\gamma = 0 \quad (5)$$

where $\psi_0(z)$ is the digamma function, which denotes the first-order derivative of $\log\Gamma(z)$. After some mathematical manipulation, the ML estimate of γ is obtained by the root of the single nonlinear equation,

$$\psi_0\Big(\frac{\frac{1}{N}\sum_{i=1}^{N}|x_i|^\gamma}{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i|^\gamma\log\frac{|x_i|^\gamma}{|x_j|^\gamma}}\Big) - \frac{1}{N}\sum_{i=1}^{N}\log|x_i|^\gamma + \log\Big(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i|^\gamma\log\frac{|x_i|^\gamma}{|x_j|^\gamma}\Big) = 0. \quad (6)$$

Given an estimate of γ, it is straightforward to derive the estimates for η and β.

Since, however, it is difficult to solve (6) analytically, Shin et. al. [5] employ a gradient ascent algorithm to obtain the estimate of γ, and determine the estimates of η and β based on the obtained value of γ. From now on, let us denote the estimates of γ, η and β by $\hat{\gamma}$, $\hat{\eta}$ and $\hat{\beta}$, respectively. The large sample size and the reasonable initial estimates yield a satisfactory estimation of the parameters via the gradient ascent algorithm despite the persistent divergence of iterative numerical methods and the possibility of multiple solutions. Our previous work suggested an on-line algorithm with a forgetting scheme which emphasizes the data incoming most recently. To estimate the relevant parameters, only three statistics should be computed over the given data, $\frac{1}{N}\sum_{i=1}^{N}|x_i|^{\hat{\gamma}}$, $\frac{1}{N}\sum_{i=1}^{N}\log|x_i|^{\hat{\gamma}}$, and $\frac{1}{N}\sum_{i=1}^{N}|x_i|^{\hat{\gamma}}\log|x_i|^{\hat{\gamma}}$. For the implementation of an on-line algorithm, these statistics are modified to incorporate a forgetting factor λ, i.e.,

$$\begin{aligned} S_1(n) &= (1-\lambda)S_1(n-1) + \lambda|x_n|^{\hat{\gamma}(n)} \\ S_2(n) &= (1-\lambda)S_2(n-1) + \lambda\log|x_n|^{\hat{\gamma}(n)} \\ S_3(n) &= (1-\lambda)S_3(n-1) + \lambda|x_n|^{\hat{\gamma}(n)}\log|x_n|^{\hat{\gamma}(n)} \end{aligned} \quad (7)$$

In our experiments, the initial value for $\hat{\gamma}$ is set to 1, which specifies the Laplacian or Gamma pdf, for both the noisy speech and the noise. Once $\hat{\gamma}$ is given, we can obtain $\hat{\eta}$ and $\hat{\beta}$ from (3) and (4) such that

$$\psi_0(\hat{\eta}(n)) - \log\hat{\eta}(n) = S_2(n) - \log S_1(n) \quad (8)$$

$$\hat{\beta}(n) = \frac{\hat{\eta}(n)}{S_1(n)} \quad (9)$$

by taking the forgetting scheme into consideration. Since $\psi_0(z) - \log z$ is a monotonically increasing function of z, the value of $\hat{\eta}$ can be uniquely determined if the solution exists. The value of $\hat{\gamma}$ is updated at each time based on the gradient ascent approach given as follows:

$$\hat{\gamma}(n+1) = \hat{\gamma}(n) + \mu\phi(\hat{\gamma}(n), \hat{\eta}(n), \mathbf{x}) \quad (10)$$

where μ is a learning rate and $\phi(\hat{\gamma}(n), \hat{\eta}(n), \mathbf{x})$ is an on-line version of the gradient of the 'average' log-likelihood function with respect to γ. The 'average' log-likelihood function is given as (2) divided by N, and its gradient with respect to γ equals to the left-hand side of (5). Using (3), (4), (5) and (7), the on-line version of the gradient is given by

$$\phi(\hat{\gamma}(n), \hat{\eta}(n), \mathbf{x}) = \frac{1}{\hat{\eta}(n)} + S_2(n) - \frac{S_3(n)}{S_1(n)}. \quad (11)$$

As we can see, the estimation procedure is not computationally expensive if we store the values of the function $\psi_0(z) - \log z$ or the inverse of it on a table.

3. LIKELIHOOD RATIO TEST BASED ON GΓD

VAD can be considered as a hypothesis test where one hypothesis (H_0) states that the input signal consists of a pure noise and the other (H_1) indicates that the input is a mixture of both the active speech and noise. The distributions for the noise and noisy speech spectra are modeled by separate GΓD's, and LRT is performed for each frame of the input signal.

In our approach, what is distinguished from the other conventional VAD algorithms is that the noisy speech spectra distribution represents not only the active speech regions but also the inactive speech regions. Even though this approach may cause a biased estimate of the likelihood ratio value, we have found that it enables a more robust parameter estimation in noisy environment. We assume that the real and imaginary parts of the DFT coefficient are statistically independent [2] and distributed according to the same GΓD, i.e.,

$$p(X_k) = \frac{\gamma^2\beta^{2\eta}}{4\Gamma(\eta)^2}|X_{k,R}X_{k,I}|^{\eta\gamma-1} \cdot \exp(-\beta|X_{k,R}|^\gamma - \beta|X_{k,I}|^\gamma) \quad (12)$$

for both the noise and noisy speech. This is equivalent to the assumption that both the real part and the imaginary part are the realization of the same random variable distributed according to GΓD, i.e., the data set $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$ can be substituted with $\{X_{k,R}(1), X_{k,I}(1), X_{k,R}(2), X_{k,I}(2), \cdots, X_{k,R}(\frac{N}{2}), X_{k,I}(\frac{N}{2})\}$.

First, the parameters of the specified GΓD's should be estimated. For the distribution of noisy speech spectra, the

parameter estimation procedure is the same to the one described in the previous section. On the other hand, for the distribution of noise, we do not have a knowledge as to which frame contains active speech and which does not, or how large portion of the given input signal contributes to noise estimation. Previous studies [1], [2] compute variously defined signal-to-noise ratio (SNR) and use it to estimate the noise power from the noisy speech spectra directly. The procedure is considered rather simple since only the variances are required to be estimated. In contrast, we need to estimate all three statistics, S_1, S_2, S_3 in (7), and we can not solely rely on SNR. Here, we take the global speech absence probability (GSAP) as a measure of speech inactivity, and incorporate it into a forgetting scheme. The GSAP is given by [6]

$$\begin{aligned} P(H_0|\mathbf{X}) &= \frac{p(\mathbf{X}|H_0)P(H_0)}{p(\mathbf{X})} \\ &= \frac{p(\mathbf{X}|H_0)P(H_0)}{p(\mathbf{X}|H_0)P(H_0)+p(\mathbf{X}|H_1)P(H_1)} \\ &= \frac{1}{1+\frac{P(H_1)}{P(H_0)}\prod_{k=1}^{M}\Lambda_k}. \end{aligned} \tag{13}$$

where $\mathbf{X} = [X_1, X_2, \cdots X_M]$ when M indicates the total number of spectral bins and $P(H_0)(= P(H_1))$ represents the *a priori* probability of speech absence [6].

Given the GSAP, the statistics in (7) are modified to incorporate a measure of speech activity under the forgetting scheme such that

$$\begin{aligned} S_1(n) &= (1-\lambda P)S_1(n-1) + \lambda P|x_n|^{\hat{\gamma}(n)} \\ S_2(n) &= (1-\lambda P)S_2(n-1) + \lambda P \log|x_n|^{\hat{\gamma}(n)} \\ S_3(n) &= (1-\lambda P)S_3(n-1) + \lambda P|x_n|^{\hat{\gamma}(n)}\log|x_n|^{\hat{\gamma}(n)} \end{aligned} \tag{14}$$

where P represents the computed GSAP when estimating the noise spectra distribution and it is set to 1.0 to update the estimates for the noisy speech spectra distribution and λ is a forgetting factor. Once we obtain the estimated statistics S_1, S_2, S_3 through (14), we can estimate η, β, γ by means of (8), (9), (10) and (11) for both the noisy speech and noise. For active speech period where GSAP has a small value near to zero, the statistics, S_1, S_2, S_3 are updated very slowly for the noise spectra distribution while the estimate for the parameters of the noisy speech distribution evolves rather fast.

Given the parameters of GΓD, the likelihood ratio for the *k*-th DFT coefficient is given by

$$\begin{aligned} \Lambda_k &= \frac{p(X_k|H_1)}{p(X_k|H_0)} \\ &= \frac{\hat{\gamma}_S^2\hat{\beta}_S^{2\hat{\eta}_S}\Gamma(\hat{\eta}_N)^2}{\hat{\gamma}_N^2\hat{\beta}_N^{2\hat{\eta}_N}\Gamma(\hat{\eta}_S)^2}|X_R X_I|^{\hat{\eta}_S\hat{\gamma}_S-\hat{\eta}_N\hat{\gamma}_N} \\ &\quad \cdot \mathrm{e}^{(-\hat{\beta}_S(|X_R|^{\hat{\gamma}_S}+|X_I|^{\hat{\gamma}_S})+\hat{\beta}_N(|X_R|^{\hat{\gamma}_N}+|X_I|^{\hat{\gamma}_N}))} \end{aligned} \tag{15}$$

where the subscript N indicates parameters related to the pdf of the noise spectra while the subscript S indicates those corresponding to the pdf of the noisy speech spectra. The final decision rule for VAD is given as follows:

$$\log\Lambda = \sum_{k=0}^{M-1}\log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \tag{16}$$

The decision threshold ξ as well as μ and λ which control the rate of parameter update are determined according to a SNR-based rule which will be described in the next section. To further enhance the performance of VAD, $\log\Lambda$ is modified using hangover scheme proposed in [1], and then smoothed using a forgetting scheme similar to that in [7]:

$$\Psi(n) = (1-\lambda_\Lambda)\Psi(n-1) + \lambda_\Lambda \log\Lambda \tag{17}$$

where λ_Λ is a smoothing factor.

4. EXPERIMENTAL RESULTS

To compare the performance of the proposed algorithm with that of the conventional algorithms, we evaluated speech detection error probability (P_e), where both false alarms and missing errors are considered. In our experiments, speech data spoken by 4 male and 4 female speakers were sampled at 8000 Hz. The total length of the speech material was 456 s. To obtain P_e, we made reference decisions on a clean speech material by labeling manually at every 10 ms frame. The percentage of the hand-marked speech frames was 58.2% which consisted of 44.8% voiced sounds and 13.4% unvoiced sounds frames. In order to make noisy environments, we added the vehicular and office noises to the clean speech data by varying SNR.

The threshold, ξ as well as the smoothing parameter of test statistic, λ_Λ, the forgetting factor of statistics in (14) for noisy speech, λ, the learning rate of γ for noisy speech, μ, the ratio of λ for noisy speech to that for noise, R_λ, and the ratio of μ for noisy speech to that for noise, R_μ were determined to minimize P_e. The forgetting factor λ used to update the noise spectra distribution was set to be higher than that for the noisy speech distribution to make the effective averaging interval lengths equal since for noise, $\lambda P(H_0|\mathbf{X})$ played a role of a forgetting factor. The learning rate μ for the noise spectra was chosen smaller than that of the noisy speech spectra based on the assumption that the background noise characteristic evolves more slowly. These factors are adaptively determined based on SNR. λ_Λ should be increased as SNR increases, since for low SNR, more smoothing is needed. On the other hand, λ and μ should be set higher when the SNR is low to enable a fast update of

noise	vehicle			office		
SNR(dB)	5	10	15	5	10	15
G.729B	27.49%	23.45%	19.76%	26.43%	22.72%	19.26%
AMR 2	8.09%	6.91%	6.29%	16.24%	14.77%	15.43%
Laplacian	11.48%	8.60%	6.91%	18.43%	16.45%	17.25%
Gamma	11.84%	9.24%	7.49%	23.54%	21.01%	18.96%
GΓD	6.41%	5.85%	5.38%	13.47%	14.60%	18.34%

Table 1. P_e of the proposed GΓD, Laplacian, Gamma-based, AMR VAD option 2 and G.729 Annex B VAD's for the various environmental conditions

the statistics. R_λ should be decreased as SNR goes lower because adaptability becomes more important not only for noise but also for noisy speech in a low SNR environment. In contrast, ξ should be larger in higher SNR conditions since the estimates for the noise spectra distribution are unreliable. Factor values used in the experiment were $\lambda_\Lambda \in [0.04, 0.2]$, $\lambda \in [0.022, 0.028]$, $\mu \in [0.006, 0.0085]$, $R_\lambda \in [1.05, 1.45]$ and $R_\mu = 0.7$.

The detection results are summarized in Table 1. From the experimental results, it is evident that not only the proposed VAD algorithm based on GΓD outperforms algorithms utilizing other commonly used pdf's, but also it shows better performance than the standard VAD algorithms such as ITU-T G.729 Annex B VAD [8] and ETSI AMR VAD option 2 [9] in most of the environmental conditions.

5. CONCLUSION

We have proposed an approach to apply the complex GΓD to VAD based on LRT. The distribution of noisy speech including inactive speech periods and that of noise spectra are modeled by GΓD's where the parameters are estimated through the on-line parameter estimation algorithm incorporating GSAP as a measure of speech inactivity. It has been found that the VAD algorithm based on GΓD outperformed those based on other widely used pdf's and the standard VAD algorithms including G.729 Annex B VAD and AMR VAD option 2 in a number of experiments. Further improvement is expected if we incorporate some feature used in standard VAD's, such as channel energy, channel SNR, and pitch lag.

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] J. -H. Chang, J. W. Shin and N. S. Kim, "Likelihood ratio test with complex Laplacian model for voice activity detection," *Proc. Eurospeech*, pp. 1065-1068, Geneva, Switzerland, Aug. 2003.

[3] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204-207, Jul. 2003.

[4] R. Martin, "Speech enhancement using short time spectral estimation with Gamma distributed priors," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. I-253 - I-256, Orlando, FL, USA, May 2002.

[5] J. W. Shin, J. -H. Chang and N. S. Kim, "Statistical modeling of speech signals based on generalized gamma distribution," to appear in *IEEE Signal Processing Letters*.

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

[7] Y. D. Cho, K. Al-Naimi and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 7-11, Salt Lake City, Utah, USA, May 2001.

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

[9] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech teaffic channels," *ETSI EN 301 708 v7.1.1*, Dec. 1999.