Voice activity detector employing generalised Gaussian distribution


A novel approach to a voice activity detector (VAD) in noisy environments is presented. The generalised Gaussian distribution (GGD) is employed as a parametric model for noisy speech, which enables tuning to the actual data. According to the experimental results, it was discovered that the proposed GGD model is more effective for the VAD algorithm compared to the conventional Laplacian model.

Introduction: In recent years, there has been a great deal of interest in low bit-rate speech coding techniques. For an efficient use of the limited bandwidth resources, it is indispensable to describe the speech signal with minimal bits. For this, a voice activity detector (VAD) has become an indispensable part of the variable-rate speech coding. One of the successful applications used to improve the performance of the VAD is the statistical model-based technique in which the likelihood ratio test (LRT) is applied to a set of hypotheses [1]. A prevailing way to characterise the statistical modelling of the speech signal is to apply the complex Gaussian model in the discrete Fourier transform (DFT) domain [1, 2]. Recently, Chang and Kim further improved this statistical model-based VAD technique by incorporating a complex Laplacian model for which the Laplacian provides a better model of the distribution of noisy speech spectra than the conventional Gaussian [3]. Alternatively, it is shown that the generalised Gaussian distribution (GGD), which includes the Gaussian and the Laplacian distribution as special cases, is a better model compared to the Laplacian for the discrete cosine transform (DCT) coefficients of image [4]. Also, the distribution of DCT coefficients of speech signal have been studied, all of which prefer GGD than the Gaussian model [5].

In this Letter, we propose a novel VAD technique in which GGD is employed for the noisy speech distribution. Since the distribution shape of noisy speech spectra than the conventional Gaussian model [3].

Generalised Gaussian distribution: We first briefly review GGD which is given by [4, 5]

\[
f(x) = \frac{\nu(a)}{2\Gamma(1/\nu)} \exp\left\{ -\frac{\left| x^\nu \right|^\nu}{\sigma^\nu} \right\}
\]

with \( a = \frac{\Gamma(3/\nu)}{\Gamma(1/\nu)} \) \( \) (2)

where \( \Gamma() \) denotes the gamma function, \( \nu \) a shape parameter controlling the distribution shape and \( \sigma \) is the standard deviation. If a random variable follows GGD, then \( \sigma \) equals the square root of the variance of the random variables. Also, it is observed that for \( \nu = 1 \) or 2 GGD becomes the Laplacian and Gaussian densities, respectively.

The shape parameter \( \nu \) is estimated successfully by matching the mean square value and variance of the data set to a GGD using Mallat’s method [6]. Letting \( m_1 = 1/\sigma \sum_{i=1}^N |x_i| \) and \( m_2 = 1/\sigma^2 \sum_{i=1}^N |x_i|^2 \) which denote moment parameters, the shape parameter \( \nu \) is estimated by solving the following equation:

\[
\hat{\nu} = \nu = F^{-1}\left( m_1/m_2 \right)
\]

where

\[
F(\nu) = \frac{\Gamma(2/\nu)}{\Gamma(1/\nu)\Gamma(3/\nu)}
\] (4)

To find a model among the distributions that fits better a given noisy speech spectra, various goodness-of-fit (GOF) experiments such as the Kolmogorov-Smirnov (KS) test have been carried out [7]. From the KS test results, it is found that the GGD model is more suitable for the DFT coefficients of noisy speech than the Laplacian and Gaussian.

Voice activity based on GGD: For VAD, 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 indicate respectively voice activity and inactivity, it is assumed that

\[
H_0; \text{speech absent: } X = N
\] (5)

\[
H_1; \text{speech present: } X = N + S
\] (6)

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

The above statistical model is completed with an appropriate specification of the DFT coefficients’ distribution. In this Letter, we consider three different probability density functions (pdfs) for the candidate distribution. The first one is the complex Gaussian pdf which is applied most widely to characterise the DFT coefficients’ distribution in a number of speech analyses [1, 2]. The second one is the complex Laplacian pdf which was proposed recently by our previous work [3] and resulted in better performance than the Gaussian case.

If \( X_{l,k} \) and \( X_{l,i} \) denote the real and imaginary parts of the DFT coefficients of \( X_l \), we derive the complex GGD on the DFT domain using the independent assumption of each part [3] such that

\[
p(X_l) = p(X_{l,k}) \cdot p(X_{l,i})
\]

\[
= \frac{\nu_{l,k}}{4\sigma_{l,k}^2} e^{\frac{-x_{l,k}^2}{2\sigma_{l,k}^2}} \times \exp\left\{ -\frac{\left| x_{l,i}^\nu_{l,i} \right|^\nu_{l,i}}{\sigma_{l,i}^\nu_{l,i}} \right\}
\]

\[
= \frac{\nu_{l,i}}{4\sigma_{l,i}^2} e^{\frac{-x_{l,i}^2}{2\sigma_{l,i}^2}} \times \exp\left\{ -\frac{\left| x_{l,k}^\nu_{l,k} \right|^\nu_{l,k}}{\sigma_{l,k}^\nu_{l,k}} \right\}
\]

where \( \sigma_{l,k} \) denotes the variance of \( X_{l,k} \).

From (7), the distributions of the DFT coefficients under the respective hypothesis (\( H_0, H_1 \)) are given by

\[
p(X_l|H_0) = \frac{\nu_{l,k}}{4\sigma_{l,k}^2} e^{\frac{-x_{l,k}^2}{2\sigma_{l,k}^2}} \times \exp\left\{ -\frac{\left| x_{l,i}^\nu_{l,i} \right|^\nu_{l,i}}{\sigma_{l,i}^\nu_{l,i}} \right\}
\]

\[
p(X_l|H_1) = \frac{\nu_{l,i}}{4\sigma_{l,i}^2} e^{\frac{-x_{l,i}^2}{2\sigma_{l,i}^2}} \times \exp\left\{ -\frac{\left| x_{l,k}^\nu_{l,k} \right|^\nu_{l,k}}{\sigma_{l,k}^\nu_{l,k}} \right\}
\]

where \( \nu_{l,k} \) and \( \nu_{l,i} \) are shape parameters related to \( H_0 \) and \( H_1 \) of noisy speech on 8th frequency bin, respectively. Also, \( \lambda_{l,k} \) and \( \lambda_{l,i} \) denote the variances of \( N_{l,k} \) and \( N_{l,i} \), respectively.

For the tracking the time-varying shape parameters \( \nu_{l,k} \) and \( \nu_{l,i} \), the moment parameters are modified to incorporate a forgetting factor which emphasizes the data incoming most recently. Actually, we employ the different online update algorithm for robust determination of shape parameters depending on the respective hypotheses. For computing \( \nu_{l,k} \) using (3), we employ the long-term smoothed moment parameters such that

\[
m_{l,k,t} = (1 - \lambda) m_{l,k,t-1} + \lambda |X_l(t)|
\]

\[
m_{l,i,t} = (1 - \lambda) m_{l,i,t-1} + \lambda \frac{|X_l(t)|}{|X_l(t)|}
\]

where \( t \) denotes the frame index and \( \lambda \) is the forgetting factor which was set to 0.004 in the experiment. In the case of \( \nu_{l,i} \), we update the moment parameters during the speech absence as the following:

\[
m_{l,k,t} = (1 - \lambda) m_{l,k,t-1} + \lambda \rho(H_0|X_l(t)) \cdot |X_l(t)|
\]

\[
m_{l,i,t} = (1 - \lambda) m_{l,i,t-1} + \lambda \rho(H_0|X_l(t)) \cdot |X_l(t)|
\]

where \( \rho(H_0|X_l(t)) = 1 - \rho(H_1|X_l(t)) \) is the global speech absence probability (GASP) proposed in [2] (see Appendix for details of \( \rho(H_0|X_l(t)) \)).

Based on the assumed statistical models, we compute the LR for the 8th frequency bin as follows
Experimental results: In order to evaluate the performance of each statistical model, we evaluated speech detection error probability \( P_e \), which is the sum of false alarm and missing probabilities. In our experiments, speech data spoken by four male and four female speakers was sampled at 8000 Hz. To evaluate \( P_e \), we made reference decisions on a clean speech material of 456 s long by labelling manually at every 10 ms frame. The percentage of the hand-marked active speech frames was 58.2% which consisted of 44.8% voiced sounds and 13.4% unvoiced sounds. In order to make noisy environments, we added the vehicular and office noises to the clean speech data by varying SNR.

The final decision for VAD is made from the geometric mean of the LRs computed for the individual frequency bins, which is obtained by

\[
\log \Lambda = \sum_{k \in \mathbb{S}} \log \Lambda_k \geq \eta \quad H_1
\]

where \( \eta \) is a threshold. The robust estimation of \( \hat{\lambda}_{a,k}, \hat{\lambda}_{s,k} \) and \( \hat{\epsilon}_k \) in (16) also plays an essential role in the performance of VAD. In this Letter, we follow the parameter estimation procedure proposed in [2, 3].

Table 1: \( P_e \) of proposed GGD, Laplacian, Gaussian-based, and G.729 Annex B VADs for various environmental conditions

<table>
<thead>
<tr>
<th>Noise</th>
<th>Vehicle</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR, dB</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>G.729B</td>
<td>27.49</td>
<td>23.45</td>
</tr>
<tr>
<td>Gaussian</td>
<td>12.58</td>
<td>9.70</td>
</tr>
<tr>
<td>Laplacian</td>
<td>11.48</td>
<td>8.60</td>
</tr>
<tr>
<td>GGD</td>
<td>8.54</td>
<td>7.99</td>
</tr>
</tbody>
</table>

Conclusions: We have presented an approach to incorporate the complex GGD to VAD. A major advantage of GGD is its capability to cover the Laplacian and Gaussian as special cases in a parametric way. The proposed approach has been found to improve the detection performance with an online estimation for the shape parameters of GGD. The proposed complex GGD model-based VAD results in a better performance than the conventional Laplacian and Gaussian model-based one in various environments.

Appendix: GSAP is defined in a frame globally as follows [2]:

\[
P(H_0|X(t)) = \frac{p(X(t)|H_0)p(H_0)}{p(X(t))} = \frac{p(X(t)|H_0)p(H_0) + p(X(t)|H_1)p(H_1)}{1 + [P(H_1)/P(H_0)][\Pi_{k=1}^{M} \Lambda_k]}
\]

where \( P(H_0|X(t)) \) approaches to one in the case of speech absence and zero during the speech presence on a current \( n \)th frame.

References