Modifying Voice Activity Detection in Low SNR by correction factors

H. Farsi, M. A. Mozaffarian, H. Rahmani
Department of Electrical Engineering
University of Birjand
P.O. Box: +98-97175-376
IRAN
hfarsi@birjand.ac.ir  amoazzafarian@birjand.ac.ir

Abstract: In this paper, by using a new estimation algorithm, we present a novel method to improve the performance of VAD systems in various noisy environments. We study the performance of Gamma, Laplacian and Gaussian distributions in low SNR and then modify the best one with correction factors.

Key-Words: Voice Activity Detection, Low SNR, statistical model, correction factors,

1 Introduction
The process of separating conversational speech and silence is called the voice activity detection (VAD) [1]. In communications systems based on variable bit rate speech coders, it represents the most important block, reducing the average bit rate; in a cellular radio system using the discontinuous transmission (DTX) mode, a VAD is able to increase the number of users and power consumption in portable equipment. Unfortunately, a VAD is far from efficient, especially when it is operating in adverse acoustic conditions. In this paper, the candidate models are the distribution of the spectral components under various noisy conditions. Not only the traditional Gaussian PDF but also the complex Laplacian and Gamma PDFs are applied to represent the distribution of each Discrete Fourier Transform (DFT) coefficients. We extend the PDFs described in [2] in low SNRs. At the end, we consider a set of correction coefficients to improve the performance of the estimation.

2 Statistical models for noisy speech
In this section we extend the hypothesis of Gaussian PDF and study the best performance of Gamma, Laplacian and Gaussian distributions especially in low SNR. Then we evaluate the error function of the best distribution to modify the false alarm. The first distribution is Gaussian PDF. We assume that the noise signal \( n(t) \) is added to the speech signal \( x(t) \), with their sum being denoted by \( y(t) \) in time domain. \( y(t) \) is transformed by the Discrete Fourier Transform (DFT) as follows:

\[
Y(t) = X(t) + N(t)
\]

Where:

\[
Y(t) = [Y_1(t), Y_2(t), \ldots, Y_m(t)],
\]

\[
X(t) = [X_1(t), X_2(t), \ldots, X_m(t)],
\]

\[
N(t) = [N_1(t), N_2(t), \ldots, N_m(t)].
\]

denote the DFT factors of the noisy speech signal, clean speech, and the added noise. Given two classes, \( H_0 \) and \( H_1 \) which, respectively, indicate speech presence and absence, it is assumed that:

\[
H_0: \text{speech absent: } Y_k(t) = N_k(t)
\]

\[
H_1: \text{speech present: } Y_k(t) = X_k(t) + N_k(t)
\]

With the Gaussian PDF assumption, the distributions of the noisy spectral components conditioned on both hypotheses are given by:

\[
p_G(Y_k|H_0) = \frac{1}{\pi \lambda_{n,k}} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{n,k}} \right\}
\]

\[
p_G(Y_k|H_1) = \frac{1}{\pi \lambda_{x,k} \lambda_{n,k}} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{x,k} + \lambda_{n,k}} \right\}
\]

where \( \lambda_{x,k} \) and \( \lambda_{n,k} \) indicate the variances of noise and speech for the individual frequency band, respectively.

The second distribution is the complex Laplacian PDF. The real and imaginary parts of each DFT coefficients are assumed to be distributed according to a real Laplacian PDF. Let \( X_k(R) \) and \( X_k(I) \) denote the real and imaginary parts, respectively, of the DFT coefficients \( X_k \). If both the real and imaginary parts have the same variances and assume to be independent [3], the distribution \( p(X_k) \) of \( X_k \) turns out to be:

\[
p_L(X_k) = p_L(X_k(R)) \cdot p_L(X_k(I)) = \frac{1}{\sigma_x} \exp \left\{ -\frac{2(|X_k(R)|^2 + |X_k(I)|^2)}{\sigma_x^2} \right\}
\]

From this equation, the distributions of the DFT coefficients under the respective hypotheses are given by [4]:

\[
p_L(X_k|H_0) = \frac{1}{\lambda_{n,k}} \times \exp \left\{ -\frac{2(|X_k(R)|^2 + |X_k(I)|^2)}{\sqrt{\lambda_{n,k}}} \right\}
\]

\[
p_L(X_k|H_1) = \frac{1}{\lambda_{n,k} + \lambda_{x,k}} \times \exp \left\{ -\frac{2(|X_k(R)|^2 + |X_k(I)|^2)}{\sqrt{\lambda_{n,k} + \lambda_{x,k}}} \right\}
\]
The last statistical model is described in terms of the complex Gamma PDF. If the real and imaginary parts assumed to be independent of each other as in the Laplacian case, the distribution of a DFT coefficient $X_k$ is then given by:

$$p_M(X_k) = \frac{\sqrt{6}}{8\pi \sigma_X} |X_k(\Omega)|^{0.5} |X_k(i)|^{0.5} \times \exp\left\{ -\frac{\sqrt{3}(|X_k(\Omega)|+|X_k(i)|)}{\sqrt{2}\sigma_X} \right\}$$

(7)

Applying this equation in two hypotheses $H_0$ and $H_1$, we have the distributions of the DFT coefficients as follows:

$$p_M(X_k|H_0) = \frac{\sqrt{6}}{8\pi \sqrt{\lambda_{n,k}}} |X_k(\Omega)|^{0.5} |X_k(i)|^{0.5} \times \exp\left\{ -\frac{\sqrt{3}(|X_k(\Omega)|+|X_k(i)|)}{\sqrt{2}\lambda_{n,k}} \right\}$$

(8)

$$p_M(X_k|H_1) = \frac{\sqrt{6}}{8\pi \sqrt{\lambda_{n,k} + \lambda_{x,k}}} |X_k(\Omega)|^{0.5} |X_k(i)|^{0.5} \times \exp\left\{ -\frac{\sqrt{3}(|X_k(\Omega)|+|X_k(i)|)}{\sqrt{2}(\lambda_{n,k} + \lambda_{x,k})} \right\}$$

(9)

3 Modifying statistical model for nonstationary noise

In [4] it has been shown a good performance in high SNR, because the threshold depends only on the background noise statistics. The lower variance in a particular spectral bin requires the lower threshold. The system performance will be better under the less time variable background noise. However, as the SNR becomes lower, the fundamental assumption in which there will be a significant shift in mean during periods of speech becomes weaker. The hangover scheme in [4] caused a lower performance and time consumption in non-stationary noise.

We present the overall performance of the proposed statistical model-based VAD. The values of speech detection probability ($P_d$) for these three models have been shown in Table 1 where the VAD algorithms were applied to the speech data corrupted by the aforementioned noises at a variety of SNRs (-10, -5, 0 and 5 dB). The choice of the value of parameters $k_L$ and $k_M$ for Laplacian and Gamma models, respectively - that describe above - in 5dB SNR shows that the best choices for these parameters are $k_L=0.9$ and $k_M=0.9$ for Laplacian and Gamma model respectively [4]. From the obtaining results, we could obtain the following observations:

1- In the case of the white noise, the Laplacian model-based VAD algorithm outperformed the other approaches. Also, the Gamma model-based resulted in a better performance than that of the Gaussian model in the most tested conditions.

2- In contrast, from the $P_d$’s shown in Table 1 looks relatively close to each other in the case of the vehicular noise. However, it is observed that the Gamma model-based VAD algorithm demonstrates a slightly better performance than the other models.

3- As shown in Table 1, for the babble noise, the VAD algorithm incorporating the Laplacian model yielded a performance superior to both the Gamma and Gaussian PDFs. Moreover, the performance difference became larger as the SNR decreased.

4 Correction factors

In this section, we propose a technique to adopt various factors for the Likelihood ratios (LRs) such as $c_k \log \Lambda_k$, which shows below, as we believe that incorporation of the different contributions of the LRs will increase the performance of the VAD:

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

(10)

where $\xi_k = \lambda_{x,k}/\lambda_{n,k}$ and $\gamma_k = Y_k/\lambda_{n,k}$ denote the a priori signal-to-noise ratio (SNR) and the a posteriori SNR, respectively [4]. The a posteriori SNR $\gamma_k$ is estimated using $\lambda_{n,k}$, and the a priori SNR is estimated by the well-known Direct Decision (DD) method as follows [5]:

$$\xi_k = \frac{\lambda_{n,k}(t-1)^2}{\lambda_{n,k}(t-1)} + (1 + \alpha) P[y_k(t) = 1]$$

(11)

where $\hat{Y}_k(t-1)^2$ is the speech spectral amplitude estimate of the previous frame obtained using the minimum mean-square error (MMSE) estimator [3]. Also, $\alpha$ is a weight determined in the range (0.95, 0.99) [1]. The function $P[x]=0$ if $x<0$ and $P[x]=1$ otherwise. The final decision in the conventional statistical model-based VADs has been established from the geometric mean of the LRs computed for the individual frequency bins [7] and is obtained by:

$$\log \Lambda = \frac{1}{M} \sum_{k=1}^{M} \log \Lambda_k(t) \approx \sum_{k=1}^{M} \eta_k$$

(12)

where an input frame is classified as speech presence if the geometric mean of the LRs is greater than a certain threshold value $\eta$ and speech absent otherwise. The factors $c_k$ is needed to satisfy the following conditions:

$$\sum_{k=1}^{M} c_k = 1, c_k \geq 0$$

(13)

Let $\Lambda_c = \frac{1}{M} \sum_{k=1}^{M} c_k \log \Lambda_k$ give the threshold value where $c_k$ is:

$$c_k = \frac{\exp(c_k)}{\sum_{k=1}^{M} \exp(c_k)}$$

(14)

We therefore adopt the following parameter transformation, which is inversely transformed to $c_k$:

$$\hat{c}_k = \log c_k$$

(15)
Tab.1 Speech detection probability in white, vehicle and babble noise in 5 dB, 0 dB, -5 dB and -10 dB.

<table>
<thead>
<tr>
<th>Noise Distribution</th>
<th>False-alarm prob.</th>
<th>-10dB</th>
<th>-5dB</th>
<th>0dB</th>
<th>5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>White</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.05</td>
<td>0.68</td>
<td>0.73</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>White</td>
<td>0.1</td>
<td>0.7</td>
<td>0.74</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>White</td>
<td>0.25</td>
<td>0.72</td>
<td>0.76</td>
<td>0.8</td>
<td>0.84</td>
</tr>
<tr>
<td>White</td>
<td>0.5</td>
<td>0.73</td>
<td>0.78</td>
<td>0.82</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Vehicle</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.05</td>
<td>0.58</td>
<td>0.64</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.1</td>
<td>0.65</td>
<td>0.69</td>
<td>0.91</td>
<td>0.98</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.25</td>
<td>0.76</td>
<td>0.79</td>
<td>0.92</td>
<td>0.99</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.5</td>
<td>0.85</td>
<td>0.91</td>
<td>0.94</td>
<td>1</td>
</tr>
<tr>
<td><strong>Babble</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Babble</td>
<td>0.05</td>
<td>0.4</td>
<td>0.48</td>
<td>0.5</td>
<td>0.51</td>
</tr>
<tr>
<td>Babble</td>
<td>0.1</td>
<td>0.5</td>
<td>0.55</td>
<td>0.6</td>
<td>0.64</td>
</tr>
<tr>
<td>Babble</td>
<td>0.25</td>
<td>0.61</td>
<td>0.74</td>
<td>0.76</td>
<td>0.8</td>
</tr>
<tr>
<td>Babble</td>
<td>0.5</td>
<td>0.75</td>
<td>0.82</td>
<td>0.86</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Let denote the set of estimations for the transformed factors at time. Then, it is updated based on the steepest descent algorithms as follows:

\[
\hat{c}_k(t + 1) = \hat{c}_k(t) - \varepsilon \frac{\partial L(t)}{\partial \hat{c}_k} |_{\hat{c}_k=\hat{c}_k(t)}
\]  

The GPD approach approximates the empirical classification error by a smooth objective function, which is the 0-1 step loss function defined by:

\[
\text{Let denote the set of estimations for the transformed factors at time. Then, it is updated based on the steepest descent algorithms as follows:}
\]
Fig. 1 ROC curves and modified ROC curves for white noise in 5 dB, 0 dB, -5 dB, and -10 dB.

---

Fig. 2 ROC curves and modified ROC curves for vehicle noise in 5 dB, 0 dB, -5 dB, and -10 dB.

---

Fig. 3 ROC curves and modified ROC curves for babble noise in 5 dB, 0 dB, -5 dB, and -10 dB.

\[
L(t) = \begin{cases} 
\frac{1}{1 + \exp[-2y(\theta - \Lambda_c)]} & \text{if } H_1 \text{ occurred} \\
\frac{1}{1 + \exp[2y(\theta - \Lambda_c)]} & \text{if } H_0 \text{ occurred}
\end{cases} \tag{17}
\]

where \(\gamma\) denotes the gradient of the sigmoid function. Therefore the reformed system performed by:

\[
\Lambda_c = \frac{1}{M} \sum_{k=1}^{M} c_k \log \Lambda_k \geq H_2 \eta \tag{18}
\]

5 Experimental results

In order to evaluate the performance of the proposed algorithm, we added the white, vehicular, and babble noises from the NOISEX-92 database [7] to the clean speech with varying SNR. The VAD test was carried out for each 10ms frame in length.

The parameters used for defining the objective function \(L\) were selected such that \(\gamma = 1\) and the step size for parameter update was set to \(\epsilon = 1 - (t)/4000\). In practice, a threshold value of the combined score was set to 0 as the experimentally chosen boundary in the middle of \(\Lambda_s\) stemming from speech and \(\Lambda_n\) stemming from noise. As a result, Fig.1, 2 and 3 show the ROC curves for the proposed algorithm compared with the conventional Laplacian methods using white, vehicle and babble noises, respectively, in 5 dB, 0 dB, -5 dB and -10 dB.

Finally, among the different sets of the factors, we selected only a single set of the factors as a representative case which is obtained based on an observation that the weights under each training condition seem to be quite similar.

6 Conclusion

An extensive study and experiments on the statistical models under Low SNR conditions have made it possible to understand that the complex Laplacian and Gamma PDFs could be strong candidates for a parametric representation of the noisy speech spectra distribution. We select the Laplacian model and trying to improve its PDF. The results show a good modification especially in case of Gaussian and babble noise.

References:


