Comparison of Statistical Model-Based Voice Activity Detectors for Mobile Robot Speech Applications ⋆

Ivan Marković * **Hrvoje Domitrović** * **Ivan Petrović** *

* *University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia, (e-mail: {ivan.markovic, hrvoje.domitrovic, ivan.petrovic}@fer.hr)*

Abstract: This paper deals with the problem of voice activity detection in adverse acoustic conditions, namely high and varying noise scenarios. For robotic applications, we need the voice activity detector to be computationally light, robust to varying levels of background noise, and have a low latency, especially if we are tracking moving speakers. We analyze three voice activity detectors—two model the discrete Fourier transform coefficients by Gaussian and generalized Gaussian distribution, while the third models the spectral envelope as having either Rayleigh or Rice distribution—and we present them in a unifying and consistent manner, with respect to a statistical hypotheses ratio measure and a joint noise spectrum estimation algorithm. Moreover, we compare the performance under various noise conditions; three types of noises, three different signal-to-noise ratios and six different speakers, by means of receiver operating characteristic curves and area under a curve score. The results showed that the Rayleigh-Rice model had on average better results and medium computational demand.

Keywords: voice activity detection, statistical model-based detectors, receiver operating characteristic curves.

1. INTRODUCTION

Voice activity detection is a technique in speech processing by which presence of speech is detected in a given signal frame. This problem can be seen as a dual hypothesis problem, where a signal frame is classified as either containing speech or containing noise. In a voice activity detector (VAD), the absence of speech usually presumes presence of noise only. This system is not only of great importance for many applications, like mobile telephony, Internet telephony, hearing aid devices, but also for robotics if speech oriented systems are utilized like speaker localization, speech and speaker recognition. For most of the stated research problems, it is indispensable to save on bandwidth resources by coding noise with significantly less bits, while for others it is mandatory to completely ignore frames with noise.

A VAD must provide a robust and reliable decision procedure in varying acoustical conditions. This task gets quite formidable with the varying level and type of background noise, like in the case of a mobile robot. Furthermore, voice activity detection often serves as a front-end algorithm for other applications and it is difficult to set algorithm constrains without knowing what the total system will be like. But for some applications, like speaker tracking, the detector should be computationally light and have low latency in order to keep a feasible track, while for applications like speech recognition, where a big delay is already present, low latency and real-time operation might not be so crucial.

Approaches to voice activity detection mostly differ in the type of the extracted features and in the decision models used to reach a speech/non-speech decision based on those features. A lot of attention is given to statistical model-based VADs, in which certain probabilistic properties are assumed on the coefficients of the discrete Fourier transform (DFT). For an example, in Sohn et al. (1999) they are assumed to have Gaussian distribution and this approach was further developed in Cho et al. (2001); Chang and Kim (2003); Chang et al. (2004); Ramírez et al. (2007) and Górriz et al. (2010). Furthermore, special attention is given to derivation of various noise robust features and decision rules in Woo et al. (2000); Li et al. (2002) and Marzinzik and Kollmeier (2002). However, VAD is not necessarily limited to single channel processing. In Valin et al. (2007) the authors used multichannel post-filter for speech recognition of multiple speakers. Although the multichannel approach can undoubtedly further enhance VADs, in the present paper we have focused on the comparison of statistical model-based VADs.

In this paper, we have implemented and compared three statistical model-based VAD algorithms that were originally presented by Sohn et al. (1999); Chang et al. (2004) and Mumolo et al. (2003), respectively. The three VAD algorithms are presented in a unifying and consistent manner by using a joint noise spectrum estimation technique. Performance of the algorithms is tested and compared un-

⋆ This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No. 036-0363078-3018 and European Community's Seventh Framework Programme under grant agreement no. 285939 (ACROSS).

der varying noise conditions, namely three types of noises and three different signal-to-noise ratios (SNRs).

The rest of the paper is organized as follows. Section 2 presents the statistical model-based VADs. In Section 3, the implemented algorithms for noise spectrum estimation and *a priori* signal-to-noise ratio are presented. Section 4 presents the experimental comparison of the algorithms, and Section 5 concludes the paper.

2. STATISTICAL MODEL-BASED DETECTORS

These VADs rely on statistical modeling of the DFT coefficients. All the statistical model-based VADs assume a two hypotheses scenario. Since speech is degraded with uncorrelated additive noise, the two hypotheses are as follows:

$$\begin{aligned} H_0 &: \text{speech absent} \Rightarrow \mathbf{X} = \mathbf{N} \\ H_1 &: \text{speech present} \Rightarrow \mathbf{X} = \mathbf{N} + \mathbf{S}, \end{aligned} \tag{1}$$

where $\mathbf{X} = [X_0, X_1, \ldots, X_{K-1}]^{\mathrm{T}}$, $\mathbf{N} = [N_0, N_1, \ldots, N_{K-1}]^{\mathrm{T}}$ and $\mathbf{S} = [S_0, S_1, \ldots, S_{K-1}]^{\mathrm{T}}$ are the DFT coefficients of a K-point DFT of the noisy speech, noise, and clean speech, respectively.

The form of the probability density function (pdf) of $\mathbf{X}$ conditioned on the hypotheses, i.e. $p(\mathbf{X}|H_0)$ and $p(\mathbf{X}|H_1)$, depends on the distribution used to model each DFT coefficient. In this paper three different distributions are presented and analyzed.

After the pdfs $p(\mathbf{X}|H_0)$ and $p(\mathbf{X}|H_1)$ are determined, usually a likelihood ratio (LR) on all the DFT coefficient indices k is calculated:

$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)}, \tag{2}$$

where Λ_k becomes a vector of length K. This information is then used to calculate the geometric mean which is then compared to a certain threshold in order to reach a final decision in favor of either the hypothesis H_0 or H_1:

$$\log \Lambda = \frac{1}{K} \sum_{k=0}^{K-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3}$$

The goal of statistical model-based VADs is to model the DFT coefficients, or their derivatives, as faithfully as possible and in essence, the advantage of each model would be in the representation faithfulness thereof.

2.1 Gaussian distribution statistical model

This VAD was first proposed by Sohn et al. (1999), where the DFT coefficients are asymptotically independent and zero-mean complex Gaussian random variables. Let us look at the DFT of the clean speech signal. In the complex Gaussian speech model, both the real and the imaginary parts of the DFT, $S_k = S_{R,k} + jS_{I,k}$, are independent zero-mean Gaussian random variables, each with a variance of $\lambda_{s,k}/2$. The pdfs of the coefficients are:

$$p(S_{R,k}) = \frac{1}{\sqrt{\pi \lambda_{s,k}}} \exp\left\{ -\frac{S_{R,k}^2}{\lambda_{s,k}} \right\}, \tag{4}$$

$$p(S_{I,k}) = \frac{1}{\sqrt{\pi \lambda_{s,k}}} \exp\left\{ -\frac{S_{I,k}^2}{\lambda_{s,k}} \right\}. \tag{5}$$

Since real and imaginary coefficients are independent, joint pdf can be written in the following form:

$$\begin{aligned} p(S_k) = p(S_{R,k}) p(S_{I,k}) &= \frac{1}{\pi \lambda_{s,k}} \exp\left(-\frac{S_{R,k}^2 + S_{I,k}^2}{\lambda_{s,k}} \right) \\ &= \frac{1}{\pi \lambda_{s,k}} \exp\left(-\frac{|S_k|^2}{\lambda_{s,k}} \right). \end{aligned} \tag{6}$$

Similar derivation can be done for the pdf of the noise coefficients.

When both speech and noise are present, we have for each coefficient a sum of independent Gaussian variables, thus resulting with a pdf of variance $\lambda_{x,k} = \lambda_{n,k} + \lambda_{s,k}$. Hence, the conditional pdfs of X_k on hypotheses H_0 and H_1 are as follows:

$$p(X_k|H_0) = \frac{1}{\pi \lambda_{n,k}} \exp\left\{ -\frac{|X_k|^2}{\lambda_{n,k}} \right\}, \tag{7}$$

$$p(X_k|H_1) = \frac{1}{\pi(\lambda_{n,k} + \lambda_{s,k})} \cdot \exp\left\{ -\frac{|X_k|^2}{\lambda_{n,k} + \lambda_{s,k}} \right\}. \tag{8}$$

Under the Gaussian distribution model, the LR is simply calculated as the ratio of (8) and (7):

$$\Lambda_k^{\mathrm{GD}} = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{ \frac{\gamma_k \xi_k}{1+\xi_k} \right\}, \tag{9}$$

where $\xi_k = \lambda_{s,k}/\lambda_{n,k}$ is the *a priori* SNR, and $\gamma_k = |X_k|^2/\lambda_{n,k}$ is the *a posteriori* SNR. The algorithms for estimation of these values are presented in Section 3.

2.2 Generalised Gaussian distribution statistical model

In Chang and Kim (2003) statistical model-based VAD was improved by incorporating a complex Laplacian model. The analysis in the latter paper showed that the Laplacian provides a better model of the distribution of noisy speech spectra, than the Gaussian model. Furthermore, VAD based on generalized Gaussian distribution (GGD), which includes the Gaussian and Laplacian model as special cases, was proposed by Chang et al. (2004), where it was also experimentally verified that VAD based on GGD outperforms the VAD based on the Laplacian model. Following the same train of thought as in Section 2.1, joint GGD of the DFT coefficients for clean speech signal is given by:

$$\begin{aligned} p(S_k) =& \frac{\nu^2 \alpha^2(\nu)}{4 \lambda_{s,k} \Gamma^2(1/\nu)} \cdot \\ & \cdot \exp\left\{ -\alpha^{\nu}(\nu) \left[\left| \frac{S_{R,k}}{\sqrt{\lambda_{s,k}}} \right|^{\nu} + \left| \frac{S_{I,k}}{\sqrt{\lambda_{s,k}}} \right|^{\nu} \right] \right\}, \end{aligned} \tag{10}$$

with

$$\alpha(\nu) = \sqrt{\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}}, \tag{11}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and ν denotes parameter controlling the distribution shape. For $\nu = 1$ and 2 the GGD becomes the Laplacian and Gaussian density, respectively.

The shape parameter ν needs to be continuously estimated. By letting m_1 and m_2 be the first and the second moment of $|X_k|$ (cf. Chang et al. (2004)), ν can be estimated by solving the following equation:

$$\hat{\nu} = F^{-1}\left(\frac{m_1}{m_2}\right), \tag{12}$$

where

$$F(x) = \frac{\Gamma(2/x)}{\sqrt{\Gamma(1/x)\Gamma(3/x)}}. \tag{13}$$

The (12) is the inverse of (13) and is usually solved by precomputing a lookup table.

From the previous discussion we can write the distribution of X_k conditioned on the hypotheses H_0 and H_1 as follows:

$$p(X_k|H_0) = \frac{\nu_{n,k}^2\alpha^2(\nu_{n,k})}{4\lambda_{n,k}\Gamma^2(1/\nu_{n,k})}\exp\left\{-\alpha^{\nu_{n,k}}(\nu_{n,k}) \cdot\left[\left|\frac{X_{R,k}}{\sqrt{\lambda_{n,k}}}\right|^{\nu_{n,k}} + \left|\frac{X_{I,k}}{\sqrt{\lambda_{n,k}}}\right|^{\nu_{n,k}}\right]\right\} \tag{14}$$

$$p(X_k|H_1) = \frac{\nu_{s,k}^2\alpha^2(\nu_{s,k})}{4(\lambda_{s,k}+\lambda_{n,k})\Gamma^2(1/\nu_{s,k})}\exp\left\{-\alpha^{\nu_{n,k}}(\nu_{n,k}) \cdot\left[\left|\frac{X_{R,k}}{\sqrt{\lambda_{s,k}+\lambda_{n,k}}}\right|^{\nu_{n,k}} + \left|\frac{X_{I,k}}{\sqrt{\lambda_{s,k}+\lambda_{n,k}}}\right|^{\nu_{n,k}}\right]\right\}, \tag{15}$$

where $\nu_{n,k}$ and $\nu_{s,k}$ are shape parameters related to H_0 and H_1 of noisy speech on frequency bin k, respectively. In order to compute these parameters, the corresponding $(m_{1,k}^n, m_{2,k}^n)$ and $(m_{1,k}^s, m_{2,k}^s)$ are calculated recursively from $|X_k|$ as proposed in Chang et al. (2004).

Finally, we can write the LR for the GGD model:

$$\begin{aligned}\Lambda_k^{\mathrm{GGD}} =& \frac{1}{1+\xi_k}\cdot\frac{\nu_{s,k}^2\alpha^2(\nu_{s,k})\Gamma^2(1/\nu_{n,k})}{\nu_{n,k}^2\alpha^2(\nu_{n,k})\Gamma^2(1/\nu_{s,k})}\exp\Bigg\{\\ &-\alpha^{\nu_{s,k}}(\nu_{s,k})\left[\frac{|X_{R,k}|^{\nu_{s,k}}+|X_{I,k}|^{\nu_{s,k}}}{\left(\sqrt{\lambda_{n,k}(1+\xi_k)}\right)^{\nu_{s,k}}}\right] + \\ &+\alpha^{\nu_{n,k}}(\nu_{n,k})\left[\frac{|X_{R,k}|^{\nu_{n,k}}+|X_{I,k}|^{\nu_{n,k}}}{\left(\sqrt{\lambda_{n,k}}\right)^{\nu_{n,k}}}\right]\Bigg\}.\end{aligned} \tag{16}$$

2.3 Rayleigh and Rice distribution statistical model

In the approach proposed by Mumolo et al. (2003), derived from McAulay and Malpass (1980), the DFT coefficients are still modelled as having a Gaussian distribution, but instead of using their joint distribution, the distribution of the signal envelope is used. The envelope of a signal, $|X_k| = \sqrt{X_{R,k}^2 + X_{I,k}^2}$, is actually the euclidean norm of the real and imaginary coefficients. Therefore, instead of looking at the distribution of the coefficients, the distribution of the signal envelope is analysed.

Under hypothesis H_0 the signal is only noise, which means that the DFT coefficients are both independent, zero-mean Gaussian variables with variance $\lambda_{n,k}/2 = \mathrm{E}[|N_k|^2]$. Under that assumption, the pdf of the euclidean distance of such DFT coefficients is a Rayleigh distribution:

$$p(X_k|H_0) = \frac{2|X_k|}{\lambda_{n,k}}\exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\}. \tag{17}$$

Under hypothesis H_1, the envelope is the euclidean norm of two independent, non-zero-mean Gaussian variables. Such pdf is a Rician:

$$\begin{aligned}p(X_k|H_1) =& \frac{2|X_k|}{\lambda_{n,k}}\exp\left\{-\frac{1}{\lambda_{n,k}}\left(|X_k|^2+|A_k|^2\right)\right\} \cdot I_0\left\{\frac{2|A_k||X_k|}{\lambda_{n,k}}\right\}\\ =& \frac{2|X_k|}{\lambda_{n,k}}\exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}-\xi_k\right\} \cdot I_0\left\{2\sqrt{\xi_k\frac{|X_k|^2}{\lambda_{n,k}}}\right\},\end{aligned} \tag{18}$$

where A_k is the amplitude of the clean speech spectrum, $\xi_k = |A_k|^2/\lambda_{n,k}$ is the *a priori* SNR and $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero. In Mumolo et al. (2003) this VAD was implemented by calculating the *a posteriori* probability $p(H_1|X_k)$ of voice activity from (17) and (18) via Bayes' formula. Since in this paper the *a priori* SNR estimation, presented in Section 3, for all frequency bins is implemented, we are proposing the LR instead of the *a posteriori* probability $p(H_1|X_k)$.

Finally, we derive the LR for Rayleigh and Rice distribution (RRD) model:

$$\Lambda_k^{\mathrm{RRD}} = \exp\{-\xi_k\}\, I_0\left\{2\sqrt{\xi_k\gamma_k}\right\}. \tag{19}$$

3. NOISE SPECTRUM ESTIMATION

We can see from previous sections that all VADs require estimation of the noise spectrum $\lambda_{n,k}$ and the *a priori* SNR ξ_k. First we shall address the estimation of $\lambda_{n,k}$ and then the estimation of ξ_k.

In most VADs the noise spectrum estimation is done in a way to assume that in the first several frames only noise is present and for that time $\lambda_{n,k}$ is estimated by time averaging the spectrum of the recorded signal. Then, the VAD itself is used to discriminate between frames where speech is present and where only noise is present. When only noise is detected, $\lambda_{n,k}$ is again estimated in a time-averaging fashion.

In this paper the *minima-controlled recursive averaging* (MCRA) algorithm, proposed by Cohen and Berdugo (2001) and Cohen (2003), is used since it performs well in varying noise situations and it allows estimation from all frames, and not just the ones where no speech is detected.

3.1 Minima-controlled recursive averaging

As stated earlier, a common technique for noise spectrum estimation is to apply temporal recursive smoothing during the frames when only noise is present. Now, we have the following hypotheses:

$$\begin{aligned}H_0 &: \lambda_{n,k}(l+1) = a_n\lambda_{n,k}(l) + (1-a_n)|X_k(l)|^2,\\ H_1 &: \lambda_{n,k}(l+1) = \lambda_{n,k}(l),\end{aligned} \tag{20}$$

where $0 < a_n < 1$ is a smoothing parameter.

Let $p_{s,k}(l) = p(H_1|X_k(l))$ denote the conditional speech presence probability at time frame l. Hence, we can write (20) as follows:

$$\begin{aligned}\lambda_{n,k}(l+1) =& \lambda_{n,k}(l)p_{s,k}(l) + \left[a_n\lambda_{n,k}(l) + \right.\\ &\left. + (1-a_n)|X_k(l)|^2\right](1-p_{s,k}(l))\\ =& \tilde{a}_{n,k}(l)\lambda_{n,k}(l) + (1-\tilde{a}_{n,k}(l))|X_k(l)|^2,\end{aligned} \tag{21}$$

where

$$\tilde{a}_{n,k}(l) = a_n + (1 - a_n)p_{s,k}(l) \quad (22)$$

is a time-varying smoothing parameter. We can see that the noise spectrum is estimated by averaging past power spectral values, using a smoothing parameter that is adjusted by the speech presence probability $p_{s,k}(l)$. In order to determine $p_{s,k}(l)$, speech absence is calculated by looking at the ratio of the local energy of the noisy signal and its minimum within a certain time frame.

Firstly, the squared magnitude of the spectrum is defined:

$$S_{f,k}(l) = |X_k(k)|^2, \quad (23)$$

which could be smoothed in the frequency domain, but we have omitted this step due to the increase it brings to computational complexity. However, we do smooth the spectrum in the time domain:

$$S_k(l) = \alpha_s S_k(l-1) + (1 - \alpha_s)S_{f,k}(l), \quad (24)$$

where $0 < \alpha_s < 1$ is a smoothing parameter. The minimum of the local energy of the noisy signal is calculated by first initializing the minimum and temporary local variable: $S_{\min,k}(0) = S_k(0)$ and $S_{\text{tmp},k}(0) = S_k(0)$, respectively. Then, the minimum value of the squared amplitude spectrum is tracked in time:

$$S_{\min,k}(l) = \min\{S_{\min,k}(l-1), S_k(l)\}, \quad (25)$$

$$S_{\text{tmp},k}(l) = \min\{S_{\text{tmp},k}(l-1), S_k(l)\}. \quad (26)$$

Whenever the number of frames reaches an arbitrarily chosen M, the temporary variable is initialized by:

$$S_{\min,k}(l) = \min\{S_{\text{tmp},k}(l-1), S_k(l)\}, \quad (27)$$

$$S_{\text{tmp},k}(l) = S_k(l). \quad (28)$$

We can see that the parameter M determines the scope of the local minima search, and that the temporary variable insures that the minimum will be adapted to a change in the noise level.

For calculating the conditional speech presence probability $p_{s,k}(l)$ a decision rule based on the ratio of the local energy of the noisy signal and its minimum, $S_{r,k}(l) = S_k(l)/S_{\min,k}(l)$, is needed:

$$S_{r,k}(l) \underset{H_0}{\overset{H_1}{\gtrless}} \delta. \quad (29)$$

In Cohen and Berdugo (2001) the following estimator for $p_{s,k}(l)$ was proposed:

$$p_{s,k}(l) = \alpha_p p_{s,k}(l-1) + (1 - \alpha_p)I_k(l), \quad (30)$$

where $0 < \alpha_p < 1$ is a smoothing parameter, and $I_k(l)$ is an indicator function for the result in (29), i.e. $I_k(l) = 1, \forall k$ if $S_{r,k}(l) > \delta$ and $I_k(l) = 0, \forall k$ if $S_{r,k}(l) < \delta$. At this point, we have calculated all the variables needed for the estimation of the noise spectrum via (21).

3.2 Decision directed a priori SNR estimation

The decision directed (DD) estimation approach for the estimation of ξ_k, the *a priori* SNR, was proposed in Ephraim and Malah (1984). Firstly, the Wiener gain is introduced as the following ratio:

$$\zeta_k = \frac{\xi_k}{\xi_k + 1}. \quad (31)$$

Now, we can define the estimator for ξ_k:

$$\xi_k(l) = \alpha_a \zeta_k(l-1)^2 \gamma_k(l-1) + (1 - \alpha_a)\max\{\gamma_k(l) - 1, 0\}, \quad (32)$$

where $0 < \alpha_a < 1$ is a smoothing parameter.

The noise spectrum $\lambda_{n,k}$ and the *a priori* SNR ξ_k are continuously updated via the MCRA and DD methods, respectively, and are afterwards used in the VAD algorithms.

4. EXPERIMENTAL COMPARISON OF THE VAD ALGORITHMS

In order to analyze the VADs and their performance, we used the NOIZEUS speech corpus by Hu and Loizou (2007). Although the corpus was originally created for testing speech enhancement algorithms, we used it for the following reasons: (i) the recordings are of high quality and were made in a sound-proof booth, (ii) it offers eight different types of noises from AURORA database by Pearce and Hirsch (2000) which corrupt the original recordings at four different SNR levels, (iii) the recordings were made by six different speakers – three male and three female, (iv) it uses the IEEE sentence database which contains phonetically-balanced sentences with relatively low word-context predictability, and (v) the corpus is available to researchers free of charge.

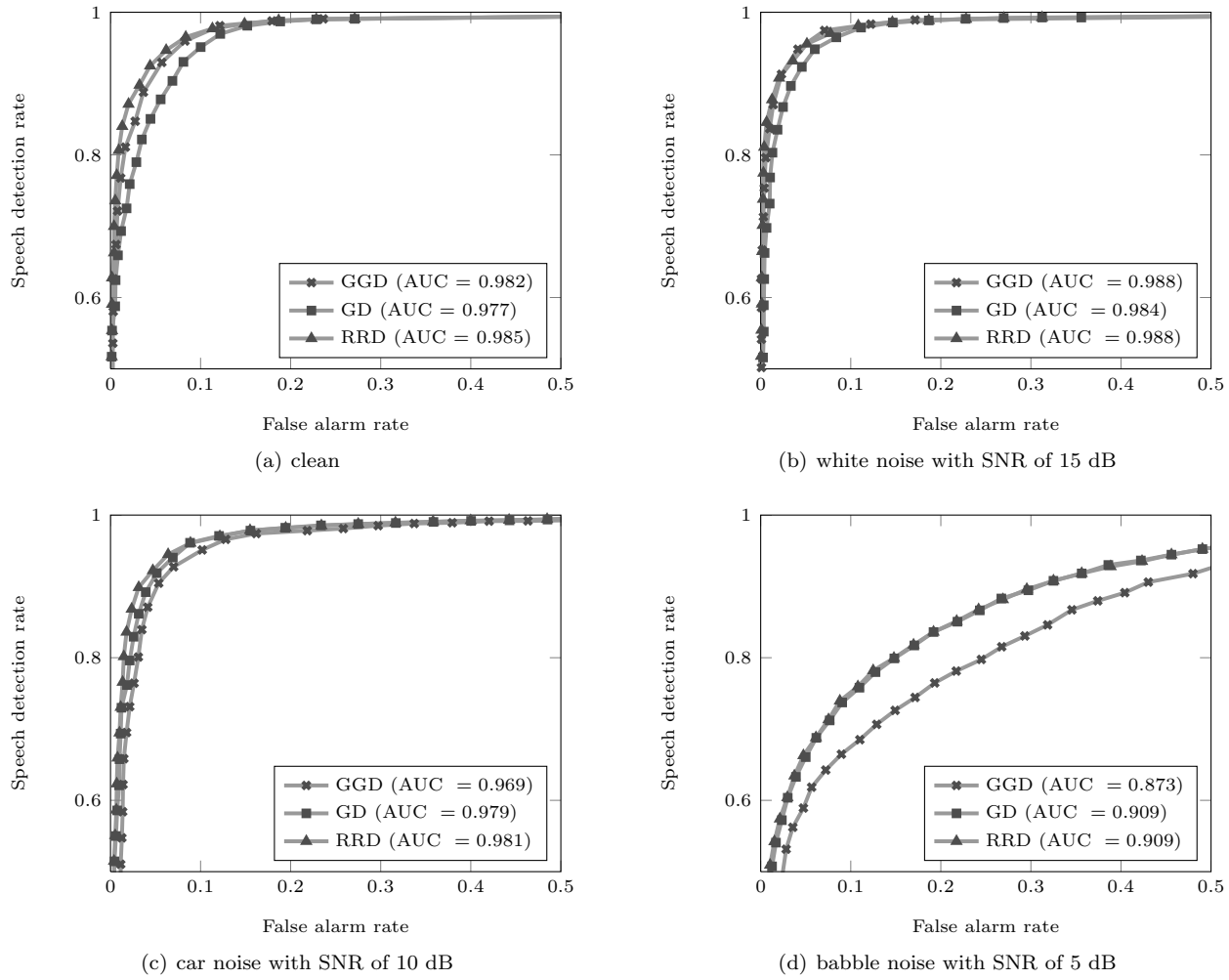
The recordings were sampled at the rate of 25 kHz and were later downsampled to 8 kHz. The total length of all the recordings was 80.04 s, which offered, with overlap and frame length of $L = 256$, in total 5000 frames for detection. The percentage of the speech segments amounted to 61.28%, which is as twice as high as compared to Sohn et al. (1999), and Chang and Kim (2003), but less than 5% higher than in the cases of Chang et al. (2004) and Górriz et al. (2010).

Usually, in order to test the algorithms, the speech segments are hand-labeled. However, in the present work we used signal energy calculated via Parseval's theorem as the indicator of speech presence, which enabled automatic frame labeling. We find this approach justifiable in the case of the NOIZEUS corpus, since the clean recordings were made in a sound-proof booth resulting with the speech-absent frames having energy a thousand times lower than the weakest speech frame.

4.1 Receiver operating characteristics

When analyzing detector performance it is common to utilize receiver operating characteristics (ROC) curves. The ROC curves give a practical representation of the detector performance, by depicting the relationship of the following two rates: speech detection rate (SDR)—the rate of correctly detected speech frames in speech-labeled frames, and false alarm rate (FAR)—the rate of frames detected as speech in noise-labeled frames. To compare different detectors it is practical to reduce the information in the ROC curve to a single scalar value. A common method is to evaluate the area under an ROC curve (AUC).

The results are shown in Fig. 1. By analyzing Figures 1(a) to 1(b) we can see that in the lower SNR scenarios the GGD and RRD mostly outperform the GD VAD . On the contrary, in Fig. 1(d) under very low SNR, the GD and RRD VAD show similar performance, and basically better results than the GGD VAD. More elaborate methods for estimation of the shape parameter ν (cf. Kokkinakis



(a) clean

(b) white noise with SNR of 15 dB

(c) car noise with SNR of 10 dB

(d) babble noise with SNR of 5 dB

Fig. 1. ROC curves for the three VADs. Each figure represents a different type of noise and a different SNR level.

and Nandi (2005)) could improve the performance of the GGD VAD, but comparison of such methods is out of the scope of the present paper. We can see that with the changing SNR and noise type, the performance of the VADs relative to each other changes. But still, from Fig. 1, we can conclude that the RRD VAD shows equal or better performance than the other VADs in all four scenarios, and that preliminarily it seems as the best choice.

In Fawcett (2004), it is suggested to generate results from several test subsets and average these results in order to obtain a measure of variance. The ROC curves can be either averaged vertically—by fixing FAR and averaging over SDR, or by the threshold—for each threshold value an SDR–FAR pair is found and their values are averaged thus yielding both vertical and horizontal variance. The test set for this experiment was constructed by concatenating the clean signal with its corrupted versions thus, with frame length of $L = 256$ samples, yielding 50000 examples for evaluation. In the present paper we used 10-fold cross-validation procedure and threshold averaging. In Fig. 2 we can see the results of the experiment. Each point in the ROC curve also depicts a horizontal and vertical error bars which correspond to a value of three standard deviations. Moreover, in the legend we can also seesd the AUC score along with one standard deviation. By analyzing Fig. 2 we can assert that none of the detectors exhibited large deviations and thus they all performed consistently on all the subsets, and that the RRD VAD, on average, had the best performance.

Another important parameter that should be analyzed is the computational demand, since we can see that (9), (16), and (19) differ in complexity. The execution times of all the VADs (without the MCRA and the DD SNR estimation), was measured for Matlab implementations on an Intel Core2Quad processor with 2.33 GHz frequency (only one core was used). The results were as follows: the GGD, RRD, and GD VAD had the execution times of 9.70 ms, 0.37 ms, and 0.21 ms, respectively. The reason behind the much higher computational complexity of the GGD VAD lies in the need to evaluate (12) via lookup table twice. Without this step, the GGD VAD takes on average 0.90 ms, which is still twice as much as the RRD VAD. However, a faster time varying estimate of the shape parameter ν could be utilized (cf. Krupiski and Purczyski (2006)) to lower the computational complexity.

The reader should note at this point, that in the present paper we have implemented the VADs somewhat differently than when they were first proposed by Sohn et al. (1999); Chang et al. (2004) and Mumolo et al. (2003).

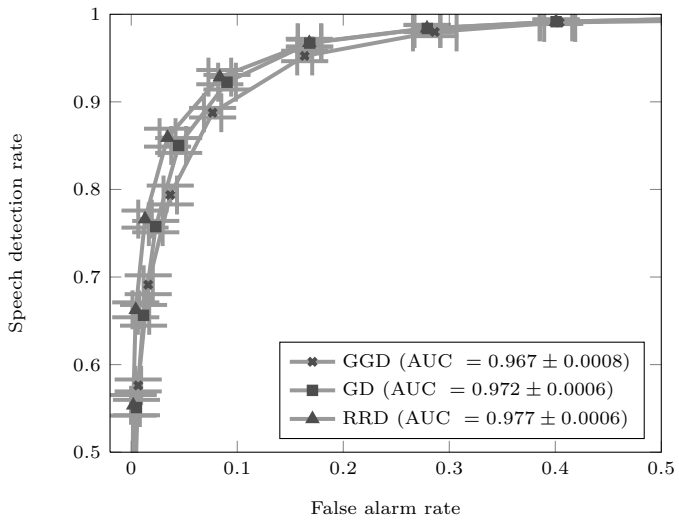


Fig. 2. Threshold averaged ROC curves with AUC scores

Mostly, the difference is in the noise spectrum and the *a priori* SNR estimation, and in the case of the RRD VAD, in the introduction of the LR for that model. Furthermore, the algorithms did exhibit some variance in performance with respect to changes in some of the smoothing parameters, but however this did not cause a change in relative performance of the detectors.

5. CONCLUSION

In this paper we have presented three different voice activity detection algorithms in an unifying and consistent manner, by incorporating noise spectrum and the *a priori* signal-to-noise ratio estimation to their respective frameworks. Furthermore, we introduced the LR for the Rayleigh and Rice distribution based detector. The decision framework was based on a statistical hypothesis ratio measure, and its geometric mean over all the DFT coefficient indices. The algorithms were tested on the NOIZEUS speech corpus which consisted of clean recordings, and its versions corrupted with three types of noises and three different SNRs. The performance analysis was conducted using threshold averaged ROC curves and AUC score. Based of the aforementioned parameters, and the computational complexity, we concluded that the VAD based on Rayleigh and Rice distribution showed the best performance on average and is the most suitable among the tested algorithms.

REFERENCES

Chang, J.H., Shin, J.W., and Kim, N.S. (2004). Voice Activity Detector Employing Generalised Gaussian Distribution. *Electronics Letters*, 40(24), 25–26.

Chang, J.H. and Kim, N.S. (2003). Voice Activity detection Based on Complex Laplacian Model. *Electronics Letters*, 39(7), 632.

Cho, Y.D., Al-Naimi, K., and Kondoz, A. (2001). Improved Voice Activity Detection Based on a Smoothed Statistical Likelihood Ratio. In *Proceeding of the International Conference on Acoustics, Speech and Signal Processing*, 737–740.

Cohen, I. (2003). Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging. *Speech and Audio Processing*, 11, 466–475.

Cohen, I. and Berdugo, B. (2001). Speech Enhancement for Non-Stationary Noise Environments. *Signal Processing*, 81, 283–288.

Ephraim, Y. and Malah, D. (1984). Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *Speech and Signal Processing*, 1109–1121.

Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Labs Tech Report.

Górriz, J.M., Ramírez, J., Lang, E.W., Puntonet, C.G., and Turias, I. (2010). Improved Likelihood Ratio Test Based Voice Activity Detector Applied to Speech Recognition. *Speech Communication*, 52, 664–677.

Hu, Y. and Loizou, P.C. (2007). Subjective Comparison and Evaluation of Speech Enhancement Algorithms. *Speech Communication*, 49(7), 588–601.

Kokkinakis, K. and Nandi, A.K. (2005). Speech Modelling based on Generalized Gaussian Probability Density Functions. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 1, 381–384.

Krupiski, R. and Purczyski, J. (2006). Approximated Fast Estimator for the Shape Parameter of Generalized Gaussian Distribution. *Signal Processing*, 86(2), 205–211.

Li, Q., Zheng, J., Tsai, A., and Zhou, Q. (2002). Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition. *IEEE Transactions on Speech and Audio Processing*, 10(3), 146–157.

Marzinzik, M. and Kollmeier, B. (2002). Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics. *IEEE Transactions on Speech and Audio Processing*, 10(6), 341–351.

McAulay, R. and Malpass, M. (1980). Speech Enhancement Using a Soft-Decision Noise Suppression Filter. *IEEE Transactions on Acoustics Speech and Signal Processing*, 28, 137–145.

Mumolo, E., Nolich, M., and Verchelli, G. (2003). Algorithms for Acoustic Localization Based on Microphone Array in Service Robotics. *Robotics and Autonomous Systems*, 42(2), 69–88.

Pearce, D. and Hirsch, H.G. (2000). The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In *ISCA ITRW ASR2000*, 29–32.

Ramírez, J., Segura, J.C., Górriz, J.M., and García, L. (2007). Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition. *IEEE Transactions on Audio Speech and Language Processing*, 15(8), 2177–2189.

Sohn, J., Kim, N.S., and Sung, W. (1999). A Statistical Model-Based Voice Activity Detection. *IEEE Signal Processing Letters*, 6(1), 1–3.

Valin, J.M., Yamamoto, S., Rouat, J., Michaud, F., Nakadai, K., and Okuno, H.G. (2007). Robust Recognition of Simultaneous Speech by a Mobile Robot. *IEEE Transactions on Robotics*, 23(4), 742–752.

Woo, K.H., Yang, T.Y., Park, K.Y., and Lee, C. (2000). Robust Voice Activity Detection Algorithm for Estimating Noise Spectrum. *Electronics Letters*, 36(2), 180–181.