applied sciences

MDPI

Article

Auditory Device Voice Activity Detection Based on Statistical Likelihood-Ratio Order Statistics

Seon Man Kim

Korea Photonics Technology Institute, Gwangju 61007, Korea; smkim@kopti.re.kr; Tel.: +82-62-605-9262; Fax: +82-605-9259

Received: 12 June 2020; Accepted: 20 July 2020; Published: 22 July 2020

Abstract: This paper proposes a technique for improving statistical-model-based voice activity detection (VAD) in noisy environments to be applied in an auditory hearing aid. The proposed method is implemented for a uniform polyphase discrete Fourier transform filter bank satisfying an auditory device time latency of 8 ms. The proposed VAD technique provides an online unified framework to overcome the frequent false rejection of the statistical-model-based likelihood-ratio test (LRT) in noisy environments. The method is based on the observation that the sparseness of speech and background noise cause high false-rejection error rates in statistical LRT-based VAD—the false rejection rate increases as the sparseness increases. We demonstrate that the false-rejection error rate can be reduced by incorporating likelihood-ratio order statistics into a conventional LRT VAD. We confirm experimentally that the proposed method relatively reduces the average detection error rate by 15.8% compared to a conventional VAD with only minimal change in the false acceptance probability for three different noise conditions whose signal-to-noise ratio ranges from 0 to 20 dB.

Keywords: voice activity detection; likelihood-ratio test; order statistics; statistical model; false rejection; auditory device; hearing aid

1. Introduction

The goal of voice activity detection (VAD) is to detect the presence or absence of speech in a sound signal. VAD is increasingly difficult in noisy situations, especially for nonstationary noise such as babble noise. VAD has steadily gained research interest in the speech community in recent years, especially for applications such as selectively encoding and transmitting data in telecommunications, estimating noise statistics in speech enhancement, and detecting endpoints in speech recognition [1–3]. We focus on VAD's function in auditory hearing aid speech processing.

Individuals with hearing impairment have difficulty understanding relevant speech content in their daily lives. Attempts have been made to address this problem using auditory devices such as hearing aids, which are widely used to compensate for hearing loss and match the dynamic range [4]. However, many individuals avoid using hearing aids, often because of noise contamination of the speech signal entering the ear; only 23% of hearing-impaired people use hearing aid auditory devices [5–7]. This has motivated progress to improve complex speech perception for hearing aid users by reducing the effects of background noise on the targeted speech signal. This improvement is usually accomplished by preserving the characteristics of speech using short-term spectral amplitude (STSA) analysis, for which statistical speech enhancement techniques including Wiener filters and minimum mean square error (MMSE) estimation have been widely used [8,9]. These techniques are strongly dependent on the a priori signal-to-noise ratio (SNR) obtained by noise power spectral densities (PSDs), which can be reliably estimated in noise-only intervals [10,11]. Eventually, in the coupled systems of speech enhancement, a priori SNRs, PSDs, and noise-only intervals, the speech-enhancing performance is strongly dependent on accurate noise-only interval estimation. Thus, the auditory hearing aid must

use reliable VAD algorithms to obtain accurate noise-only interval information, even when target speech is corrupted by nonstationary noise such as babble noise [10,11].

VAD algorithms commonly use features such as energy levels, zero-crossing rates, entropy, and likelihood ratios (LRs) [3,12,13]. LRs produce few detection errors when the estimated global (the averaged value over all frequency bins) log LR in a frame is compared to a decision threshold [12,13]—referred to as a likelihood-ratio test (LRT) [11–13]. The decision threshold can be determined by minimizing the detection error rate and combining the false rejection probability (FRP) and the false acceptance probability (FAP). Several strategies have been suggested to improve the accuracy of the statistical-model-based LRT [13–15]; however, although false rejections do frequently arise in VAD with optimized decision thresholds, they have not been investigated systematically. We argue that the false rejection rate is linked to the sparseness characteristics of the noisy speech signal because the speech activity decision is based on the global LR magnitude relative to the decision threshold. However, empty frequency bins (independent of whether speech is active) reduce the global LR, which in turn causes false rejections. We demonstrate that the existing statistical LRT-based VAD can be improved by including false rejection measures based on the sparseness characteristics of the noisy speech. In other words, this paper proposes a method for reducing the FRP of a conventional statistical LRT-based VAD based on the sparseness characteristics of speech and noise, which is performed using LR order statistics over all frequency bins.

In the last decade, as part of efforts to identify more efficient solutions via speech-processing algorithms such as speech enhancement and VAD, the research focus of the speech algorithm community has turned to deep neural network (DNN) methods, with less attention to the previously described approaches [16–19]. The DNN-based VAD method has significantly improved the performance of such conventional approaches, even though it requires a more complex neural network architecture for greater accuracy [16–19]. Despite its strengths, however, it suffers from significantly high computational costs. Thus, it is currently difficult to use DNN-based approaches for speech-processing algorithms (e.g., VAD) in auditory devices (e.g., hearing aids) that require low computational complexity in real-world applications.

The purpose of this study is to develop a VAD algorithm to be implemented in auditory devices in which filter bank algorithms must satisfy specific requirements for signal quality, computational complexity, and signal delay [20–26]. Auditory devices should have uniformly spaced narrow frequency bands and at least 60 dB stopband attenuation, preferably higher [20–22]. Moreover, the signal-processing delay should be below 10 ms to avoid the unfavorable effect to the subjective listening experience [23,24], which requires low computational cost due to the restricted battery capacity in real-world portable devices [22,25,26]. Accordingly, a discrete Fourier transform (DFT)-based uniform polyphase auditory filter bank is typically used. It is efficient, expandable into nonuniform filter banks, has low latency, and has perfect reconstruction characteristics [20,21,26]. Moreover, it has the advantage of being implemented as an overlap-and-add (OLA) with a short-term Fourier transform (STFT) [11,22]. Consequently, this study intends to employ STFT-based VAD algorithms on a uniform polyphase DFT filter bank for auditory hearing aids.

The remainder of this paper is organized as follows. Section 2 proposes a uniform polyphase DFT filter bank for implementing the proposed auditory device VAD. Section 3 briefly reviews a conventional statistical LRT-based VAD approach. Section 4 proposes how to use LR order statistics for VAD. Section 5 evaluates the performance of the proposed method. Finally, Section 6 concludes this paper.

2. Auditory Device VAD Implementation

As described in Section 1, the auditory filter bank should have uniformly spaced narrow frequency bands and at least 60 dB stopband attenuation, preferably higher [20,21]. Furthermore, the low computational cost and low time latency of less than 10 ms are demanded for the filter bank. These restrictions are satisfied with a uniform polyphase DFT filter bank based on the fast Fourier

transform (FFT). In this paper, a 32-channel filter bank is employed with an 8 ms time delay under a 16 kHz sampling rate condition [20,21,27] on which the LRT VAD is implemented, as depicted in Figure 1.

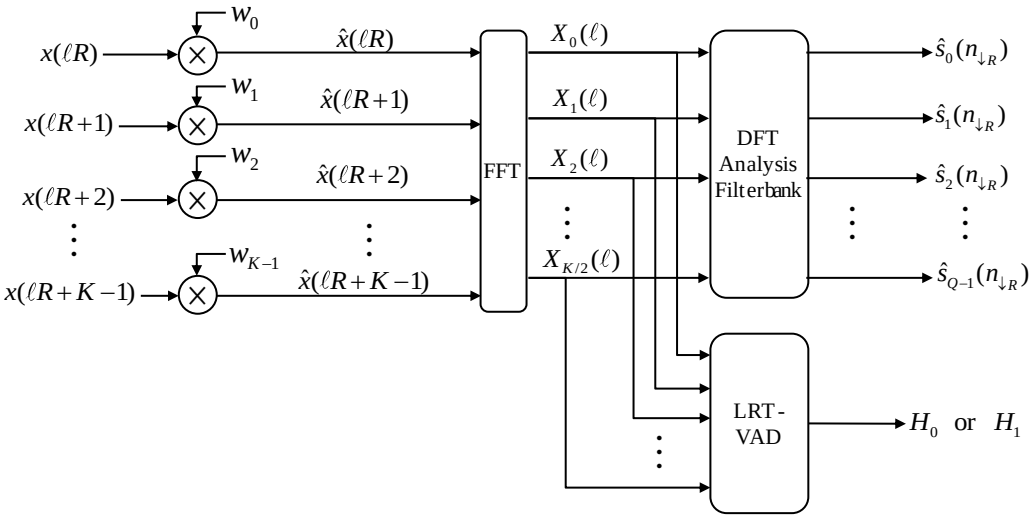

Figure 1. Block diagram of the voice activity detection (VAD) implemented on the auditory device filter bank.

We set the number of channels $Q = 32$, the frame shift length via the STFT approach R = 16, and FFT size $K = 128$ for the oversampled perfect reconstruction [22,27,28]. The input time-discrete signal $x(n)$ is buffered to form the input frame signal $\mathbf{x}$, of which length is equal to the FFT size K. In other words, the ℓth frame signal $\mathbf{x}(\ell)$ is formed as $\mathbf{x}(\ell) = [x(\ell R),\ x(\ell R+1), x(\ell R+2), \ldots, x(\ell R+K-1)]^T$, where T is the transpose operator. Then, the prototype low-pass filter (LPF) $\mathbf{w} = [w(0),\ w(1), w(2), \ldots, w(K-1)]^T$ is applied to $\mathbf{x}(\ell)$ as a window, and the windowed frame signal $\hat{\mathbf{x}}_\ell = [\hat{x}(\ell R),\ \hat{x}(\ell R+1), \hat{x}(\ell R+2), \ldots,\ \hat{x}(\ell R+K-1)]^T$ is transformed into the complex-valued spectral component $X_k(\ell)$ at the kth frequency bin $(k = 0, 1, \ldots,\ K-1)$ by an FFT. Here, 128 window sequences in [27] are used for the prototype LPF $\mathbf{w}$ of the DFT filter bank.

Whether the target speech has been activated (H_1) or not (H_0) is determined by applying a VAD algorithm to the $X_k(\ell)$ at the kth frequency bin $(k = 0, 1, \ldots, K/2)$. Simultaneously, the 16 down-sampled speech signals at the qth frequency band, $\hat{s}_q\left(n_{\downarrow 16}\right)$, can be obtained based on the real value from the complex component $X_{2q}(\ell)$, which can be used to estimate the envelope power of each band.

3. Conventional Statistical LRT-Based VAD

Assuming that speech and noise signals are additive, the detection of voice activity at the ℓ th segmented frame is accomplished by deciding upon one of two hypotheses of H_0 and H_1:

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

where $\mathbf{X}(\ell)$, $\mathbf{S}(\ell)$, and $\mathbf{N}(\ell)$ are $K/2+1$. dimensional vectors composed of k spectral components $(k = 0, 1, \ldots,\ K/2)$ of the input signal, speech, and noise, i.e., $X_k(\ell), S_k(\ell)$, and $N_k(\ell)$, respectively, such that

$$\mathbf{X}(\ell) = [\ X_0(\ell),\ \ X_1(\ell), \ldots, X_{K/2}(\ell)]^T \tag{2}$$

$$\mathbf{S}(\ell) = [\ S_0(\ell),\ \ S_1(\ell),\ \ldots, S_{K/2}(\ell)]^T \tag{3}$$

$$\mathbf{N}(\ell) = [\ N_0(\ell),\ \ N_1(\ell), \ldots, N_{K/2}(\ell)]^T. \tag{4}$$

Assuming that $S_k(\ell)$ and $N_k(\ell)$ follow complex Gaussian distributions, the probability density functions conditioned on H_0 and H_1 are given by

$$p(\mathrm{X}_k(\ell)|H_1) = \prod_{k=0}^{K-1} \frac{1}{\pi\left[\hat{\lambda}_{N,k}(\ell) + \hat{\lambda}_{S,k}(\ell)\right]} \exp\left(-\frac{|X_k(\ell)|^2}{\hat{\lambda}_{N,k}(\ell) + \hat{\lambda}_{S,k}(\ell)}\right) \tag{5}$$

and

$$p(X_k(\ell)|H_0) = \prod_{k=0}^{K-1} \frac{1}{\pi\left[\hat{\lambda}_{N,k}(\ell)\right]} \exp\left(-\frac{|X_k(\ell)|^2}{\hat{\lambda}_{N,k}(\ell)}\right) \tag{6}$$

where $\hat{\lambda}_{N,k}(\ell)$ and $\hat{\lambda}_{S,k}(\ell)$ are estimate values of the noise variance $\lambda_{N,k}(\ell)$ and speech variance $\lambda_{S,k}(\ell)$, respectively, at the kth frequency bin. Here, the hat symbol ˆ denotes an estimate value. The k th local LR $\Lambda_k(\ell)$. under $X_k(\ell)$ can be estimated [2,4] as the ratio between $p(\mathrm{X}_k(\ell)|H_1)$ and $p(\mathrm{X}_k(\ell)|H_0)$ as

$$\begin{aligned} \Lambda_k(\ell) &= \frac{p(\mathrm{X}_k(\ell)|H_1)}{p(\mathrm{X}_k(\ell)|H_0)} \\ &= \frac{1}{1+\hat{\xi}_k(\ell)} \exp\left(\frac{\hat{\gamma}_k(\ell)\cdot\hat{\xi}_k(\ell)}{1+\hat{\xi}_k(\ell)}\right) \end{aligned} \tag{7}$$

where $\hat{\xi}_k(\ell)$ is the a priori SNR estimate, which is estimated using a decision-directed (DD) approach in [2,3,10] as

$$\hat{\xi}_k(\ell) = \alpha\cdot\hat{\xi}_k(\ell-1) + (1-\alpha)\cdot\max(\hat{\gamma}_k(\ell) - 1,\ 0) \tag{8}$$

where $0 \le \alpha < 1$ is the smoothing parameter. In Equations (7) and (8), $\hat{\gamma}_k(\ell)$ is called the a posteriori SNR, which is expressed as

$$\hat{\gamma}_k(\ell) = \frac{|X_k(\ell)|^2}{\hat{\lambda}_{N,k}(\ell)} \tag{9}$$

where the noise variance estimate $\hat{\lambda}_{N,k}(\ell)$ is obtained via a recursive procedure with a smoothing parameter β, such that

$$\hat{\lambda}_{N,k}(\ell) = \beta\cdot\hat{\lambda}_{N,k}(\ell-1) + (1-\beta)\cdot|X_k(\ell)|^2 \text{if} H_0 \text{ is true.} \tag{10}$$

Then, from the averaged log value of LR $\Lambda_k(\ell)$ in Equation (7), a decision rule can be established using a decision threshold η:

$$\log\Lambda(\ell) = \frac{1}{\widetilde{K}}\sum_{k=0}^{\widetilde{K}-1}\log\Lambda_k(\ell) \begin{array}{ll} \ge & \eta \ : H_1 \\ < & \eta \ : H_0 \end{array} \tag{11}$$

where $\widetilde{K} = {}^{K}/_{2} + 1$, and $\Lambda(\ell)$ is referred to as the global LR in this paper.

4. Proposed VAD Based on LR Order Statistics

4.1. Signal Sparseness Model for LRT

The decision threshold value η in Equation (11) can be decomposed into an a priori fixed value η_0 and an increment $\Delta\eta$, i.e., $\eta = \eta_0 + \Delta\eta$, in which a larger $\Delta\eta$ subsequently leads to a larger FRP. Thus, the $\Delta\eta$ term plays a crucial role in the false reject robustness of the binary decision of speech activity in Equation (11). By incorporating $\eta = \eta_0 + \Delta\eta$ into Equation (11), we can alternatively represent this decision rule as

$$\{\log\Lambda(\ell)\}' \begin{array}{ll} \ge & \eta_0 \ : H_1 \\ < & \eta_0 \ : H_0 \end{array} \tag{12}$$

where $\{\log \Lambda(\ell)\}'$ is the attenuated version of $\log \Lambda(\ell)$ by $\Delta\eta$, i.e., $\{\log \Lambda(\ell)\}' = \log \Lambda(\ell) - \Delta\eta$. If $\{\log \Lambda(\ell)\}'$ is reduced by increasing $\Delta\eta$, the FRP increases. We argue that because of the speech and noise sparseness, $\Delta\eta$ exists in all noisy speech samples.

For the two hypotheses in Equation (1), it has been assumed in Equations (2)–(4) that a speech signal S_k is present in every frequency bin for H_1 and a noise signal N_k is present in every frequency bin for H_0 and H_1. However, speech and most types of noise (apart from white noise) do not have their energy equally distributed over all frequency bins [28,29]. Thus, to reflect the sparseness states of speech and noise, we decompose H_0 and H_1 into four states according to the presence or absence of speech and noise in the k th frequency bin, as shown in Table 1: $H_0^{\langle 1\rangle}$ or $H_1^{\langle 1\rangle}$, $H_0^{\langle 2\rangle}$ or $H_1^{\langle 2\rangle}$, $H_0^{\langle 3\rangle}$ or $H_1^{\langle 3\rangle}$, and $H_0^{\langle 4\rangle}$ or $H_1^{\langle 4\rangle}$.

Table 1. Four sparseness states of H_0 and H_1.

Speech	**Noise**	
	Present	**Absent**
Present	$H_0^{\langle 1\rangle}: N_{k^{\langle 1\rangle}}$ $H_1^{\langle 1\rangle}: N_{k^{\langle 1\rangle}} + S_{k^{\langle 1\rangle}}$	$H_0^{\langle 2\rangle}: \varepsilon_{k^{\langle 2\rangle}}$ $H_1^{\langle 2\rangle}: \varepsilon_{k^{\langle 2\rangle}} + S_{k^{\langle 2\rangle}}$
Absent	$H_0^{\langle 3\rangle}: N_{k^{\langle 3\rangle}}$ $H_1^{\langle 3\rangle}: N_{k^{\langle 3\rangle}}$	$H_0^{\langle 4\rangle}: \varepsilon_{k^{\langle 4\rangle}}$ $H_1^{\langle 4\rangle}: \varepsilon_{k^{\langle 4\rangle}}$

In the table, the minimum value of the noise components is specified as ε_k. The superscript $\langle 1\rangle$ is the state in which both the speech and noise components exist at the k th frequency bin; $\langle 2\rangle$ and $\langle 3\rangle$ are speech-only and noise-only states, respectively, and $\langle 4\rangle$ represents the states with neither speech nor noise. Based on this notation of the four sparseness states, the LRT in Equation (11) can be expressed as

$$\frac{\sum_{k^{\langle 1\rangle}} \log \Lambda_k(\ell) + \sum_{k^{\langle 2\rangle}} \log \Lambda_k(\ell) + \sum_{k^{\langle 3\rangle}} \log \Lambda_k(\ell) + \sum_{k^{\langle 4\rangle}} \log \Lambda_k(\ell)}{\widetilde{K}} \begin{matrix} H_1 \\ \geq \\ \\ < \\ H_0 \end{matrix} \eta_0 \tag{13}$$

where $\widetilde{K} = num\left(k^{\langle 1\rangle}\right) + num\left(k^{\langle 2\rangle}\right) + num\left(k^{\langle 3\rangle}\right) + num\left(k^{\langle 4\rangle}\right)$ and $num(\cdot)$ is the number of respective frequency bins. In Equation (13), $\sum\limits_{k^{\langle 3\rangle}} \log \Lambda_k(\ell)$ and $\sum\limits_{k^{\langle 4\rangle}} \log \Lambda_k(\ell)$ reduce the global LR, thus increasing the FRP. Therefore, $\sum\limits_{k^{\langle 3\rangle}} \log \Lambda_k(\ell)$ and $\sum\limits_{k^{\langle 4\rangle}} \log \Lambda_k(\ell)$ do not robustly estimate the global LR and are less suitable to calculate the LRT:

$$\frac{1}{\widetilde{K'}}\left(\sum_{k^{\langle 1\rangle}} \log \Lambda_k(\ell) + \sum_{k^{\langle 2\rangle}} \log \Lambda_k(\ell)\right) \begin{matrix} H_1 \\ \geq \\ \\ < \\ H_0 \end{matrix} \eta_0 \tag{14}$$

where $\widetilde{K'} = num\left(k^{\langle 1\rangle}\right) + num\left(k^{\langle 2\rangle}\right)$. Consequently, the solution to improve the robustness against false rejection is based on estimating $\sum\limits_{k^{\langle 1\rangle}} \log \Lambda_k(\ell) + \sum\limits_{k^{\langle 2\rangle}} \log \Lambda_k(\ell)$ from four sparseness states. Accordingly, we suggest an LR order statistics approach motivated by

$$\begin{aligned} &\sum_{k^{\langle 2\rangle}} \log \Lambda_k(\ell) \geq \sum_{k^{\langle 1\rangle}} \log \Lambda_k(\ell) \\ &\geq \sum_{k^{\langle 3\rangle}} \log \Lambda_k(\ell) \text{or} \sum_{k^{\langle 4\rangle}} \log \Lambda_k(\ell) \end{aligned} \tag{15}$$

In the next subsection, we will explore the use of LR order statistics for a false rejection of robust VAD.

4.2. VAD Based on LR Order Statistics

First, the log LR sets $\log \Lambda_0(\ell), \log \Lambda_1(\ell), \ldots, \log \Lambda_{K-1}(\ell)$ are arranged in descending order of magnitude as

$$\log \Lambda_{\{0\}}(\ell) \geq \log \Lambda_{\{1\}}(\ell) \geq \cdots \geq \log \Lambda_{\{K-1\}}(\ell) \tag{16}$$

where the subscript $\{k\}$ is the new index of the log LR after ordering. We then denote the elements of the new log LR set as $\Psi_k(\ell)$, such that $\Psi_k(\ell) \equiv \log \Lambda_{\{k\}}(\ell)$. Thus, the left side in Equation (14) can be expressed as $\sum_{k\langle 1\rangle} \log \Lambda_k(\ell) + \sum_{k\langle 2\rangle} \log \Lambda_k(\ell) = \sum_{k=0}^{\widetilde{K}'-1} \Psi_k(\ell)$. Finally, using the LR order statistics, the LRT rule becomes

$$\frac{1}{\widetilde{K}'} \sum_{k=0}^{\widetilde{K}'-1} \Psi_k(\ell) \begin{array}{ccc} \geq & \eta_0 & : H_1 \\ < & \eta_0 & : H_0 \end{array} \tag{17}$$

From Equation (17), the problem of separately estimating $\sum_{k\langle 1\rangle} \log \Lambda_k(\ell) + \sum_{k\langle 2\rangle} \log \Lambda_k(\ell)$ from the observed noisy speech is focused on the estimation of $\widetilde{K}'$, which is a tuning parameter used to control the FRP robustness. When $\widetilde{K}' = \widetilde{K}$, the proposed VAD in Equation (14) equals the conventional VAD in Equation (11).

5. Experiments and Results

We evaluated the proposed algorithm by counting detection errors and comparing it with conventional VAD under various noise types and SNR conditions. Speech utterances of approximately 57 s in duration were obtained from four speakers (two males and two females) from the TIMIT database (DB) [28] and mixed with three noises (white, babble, and Volvo noise) from the NOISEX-92 DB [29]. Based on the clean signals, 65.7% of the samples in the speech material were marked as active (49.3% voiced and 16.4% unvoiced). The noise signals were then artificially mixed additively with SNRs ranging from 0 to 20 dB with 5 dB steps. Signals were segmented using the 128-point LPF window in [27] and overlapped with each previous segment by one-eighth. We implemented the statistical LRT-based VAD method proposed in [12] with a conventional hang-over scheme to investigate how much the proposed LR order statistics approach reduced the detection error rate. Moreover, we set the $\alpha = 0.97$ in Equation (8) and $\beta = 0.98$ in Equation (10) for the experiments, which were empirically determined.

Figure 2a illustrates an example waveform of male speech at 5 dB SNR in babble noise, and Figure 2b illustrates the conventional $\log \Lambda_k$ and the proposed descending version of $\log \Lambda_k$, Ψ_k, at close to 0.6 s (a voiced interval) of this waveform. The proposed Ψ_k illustrated a concentration at low frequencies and is less distributed than $\log \Lambda_k$. Figure 2c illustrates the global log LR curves obtained from the local LRs of $\log \Lambda_k$ and Ψ_k. To quantize the tuning parameter $\widetilde{K}'$ in Equation (17), the value was converted to a percentage: $M = 100 \times \widetilde{K}'/\widetilde{K}(\%)$. The proposed method (at $M = 25\%$) produced higher log LRs than the conventional method. At $M = 100\%$ (or $\widetilde{K}' = \widetilde{K}$), the proposed method was identical to the conventional VAD in Equation (11).

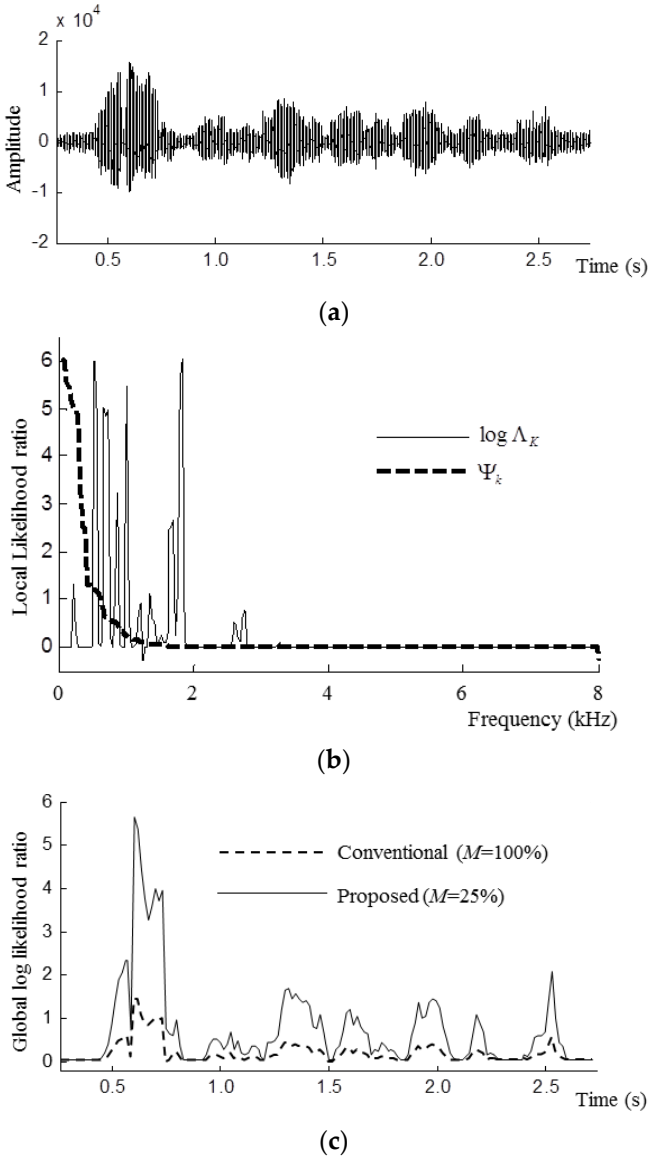

Figure 2. Step-by-step illustration of (**a**) waveform of babble noisy speech at 5 dB signal-to-noise ratio (SNR), (**b**) local likelihood ratio (LR) estimates of the frequency, and (**c**) global log LR estimates of the time.

For investigating the effect of the control constant K' in Equation (17) in detail, the FRP was measured by comparing the detection results of the proposed method to true voice activity intervals (Figure 3). The ground truth voice activity intervals were determined manually by the author. The figure illustrates FRP histograms for white noise (left) and babble (right) noise environments at 5 dB SNR for three different values of M. The FRPs decreased considerably with decreasing M for both white noise and babble noise. The optimal value of $M = 25\%$ was adopted for the subsequent experiments.

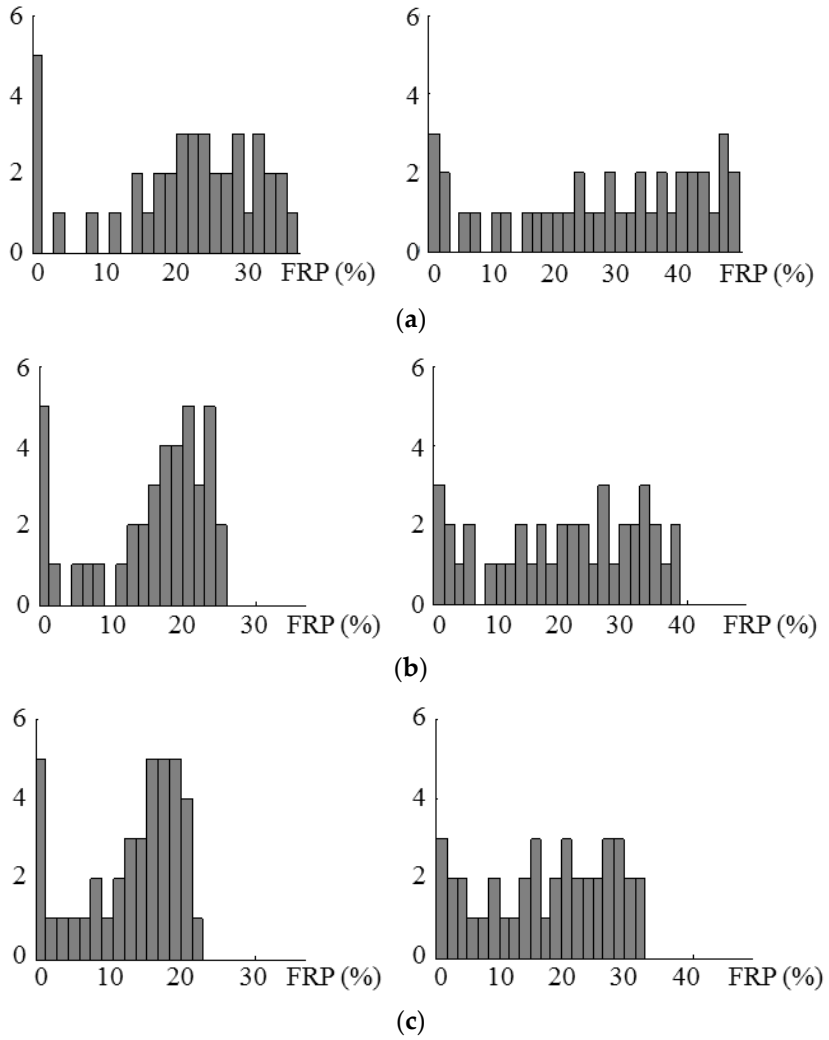

Figure 3. False rejection probability (FRP) histograms for speech in white noise (**left**) and babble noise (**right**) at 5 dB SNR at different M: (**a**) $M = 100\%$, (**b**) $M = 50\%$, and (**c**) $M = 25\%$.

We measured the FAP to investigate further the effectiveness of the proposed LR order statistics in an LRT-based VAD method. Figure 4 depicts the results of the comparison between the proposed and conventional methods. The results are presented in the form of a detection error tradeoff graph, similar to receiver operating characteristic (ROC) curves. Depicted are the results of the conventional method with $M = 100\%$ and the proposed method with $M = 25\%$ for speech in white noise and babble noise at 5 dB SNR. The detection error tradeoff curve of the proposed method was always closer to the bottom-left corner than that of the conventional VAD for both noise conditions, demonstrating that the proposed method is more robust against false detection than the conventional method.

Finally, we compared the relative detection error reduction rate (RDER) of the proposed method ($M = 25\%$) to that without under variable noise and SNR conditions. The detection error rate was defined as (FAP + FRP)/2. The decision threshold was explicitly determined to minimize the detection error under each noise and SNR condition. Table 2 illustrates that the proposed method increased the RDER in all noise environments at all SNRs. In the table, the proposed method relatively reduced the average detection error rate by 15.8% compared to a conventional VAD, with only minimal change in the false acceptance probability for three different noise conditions whose signal-to-noise ratio ranged from 0 to 20 dB. This finding implies that an LRT-based VAD employing the proposed LR order statistics can be an effective solution for attenuating the detection error by improving the false reject robustness in noise environments.

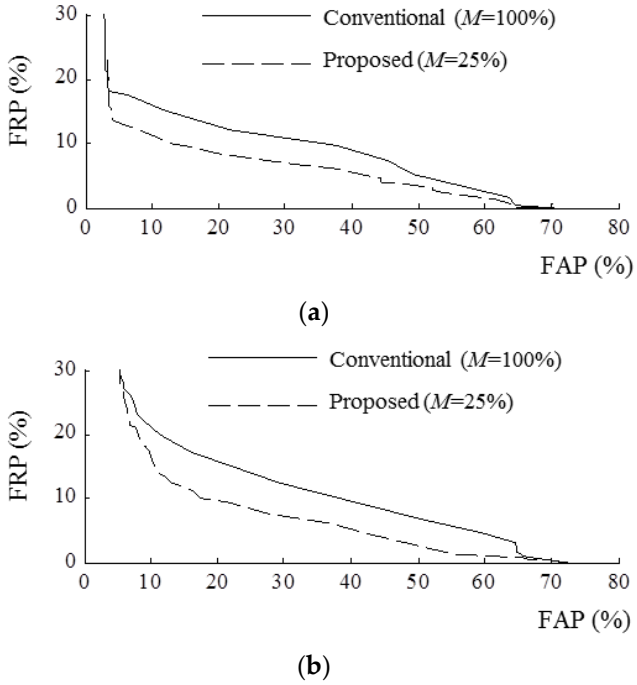

Figure 4. Detection error tradeoff curves for two noise types under the 5 dB SNR condition: (**a**) white noise and (**b**) babble noise.

Table 2. FRP, false acceptance probability (FAP), and relative detection error reduction rate (RDER) (%) of a likelihood-ratio-test (LRT)-based VAD with and without proposed LR order statistics under various noise types and SNR conditions.

Noise Types	SNR (dB)	Conventional ($M = 100\%$)		Proposed ($M = 25\%$)		RDER (%)
		FRP (%)	FAP (%)	FRP (%)	FAP (%)	
White	20	15.81	11.31	10.17	12.00	18.25
	15	16.74	3.38	11.00	4.07	25.10
	10	16.92	6.28	12.25	4.55	27.59
	5	18.49	3.59	13.08	6.48	11.41
	0	19.79	8.76	15.90	12.00	2.28
Babble	20	15.35	11.93	10.36	13.03	14.26
	15	15.40	4.76	7.49	6.62	30.01
	10	14.89	11.10	8.41	9.66	30.47
	5	21.22	9.86	14.15	11.17	18.53
	0	24.60	15.45	20.99	14.48	11.44
Volvo	20	14.93	11.52	8.88	13.38	15.84
	15	14.75	4.07	8.60	7.03	16.95
	10	14.70	4.55	10.03	5.66	18.49
	5	14.89	4.83	9.85	6.14	18.91
	0	14.98	5.17	10.96	5.45	18.56

6. Conclusions

We presented a solution for reducing the false rejection error of an LRT-based VAD in terms of auditory device speech-processing performance. The proposed LR order statistics consider that false rejections are linked to the sparseness of speech and additive noise signals. Accordingly, a spectral LRT-based VAD employing the proposed method was developed for a uniform polyphase DFT filter bank to satisfy auditory device hearing aid requirements regarding low computational cost and algorithm processing delay. Our experimental results confirmed that the LRT-based VAD having the

proposed LR order statistics relatively reduced the average detection error rate by 15.8% compared to a conventional VAD, with only minimal change in the false acceptance probability under all tested noise conditions in the tested SNR range between 0 and 20 dB.

Author Contributions: S.M.K. contributed to the research idea and the framework of this study. S.M.K. also contributed ideas on how to formulate LR order statistics for VAD and performed the experiments for the activity detection of voice in noisy environments. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2019R1F1A1063325).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Benyassine, A.; Shlomot, E.; Su, H.Y.; Massaloux, D.; Lamblin, C.; Petit, J.P. ITU-T Recommendation G729 Annex B: A silence compression scheme for use with G729 optimized for V70 digital simultaneous voice and data applications. *IEEE Commun. Mag.* **1997**, *35*, 64–73. [CrossRef]
2. Benesty, J.; Makino, S.; Chen, J. *Speech Enhancement*; Springer: New York, NY, USA, 2005.
3. ETSI Std. *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front–End Feature Extraction Algorithm; Compression Algorithms*; ETSI ES 202 050 V1.1.1 (2002–10); European Telecommunications Standards Institute: Valbonne, France, 2002.
4. Grimm, G.; Herzke, T.; Berg, D.; Hohmann, V. The master hearing aid: A PC based platform for algorithm development and evaluation. *Acta Acust. United Acust.* **2006**, *92*, 618–628.
5. Kochkin, S. MarkeTrak VII: Why my hearing aids are in the drawer: The consumers' perspective. *Hear. J.* **2000**, *53*, 34–41. [CrossRef]
6. Kochkin, S. MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids. *Hear. J.* **2007**, *60*, 24–51. [CrossRef]
7. Plomp, R. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J. Acoust. Soc. Am.* **1978**, *63*, 533–549. [CrossRef] [PubMed]
8. Healy, E.W.; Yoho, S.E.; Wang, Y.; Wang, D. An algorithm to improve speech recognition in noise for hearing-impaired listeners. *J. Acoust. Soc. Am.* **2013**, *134*, 3029–3038. [CrossRef] [PubMed]
9. Trawicki, M.B.; Johnson, M.T. Distributed multichannel speech enhancement with minimum mean-square error short-time spectral amplitude, log-spectral amplitude, and spectral phase estimation. *Signal Process.* **2012**, *92*, 345–356. [CrossRef]
10. Loizou, P.C. *Speech Enhancement: Theory and Practice*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2013.
11. Lee, S.J.; Kang, B.O.; Jung, H.; Lee, Y.; Kim, H.S. Statistical model-based noise reduction approach for car interior applications to speech recognition. *ETRI J.* **2010**, *32*, 801–809. [CrossRef]
12. Sohn, J.; Kim, N.S.; Sung, W. A statistical model–based voice activity detection. *IEEE Signal Process. Lett.* **1999**, *6*, 1–3. [CrossRef]
13. Cho, Y.D.; Kondoz, A. Analysis and improvement of a statistical model–based voice activity detector. *IEEE Signal Process. Lett.* **2001**, *8*, 276–278.
14. Ramirez, J.; Puntonet, C.G.; Segura, J.C. Generalized LRT–based voice activity detector. *IEEE Signal Process. Lett.* **2006**, *13*, 636–639.
15. Shin, J.W.; Kwon, H.J.; Jin, S.H.; Kim, N.S. Voice activity detection based on conditional MAP criterion. *IEEE Signal Process. Lett.* **2008**, *15*, 257–260. [CrossRef]
16. Lee, G.W.; Kim, H.K. Multi-Task Learning U-Net for Single-Channel Speech Enhancement and Mask-Based Voice Activity Detection. *Appl. Sci.* **2020**, *10*, 3230. [CrossRef]
17. Zazo, R.; Sainath, T.N.; Simko, G.; Parada, C. Feature learning with raw-waveform CLDNNs for voice activity detection. In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), San Francisco, CA, USA, 8–12 September 2016; pp. 8–12.
18. Kim, J.; Kim, J.; Lee, S.; Park, J.; Hahn, M. Vowel based voice activity detection with LSTM recurrent neural network. In Proceedings of the International Conference on Signal Processing Systems, Auckland, New Zealand, 21–24 November 2016; pp. 134–137.

19. Zhang, X.-L.; Wang, D. Boosting contextual information for deep neural network based voice activity detection. *IEEE/Acm Trans. Audio Speech Lang. Process.* **2016**, *24*, 252–264. [CrossRef]
20. Buchholz, J.M. A real-time hearing-aid research platform (HARP): Realization, calibration, and evaluation. *Acust. United Acust.* **2013**, *99*, 477–492. [CrossRef]
21. Kim, S.M.; Bleeck, S. An open development platform for auditory real-time signal processing. *Speech Commun.* **2018**, *98*, 73–84. [CrossRef]
22. Bäuml, R.W.; Sörgel, W. Uniform polyphase filter banks for use in hearing aids: Design and constraint. In Proceedings of the 16th European Signal Processing Conference, Lausanne, Switzerland, 25–29 August 2008; pp. 25–29.
23. Löllmann, H.; Vary, P. Low delay noise reduction and dereverberation for hearing aids. *EURASIP J. Appl. Signal Process.* **2009**, *1*, 1–9.
24. Stone, M.A.; Moore, B.C. Tolerable hearing aid delays. III. Effects on speech production and perception of across-frequency variation in delay. *Ear Hear.* **2003**, *24*, 175–183. [CrossRef]
25. Löllmann, H.W.; Vary, P. Uniform and warped low delay filter-banks for speech enhancement. *Speech Commun.* **2007**, *49*, 574–587. [CrossRef]
26. Löllmann, H.W.; Vary, P. Low delay filter-banks for speech and audio processing. In *Speech and Audio Processing in Adverse Environments*; Springer: Berlin, Germany, 2008; pp. 13–61.
27. Kim, S.M. Hearing Aid Speech Enhancement Using Phase Difference-Controlled Dual-Microphone Generalized Sidelobe Canceller. *IEEE Access* **2019**, *7*, 2169–3536. [CrossRef]
28. Garofolo, J.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L.; Zue, V. *TIMIT Acoustic–Phonetic Continuous Speech Corpus*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
29. Varga, A.; Steeneken, H.J.M. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]