Voice Activity Detection Based on Statistical Likelihood Ratio With Adaptive Thresholding

Xiaofei Li, Radu Horaud, Laurent Girin, Sharon Gannot

VOICE ACTIVITY DETECTION BASED ON STATISTICAL LIKELIHOOD RATIO WITH ADAPTIVE THRESHOLDING

Xiaofei Li[1], *Radu Horaud*[1], *Laurent Girin*[1,2]

[1]INRIA Grenoble Rhône-Alpes
[2]GIPSA-Lab & Univ. Grenoble Alpes

Sharon Gannot

Faculty of Engineering
Bar-Ilan University

ABSTRACT

Statistical likelihood ratio test is a widely used voice activity detection (VAD) method, in which the likelihood ratio of the current temporal frame is compared with a threshold. A fixed threshold is always used, but this is not suitable for various types of noise. In this paper, an adaptive threshold is proposed as a function of the local statistics of the likelihood ratio. This threshold represents the upper bound of the likelihood ratio for the non-speech frames, whereas it remains generally lower than the likelihood ratio for the speech frames. As a result, a high non-speech hit rate can be achieved, while maintaining speech hit rate as large as possible.

Index Terms— voice activity detection, likelihood ratio test, adaptive threshold, high non-speech hit rate.

1. INTRODUCTION

Voice activity detection (VAD) classifies a noisy speech stream into speech and non-speech (i.e. noise only) segments. VAD is an essential prerequisite for many speech communication systems, such as speech recognition, noise reduction, sound source localization, etc. Briefly speaking, there are mainly three categories of VAD methods: 1) Speech content-based method. Mel-frequency cepstral coefficients feature and various classifiers, such as support vector machines [1], spectral clustering [2] and Gaussian mixture model (GMM) [3], are often used for speech and non-speech classification. Recently, the non-negative matrix factorization was also investigated for VAD [4]. All these methods need supervised learning using a fair amount of speech and noise materials. 2) Long-term statistics, such as signal variability [5] and spectral flatness [6], are defined to characterize the difference between speech and noise. These methods are unsupervised, except possibly for the setting of thresholds. 3) Energy-based methods assume that non-speech portions have significantly lower energy than speech portions. The long-term spectral divergence [7] computes the spectral envelope power to noise power ratio. The sequential GMM method in [8] recursively learns a GMM comprised of two components, which correspond to the log-energy of noise and speech, respectively. These two energy-based methods simultaneously estimate the noise power spectral density (PSD) and detect the voice activity by comparing with a threshold. The statistical likelihood ratio test [9] first estimates the noise PSD and the *a priori* SNR, and then computes the likelihood ratio. The likelihood ratio is compared with a threshold for VAD. In [10], the likelihood ratio is smoothed across temporal frames, which makes this ratio higher in the speech offset regions and lower (and flatter) in the non-speech regions. Energy-based methods are also unsupervised, and the accuracy of noise PSD estimate is here a critical factor. In contrast to the other two energy-based methods, the likelihood ratio test can achieve better performance by adopting an existing advanced noise PSD estimator, such as [11, 12, 13]. In general, VAD performance is a trade-off between speech hit rate (SHR) and non-speech hit rate (NHR), that are the percentage of correctly detected speech and non-speech frames, respectively. The likelihood ratio test method always utilizes a fixed threshold, which is not suitable for different types of noise. Since, in practice, the amplitude of the likelihood ratios of different types of noise are different due to the different estimation errors of noise PSD. Generally, the more nonstationary the noise signal, the larger the estimation error of noise PSD.

In this paper, we propose an adaptive threshold that is adaptable to various types of noise. The logarithmic smoothed likelihood ratio (log-SLR) of the noise signal is assumed to be normally distributed, however the mean (and variance) of this distribution are different for different types of noise and maybe time-varying. We propose a novel method to learn the mean and variance of the log-SLR of the noise (non-speech) portions. Based on those statistics, an adaptive threshold is set to achieve a high NHR, while maintaining SHR as large as possible. A high NHR is very useful for some applications, for example in video-conferencing systems, the false alarm rate (wrongly detected non-speech frames) should be low, and a relatively lower SHR is acceptable. Experiments show the efficiency of our adaptive threshold, compared with the sequential GMM method and the likelihood ratio test with a fixed threshold.

This research has received funding from the EU-FP7 STREP project EARS (#609465).

978-1-5090-2007-2/16/$31.00

2. SMOOTHED LIKELIHOOD RATIO

Let us consider an uncorrelated additive speech + noise mixture signal, in the STFT domain. Let us denote $\mathbf{X}_l = [X_{l,1}, \ldots, X_{l,K}]^\top$ the vector of STFT coefficients of the noisy signal at time frame l, and the same for $\mathbf{N}_l$ and $\mathbf{S}_l$ the noise and speech signals, respectively (k denotes the frequency bin). For each time frame, the VAD gives the decision between two hypotheses:

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

Let $\lambda_{n,lk} = E\{|N_{l,k}|^2\}$ and $\lambda_{s,lk} = E\{|S_{l,k}|^2\}$ denote the PSDs of the noise and speech signal, respectively. The probability density function of the measured power spectrogram $|X_{l,k}|^2$ follows an exponential distribution, with mean $\lambda_{n,lk}$ and $\lambda_{s,lk} + \lambda_{n,lk}$ for each hypothesis [14], respectively.

The log-likelihood ratio for TF bin (l, k) is [9]

$$\Lambda_{l,k} \triangleq \log\left\{\frac{p(|X_{l,k}|^2|H_1)}{p(|X_{l,k}|^2|H_0)}\right\} = \frac{\gamma_{l,k}\xi_{l,k}}{1+\xi_{l,k}} - \log\{1+\xi_{l,k}\}, \tag{2}$$

where $\gamma_{l,k} \triangleq |X_{l,k}|^2/\lambda_{n,lk}$ and $\xi_{l,k} \triangleq \lambda_{s,lk}/\lambda_{n,lk}$ are the *a posteriori* and *a priori* SNRs [15], respectively.

The noise PSD $\lambda_{n,lk}$ is either supposed to be known or it can be estimated by an existing noise PSD estimator. For instance, an MMSE-based estimator [11] is adopted in our experiments. The *a priori* SNRs $\xi_{l,k}$ can be estimated by the decision-directed method [15]. In [10] a smoothed likelihood ratio was introduced as: $\Psi_{l,k} = \kappa\Psi_{l-1,k} + (1-\kappa)\Lambda_{l,k}$, where κ is a smoothing factor. The voice activity decision is made using the averaged smoothed likelihood ratios across frequency bins:

$$\Psi_l = \sum_{k=1}^{K} \Psi_{l,k} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{3}$$

where η is a threshold. It is shown in [10] that the smoothed likelihood ratio provides a larger discrimination between speech and non-speech portions than the raw ratio.

3. ADAPTIVE THRESHOLD

Ideally, if the noise PSD is accurately estimated, the smoothed likelihood ratio of the noise frames should be close to zero (always positive). In general, the more non-stationary the noise signal, the larger the noise PSD estimation error, and the larger the fluctuation of the likelihood ratio. Therefore, a fixed threshold η is not suitable for VAD in various types of noise.

In this section, we propose an algorithm that modulates the threshold adaptively, resulting in a time-varying threshold η_l. First, we compute the log-scale smoothed likelihood ratio (log-SLR, in dB) to shrink the scale of the smoothed likelihood ratio, as:

$$Y_l = 10\log_{10}(\Psi_l). \tag{4}$$

We assume that the log-SLR of the noise frames approximately follows a Gaussian distribution, which mean μ_l and variance Σ_l will be estimated in this section. Then the adaptive threshold is empirically set as $\eta_l = \mu_l + 3\sqrt{\Sigma_l}$, which leads to a high NHR. Similarly to (3), the VAD based on log-SLR is then given by: $Y_l \underset{H_0}{\overset{H_1}{\gtrless}} \eta_l$. Fig. 1 shows an instance for the adaptive thresholding.

3.1. Mean and variance estimate

The mean μ_l and variance Σ_l are estimated as:

$$\mu_l = \begin{cases} \mu_{l-1} + \phi_l, & \text{if } Y_l > \mu_{l-1} \\ \alpha\mu_{l-1} + (1-\alpha)(Y_l + \sqrt{\frac{2}{\pi}\Sigma_{l-1}}) - \phi_l, & \text{otherwise.} \end{cases} \tag{5}$$

$$\Sigma_l = \begin{cases} \Sigma_{l-1}, & \text{if } Y_l > \mu_{l-1} \\ \alpha\Sigma_{l-1} + (1-\alpha)(Y_l - \mu_l)^2, & \text{otherwise.} \end{cases} \tag{6}$$

where α is a smoothing factor. This strategy is justified by the following considerations. If the current observation Y_l is larger than the previous mean μ_{l-1}, it is unsure that the current frame is a speech frame or non-speech frame. The mean is only updated by a small value ϕ_l (which is set as $0.002\sqrt{\Sigma_{l-1}}$), and there is no update for variance. On the one hand, the small value of ϕ_l makes the update of μ_l negligible with regard to the observation Y_l larger than the previous mean. On the other hand, when Y_l increases slowly during non-speech period, ϕ_l makes μ_l track this increase well. If $Y_l \leq \mu_{l-1}$, the current frame is noise-only with a very high probability, so we recursively update the estimation of μ_l and Σ_l. However, the averaging of smaller observations will lead to a biased estimate of μ_l. To compensate this bias, a compensation term $\sqrt{\frac{2}{\pi}\Sigma_{l-1}}$ is used, which represents the bias between μ_l and the expectation of Y_l (subject to $Y_l \leq \mu_{l-1}$) [1]. The term $-\phi_l$ is a compensation to $+\phi_l$ in the case $Y_l > \mu_{l-1}$.

The proportion of the frames with Y_l lower than μ_l is recursively characterized by:

$$h_l = \alpha h_{l-1} + (1-\alpha)\mathbf{I}\{Y_l < \mu_l\}, \tag{7}$$

where the indicator function $\mathbf{I}\{\cdot\}$ is 1 if its argument is true and 0 otherwise. Indeed, when the log-SLR of the noise frames goes down abruptly (for instance, a nonstationary noise disappears, see the log-SLR value after 30s in Fig. 1), during noise periods, most frames have a log-SLR lower than

[1] For a Gaussian variable x with mean μ and variance Σ, $\sqrt{\frac{2}{\pi}\Sigma}$ is the difference between μ and the mean value of the elements less than μ. i.e. $\mu - \frac{1}{F(\mu)}\int_{-\infty}^{\mu} xf(x) = \sqrt{\frac{2}{\pi}\Sigma}$, where $f(x)$ and $F(x)$ denote probability density function and cumulative distribution function, respectively.

the estimated mean μ_l (leading to a large h_l). However, in this case, the mean estimation μ_l is still at the level before the log-SLR decreases, thence is overestimated. This also causes the variance Σ_l to be overestimated, since most of observations are much smaller than μ_l. Thence, the estimated mean μ_l goes down very slowly due to the large compensation term $\sqrt{\frac{2}{\pi}\Sigma_{l-1}}$. To accelerate the decrease of μ_l, we introduce the new following principle: if the proportion h_l is larger than ρ_1, μ_l is updated without compensation as

$$\mu_l = \alpha\mu_{l-1} + (1-\alpha)Y_l, \text{ if } Y_l \leq \mu_{l-1} \text{ and } h_l > \rho_1. \quad (8)$$

The effect of this principle can be seen in Fig. 1, when the nonstationary noise suddenly disappears after 30s: μ_l quickly decreases in a correct manner.

In parallel, during speech periods, most frames have a log-SLR larger than μ_l (leading to a small h_l). Thence, a long period of speech segment will lead to an overestimation of μ_l due to the cumulation of ϕ_l. To prevent this case, we introduce another principle: if the proportion h_l is less than ρ_2, μ_l is updated without the compensation term ϕ_l as

$$\mu_l = \mu_{l-1}, \text{ if } Y_l > \mu_{l-1} \text{ and } h_l < \rho_2. \quad (9)$$

Combining these two new principles with (5), the update of μ_l is finally given by:

$$\mu_l = \begin{cases} \mu_{l-1}, & \text{if } Y_l > \mu_{l-1} \text{ and } h_l < \rho_2 \\ \mu_{l-1} + \phi_l, & \text{if } Y_l > \mu_{l-1} \text{ and } h_l \geq \rho_2 \\ \alpha\mu_{l-1} + (1-\alpha)Y_l, & \text{if } Y_l \leq \mu_{l-1} \text{ and } h_l > \rho_1 \\ \alpha\mu_{l-1} + (1-\alpha)(Y_l + \sqrt{\frac{2}{\pi}\Sigma_{l-1}}) - \phi_l, & \text{otherwise.} \end{cases} \quad (10)$$

Note that the variance update is still made using (6).

3.2. Safety net

The process described above is able to track a slow increase and an abrupt decrease of the noise log-SLR, no matter if speech is present or not. However, when the noise log-SLR rises rapidly (for example, a nonstationary noise appears, see the noise log-SLR value after 60s in Fig. 1), it is possible that there will be too few observations smaller than μ_{l-1}, and μ_l will not be updated anymore (in other words, it is locked). To prevent this case, the minimum value of the log-SLR over the past D frames, $y_l = \min\{Y_m\}_{m=l-D+1}^{l}$, and the median value $\bar{Y}_l = \text{median}\{Y_m\}_{m=l-D+1}^{l}$ are taken into account. The safety-net is set as:

$$\hat{\mu}_l = \begin{cases} \max\{\mu_l, y_l + \sqrt{\Sigma_l}\}, & \text{if } \bar{Y}_l < \delta \\ \mu_l, & \text{otherwise.} \end{cases} \quad (11)$$

where δ is an empirical value that is larger than the log-SLR median for most types of noise. In other words, if the median $\bar{Y}_l$ is larger than δ, the current frames are most likely speech

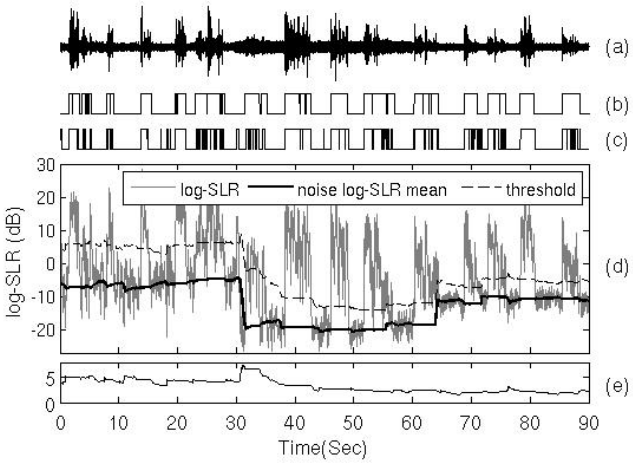

Fig. 1: An example of VAD based on the adaptive threshold. (a) Noisy speech signal: Three types of noise are added to a speech signal from the TIMIT corpus [16], with overall SNR 5dB. The noise signals are babble noise (0-30s), white noise (30-60s) and Buccaneer noise (60-90s) from NOISEX92 database [17]. (b) Ground truth voice activity labels, as provided by the word labels of TIMIT. (c) Voice activity decision based on the adaptive threshold. (d) log-SLR, estimated noise log-SLR mean and adaptive threshold. (e) Estimated noise log-SLR variance.

frames (otherwise they can be either noise frames or speech frames). Once the log-SLR of noise rises significantly, after at most D continuous noise frames, if μ_l is locked at the level before the log-SLR rise, and smaller than $y_l + \sqrt{\Sigma_l}$, it will be reset to the new level $y_l + \sqrt{\Sigma_l}$. This can be seen around the 65th second in Fig. 1.

During speech periods, if the SNR is high, the median $\bar{Y}_l$ will not be lower than δ, thence the safety net will not be activated. If the SNR is low, i.e. speech frames have a quite small log-SLR, possibly $\bar{Y}_l < \delta$, the value $y_l + \sqrt{\Sigma_l}$ is generally lower than μ_l because of the low speech power and short pauses among speech frames. During noise periods, if the log-SLR is changing slowly, obviously, the event $y_l + \sqrt{\Sigma_l} > \mu_l$ will happen with a negligible probability with regard to Gaussian distribution. Therefore, this safety net will not significantly influence the estimate of μ_l on speech periods and slowly changing noise periods.

4. EXPERIMENTS

To evaluate the proposed VAD method, various experiments are carried out.

Data set: For generating a long clean speech signal, we randomly select 300 speech sentences from 6,300 sentences of the TIMIT corpus. The silence at the beginning and at the end of each sentence is removed. Instead, a silence of an uniformly random duration from -1s to 5s is added before each sentence. Then these 300 silence-padded sentences are concatenated to be a 24min-long signal. Note that if the silence duration is less than 0s, there is actually no silence, instead an overlap between the current sentence and the pre-

vious one. The word labels from the TIMIT transcriptions are adopted to generate the labels of non-speech and speech portions with the resolution of 10ms interval. The proportion of the non-speech and speech portions are 47.5% and 52.5%, respectively. Five types of noise are tested: white, F16, Buccaneer, destroyerops and babble from the NOISEX92 database [17], where white and F16 noise are relatively stationary, and the rest are nonstationary. In addition, to simulate a practical situation where a type of noise can appear and disappear, a fusion noise is generated: five types of noise are concatenated with a random order, and two adjacent noise segments have a uniformly random overlap from 0 to 1min. Note that the duration of each segment of noise is about 2min, therefore, one type of noise could appear more than once. The noise power could significantly change when one type of noise appears or disappears. The speech signal is degraded by these noise signals with SNRs of -10:5:10 dB, respectively.

Parameter setup: The STFT is applied with a 20ms Hamming window, with 10ms-overlap, which corresponds to 320 and 160 samples with respect to the signal sampling rate of 16kHz. It means that the voice activity decision is committed every 10ms. The maximum frequency bin K for log-SLR averaging in (3) is set to 80, corresponding to 4kHz, above which the speech PSD contains lower and more diffuse energy. The smoothing factor κ is set to 0.8. The smoothing factor α is set to 0.97, which is an empirical value that provides a good update rate of μ_l and h_l. The two thresholds ρ_1 and ρ_2 are set to 0.8 and 0.02, respectively. In the safety net, the information of the past $D = 300$ (i.e. 3s) frames are used. The parameter δ is set to -2 dB, based on the log-SLR median of babble noise (which is nonstationary noise). The first frame is assumed as a noise frame, thence μ_1, Σ_1 and h_1 are initialized as Y_1, 0 and 0.5, respectively.

Comparison methods: NHR and SHR are taken as the performance metric. Two existing methods: Smoothed likelihood ratio method (SLR-Cho) [10] with a fixed threshold and the sequential GMM (SGMM) method [8] are evaluated for comparison. The proposed method replaces the fixed threshold in SLR-Cho by an adaptive one, thus the comparison of the two methods enables to quantify the effect of the adaptive threshold. The proposed method and the SGMM method are both energy-based and unsupervised VAD method, thence the comparison is fair. To obtain a good NHR for most of the test signals, the fixed threshold for SLR-Cho and the voting threshold for SGMM should be set to a sufficiently large value, in our experiments, 0.7 and 6, respectively.

Results: Fig. 2 shows the NHR and SHR for various types of noise and SNRs. Note that, since the fusion noise is generated by concatenating different types of noise with a random order, its results are obtained by averaging five runs. First, we compare SLR-Cho and the proposed method. For stationary noise (white and F16 noise), the fixed threshold is much larger than the likelihood ratio of non-speech frames. In contrast, our adaptive threshold is larger than the log-SLR of

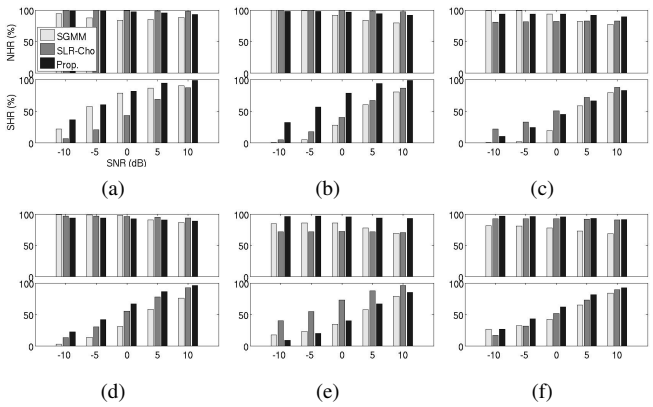

Fig. 2: NHR and SHR for various types of noise and SNRs. (a) white. (b) F16. (c) Buccaneer. (d) destroyerops. (e) babble. (f) fusion.

non-speech frames a little. Consequently, the NHR of SLR-cho are almost 100%, and ours are slightly lower. Nevertheless, the proposed method achieves much higher SHR. For nonstationary noise (Buccaneer and babble noise), the fixed threshold for SLR-Cho is lower than the likelihood ratios for a significant part of non-speech frames, which leads to the decrease of the NHR. For destroyerops noise, the adaptive threshold is close to the fixed one for SLR-Cho, and they lead to comparable results. The results for fusion noise shows that both NHR and SHR benefit from the adaptive threshold. In summary, a fixed threshold is not appropriate for various types of noise. For stationary noise, a large threshold causes the SHR to decrease significantly with only a slight increase of the NHR compared to the adaptive threshold. For nonstationary noise, a small threshold leads to a low NHR.

The SGMM method recursively estimates the log-energy of noise and speech. If the noise energy is underestimated, the NHR decreases, whereas the SHR decreases. For the white noise, the SGMM achieves comparable performance with the proposed method. The proposed method first estimates the noise PSD by an advanced estimator, and then sets the log-SLR threshold adaptively, which makes the noise level estimation more accurate and enables a good tracking for nonstationary noise. Consequently, compared with the proposed method, SGMM's SHR is much worse for F16, Buccaneer and destroyerops noise, NHR is worse for babble noise, and both NHR and SHR are worse for fusion noise.

5. CONCLUSION

In this paper, we have proposed an adaptive threshold for the likelihood ratio test of voice activity detection, which achieves a high NHR while preserving a good SHR performance. Experiments show that this adaptive threshold is able to handle a complex background noise, and obtains better VAD results for various types of noise with respect to the trade-off between NHR and SHR, compared to two baseline state-of-the-art methods.

6. REFERENCES

[1] Tomi Kinnunen, Evgenia Chernenko, Marko Tuononen, Pasi Fränti, and Haizhou Li, "Voice activity detection using MFCC features and support vector machine," in *Int. Conf. on Speech and Computer (SPECOM07)*, Moscow, Russia, 2007, vol. 2, pp. 556–561.

[2] Saman Mousazadeh and Israel Cohen, "Voice activity detection in presence of transient noise using spectral clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1261–1271, 2013.

[3] David Dov, Ronen Talmon, and Israel Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 732–745, 2015.

[4] François Germain, Dennis L Sun, and Gautham J Mysore, "Speaker and noise independent voice activity detection.," in *INTERSPEECH*, Lyon, France, 2013, pp. 732–736.

[5] Prasanta Kumar Ghosh, Andreas Tsiartas, and Shrikanth Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[6] Yanna Ma and Akinori Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–18, 2013.

[7] Javier Ramírez, José C Segura, Carmen Benítez, Angel De La Torre, and Antonio Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[8] Dongwen Ying, Yonghong Yan, Jianwu Dang, and Frank K Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, 2011.

[9] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[10] Yong Duk Cho, Khaldoon Al-Naimi, and Ahmet Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, 2001, vol. 2, pp. 737–740.

[11] Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "MMSE based noise psd tracking with low complexity," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Dallas, USA, 2010, pp. 4266–4269.

[12] Timo Gerkmann and Richard C Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[13] Xiaofei Li, Laurent Girin, Sharon Gannot, and Radu Horaud, "Non-stationary noise power spectral density estimation based on regional statistics," in *IEEE International Conference on Audio, Speech and Signal Processing*, Shanghai, China, 2016.

[14] Rainer Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[15] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[16] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[17] Andrew Varga and Herman JM Steeneken, "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 communication*, vol. 12, no. 3, pp. 247–251, 1993.