A Computation Efficient Voice Activity Detector for Low Signal-to-Noise Ratio in Hearing Aids

1st Fangqi Liu
dept. Electronic and electrical Engineering
University College London
London, United Kingdom
fangqi.liu.14@ucl.ac.uk

2nd Andreas Demosthenous
dept. Electronic and electrical Engineering
University College London
London, United Kingdom
a.demosthenous@ucl.ac.uk

***Abstract*—This paper proposes a spectral entropy-based voice activity detection method, which is computationally efficient for hearing aids. The method is highly accurate at low SNR levels by using the spectral entropy which is more robust against changes of the noise power. Compared with the traditional fast Fourier transform based spectral entropy approaches, the proposed method of calculating the spectral entropy using the outputs of a hearing aid filter-bank significantly reduces the computational complexity. The performance of the proposed method was evaluated and compared with two other computationally efficient methods. At negative SNR levels, the proposed method has an accuracy of more than 5% higher than the power-based method with the number of floating-point operations only about 1/100 of that of the statistical model based method.**

Keywords— Hearing aids, speech processing, spectral entropy, voice activity detection.

I. INTRODUCTION

Voice activity detection (VAD) is one of the essential modules of speech signal processing tasks. In the case of speech enhancement, VAD can be used to update the characteristics of the noise in specific speech enhancement algorithms or optimize the noise reduction strategy over different noise conditions. In hearing aids, noise adaptive speech enhancement requires VAD to update the noise information [1].

Earlier VAD methods are mostly based on energy levels [2], zero crossing rate [3], or cepstral feature [4] of the sampled audio signal. More advanced model-based methods focus on estimating a statistical model for the noisy signal [5]. However, these methods mainly rely on tracking the power of the signal, leading to a low estimation accuracy at low signal (speech) to noise ratio (SNR) levels. Recently, spectral entropy (SpE) has been developed for VAD [6], [7], [8]. Since SpE is more robust against the changes of noise level and more effective in expressing the characteristics of speech signals, it has a higher accuracy at low SNRs [8]. As a trade-off, the computational complexity is increased because of the signal spectrum calculation. More recently, machine learning based approaches [9], [10] have shown promising results in various noise conditions. However, the training and implementation of robust models requires a large amount of computational resource that often has implementation issues on remote devices.

Hearing aids often use VAD as part of their speech enhancement algorithms for speech intelligibility improvement, which is particularly challenging at negative SNRs. The design of VAD with high accuracy at negative SNRs is of particular interest. However, the computational resources of hearing aids are limited due to their small form factor (size)and power constraints. Their clock often works at very low rates to minimize power consumption. Hearing aids demand processing delays within a few milliseconds and reduces the number of frequency bands [11]. In addition, VAD is implemented with speech enhancement algorithms requiring more computational resources. In practice, the machine learning or complex spectrum analysis based VAD methods are often not applicable in hearing aids.

This paper presents a computationally efficient SpE based VAD. In contrast to conventional SpE methods which use the fast Fourier transform (FFT) to calculate the frequency bins, the proposed method calculates the SpE using the outputs of the hearing aid filter-bank with a small number of frequency bands and lower spectral resolution. The VAD classification threshold is the mean value of SpE for acquiring high accuracy at low SNR levels [8]. To be applicable in various noise conditions, the classification threshold is automatically adapted according to the background noise. The detection accuracy and computational efficiency of the proposed method are evaluated and compared with other computationally efficient VAD methods. The rest of the paper is organized as follows. Section II introduces the details of the SpE based VAD. Section III describes the method, dataset and parameters used for evaluation. The evaluation results are presented in Section IV and concluding remarks are drawn in Section V.

II. SPECTRAL ENTROPY-BASED VAD

A. Problem Formulation

The voice activity detection is considered as a segmentation problem in which the short frames of a sampled noisy speech signal are classified as frames that contain speech (speech frames) and frames that only contain noise (nonspeech frames). This is based on the assumption that natural speech is connected with silent pauses [12]. In real acoustic environments, clean speech is contaminated by continuous noise. Thus, the speech pauses would only contain noise that can be classified in a frame-by-frame process. The sampled noisy speech signal $y(i)$ is modelled as the sum of a clean speech signal x(i) and a noise signal $d(i)$:

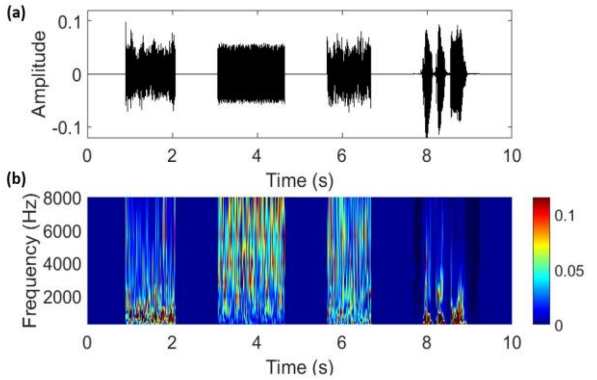

Fig. 1. (a) The waveform consists of 32-talker babble, white, pink noise, and clean speech in sequence. (b) The corresponding spectrogram of the filter-bank outputs at the frequency range between 250 Hz and 8000 Hz. The filter-bank consists of 10 frequency bands and fourth-order Butterworth band-pass filters.

$$y(i) = x(i) + d(i) \tag{1}$$

where i denotes the sample index.

B. Calculating the Spectral Entropy Using Filter-Bank Outputs

The conventional SpE based approaches [8], [13] use the probability associated with spectral energy of each frequency bin of the FFT. In contrast the proposed method uses the instantaneous power of the signal in each frequency band of the filter-bank [14] to calculate the probability. In SpE based approaches, most of the computational resources calculate the spectral characteristics of the noisy speech. Since the outputs of the original filter-bank in the hearing aid are directly used, computational complexity is significantly reduced. Previous work [15] has already shown that a nonlinear filter-bank based SpE can be used for SNR estimation with high accuracy in various noise conditions. The present paper demonstrates a linear filter-bank based SpE with a broader range of applications for different types of hearing aid.

The present paper uses a filter-bank comprising 10 frequency bands (fourth-order Butterworth band-pass filter) was used similar to that in a hearing aid model [16] which simulates the process of human auditory for speech enhancement. The central frequencies (CFs) and the bandwidth of the filter-bank follow the settings in [15]. The CFs of the band-pass filters were logarithmically spaced between 250 Hz and 8000 Hz [17], and the bandwidths (BWs) were calculated using the equivalent rectangular bandwidth (ERB) equation in [18]. As shown in Fig. 1, the filter-bank outputs show that the spectrum of the clean speech is more concentrated at lower frequency range than that of the noise. It indicates that the outputs of the filter-bank with low spectral resolution is able to reflect the spectral differences between clean speech and noise.

To calculate the SpE, the power present probability $p(k,i)$ in frequency band k at the sampling index i is calculated by normalizing the instantaneous spectral power across all frequency bands:

$$p(k,i) = \frac{S(k,i)}{\sum_{l=1}^{K} S(l,i)} \qquad k \in \{1,2,3,\dots,K\} \tag{2}$$

where the instantaneous power $S(k,i)$ is defined by:

$$S(k,i) = |Y(k,i)|^2 \tag{3}$$

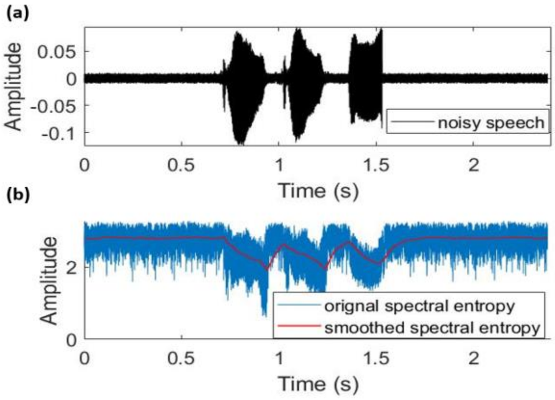

Fig. 2. (a) The utterance “three three four” spoken by a female speaker in white noise at the SNR of 15 dB. (b) The corresponding SpE and the smoothed SpE.

where

$$Y(k,i) = F(k,i) * y(i). \tag{4}$$

$F(k,i)$ is the transfer function of the band-pass filter for frequency band k, and K is the total number of frequency bands in the filter-bank. Based on the equation used in [15], the SpE [h(i)] at sampling index i is defined by:

$$h(i) = -\sum_{k=1}^{L} wf_k \, [p(k,i) log_2 \, p(k,i)]. \tag{5}$$

In contrast to [15], the present study smoothes $h(i)$ over time. This is because the SpE has a large number of variations (as shown in Fig. 2). A sudden rise of the SpE would increase the VAD detection errors. Smoothing was applied using a first-order recursive function:

$$h(i) = \lambda h(i-1) + (1-\lambda) h(i). \dots\dots\dots\dots\dots (6)$$

As shown in Fig. 2(b) the smoothed SpE of the speech signal is very different to that of the nonspeech signal. The mean value (over time) of the spectral entropy (MSpE) is used as a threshold for classification of speech and non-speech frames. The MSpE ($\bar{h}(j)$) of frame j can be calculated by:

$$\bar{h}(j) = \frac{1}{M} \sum_{i=1}^{M} h_j(i) \tag{7}$$

where M is the total number of the sampling points over the frame. In this study, the frame length was 10 ms.

C. Adaptive VAD Threshold

The proposed VAD method is designed to operate on a frame-by-frame process. In each frame, the MSpE of the sampled frame is compared with the classification threshold. In order to adapt to the background noise changes, the classification threshold is continuously updated based on the MSpE of the newly sampled frame. If the sampled frame has a MSpE lower than the threshold, the threshold is updated by smoothing the previous threshold with the MSpE of the current frame; otherwise, the threshold is updated only according to the current frame.

The threshold updating algorithm and nonspeech frame detection strategy were developed based on [2] which estimates the noise power in nonstationary noise by tracking the local minimum power. The approach in [2] was followed because it has lower computational cost and higher accuracy than other approaches [19], [20]. In contrast to [2], which tracks the local minimal power of the noisy speech, the

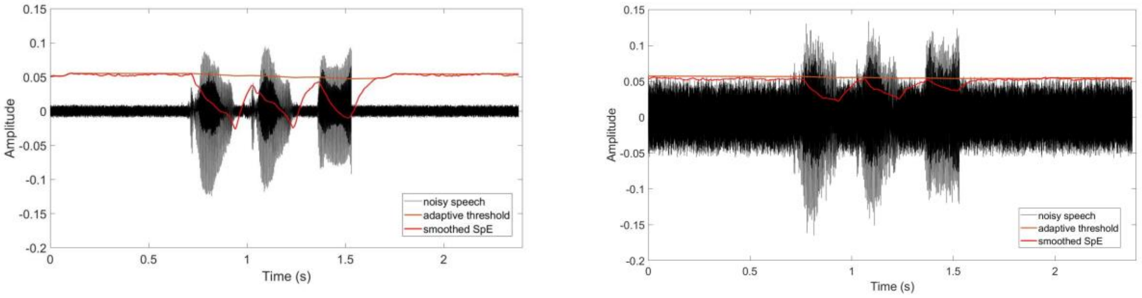

Fig. 3. The adaptive VAD threshold in comparing with the smoothed SpE and the clean speech (“three three four” spoken by a female speaker) in white noise at a SNR level of 15 dB (left) and 0 dB (right).

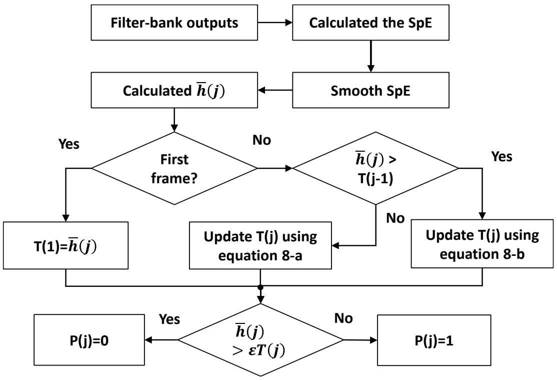

Fig. 4. The flow chart of the proposed VAD algorithm.

approach in the present paper tracks the local maximum MSpE instead because the SpE is more robust at low SNR levels. The classification threshold $T(j)$ is calculated as:

$$T(j) = \begin{cases} \alpha T(j-1) + \dfrac{1-\alpha}{1-\delta}\left(\bar{h}(j) - \delta\bar{h}(j-1)\right) & \bar{h}(j) \le T(j-1) \quad (8-a) \\ \partial \bar{h}(j) & \bar{h}(j) > T(j-1) \quad (8-\mathrm{b}) \end{cases}.$$

The initial value of T is $\bar{h}(1)$.The classification strategy can be described by the following equation:

$$P(j) = \begin{cases} 0 & \bar{h}(j) \le \varepsilon\rho(j-1) \\ 1 & \bar{h}(j) > \varepsilon\rho(j-1) \end{cases} \qquad (9)$$

where $P(j)$ is the speech presence probability of the short frame at index of j, δ and α are factors used for regulating the threshold updating speed, and ε is the decision parameter. These smoothing parameters are used to control the updating speed of the threshold which are determined by the sample rate and the frame length [2]. The parameters were obtained based on practical tests. The smoothed SpE and the updated classification threshold are plotted along with the clean speech in Fig. 3. At a SNR of either 15 dB (left) or 0 dB (right), the smoothed SpE shows apparent decrease when there is a presence of speech envelope. The classification threshold shows reliable tracking of the noise SpE.

D. Computaional Algorithm of the Proposed VAD Method

The algorithm of the proposed VAD is summarized in Fig. 4. First, calculation of the SpE uses Eqs. (2)-(5). After smoothing [Eq. (6)], the MSpE of each frame is calculated [Eq. (7)]. Then, the classification threshold is updated using Eq. (8). Finally, classifying the current frame uses the criteria in Eq. (9).

III. Dataset and Evaluation

A. Dataset

The proposed algorithm was evaluated using noisy speech generated by adding noise to clean speech. White noise, factory, 32-talker babble noise and 2 talker babble noise were used for generating noisy speech at SNR levels between 15 dB and -10 dB. The white and factory noise were drawn from NOISE92 dataset [21]. The 2- and 32-talker babble noise were generated by combining different IEEE speech sentences [15]. Each IEEE speech sentence was normalized at the same level. The clean speech was drawn from the AURORA corpus [22] with 100 utterances spoken by 25 male and 25 female speakers. The isolated utterances were connected to generate continuous noisy speech by adding a silent pause with random lengths between 0-0.4 s. The total length of the noisy speech was 160 s. The non-speech frames of the noisy speech were manually labeled with a percentage of 46%. The sample rate was 16 kHz.

B. Evaluation Method

In order to evaluate the performance of the proposed algorithm, for each type of noise the correct detection probability (Pc) and the incorrect detection probabilities (Pi) at each SNR level were evaluated by following the procedure in [6]. Pc is defined as:

$$Pc = \frac{w_c}{N} \qquad (10)$$

where w_c is the number of correctly detected frames and N is the total number of labeled speech frames. Pi is defined as:

$$Pi = \frac{w_i}{M} \qquad (11)$$

where w_i is the number of incorrectly detected frames and M is the total number of both the speech and nonspeech frames. The VAD test was carried out for each 10 ms frame. The parameters used in the present study for evaluation are listed in Table I. These parameters are determined by the sample rate and frame length and are obtained based on practical tests [2].

TABLE I. Parameters Used in the Evaluation

Parameter	***Value***	***Equation***
λ	0.9996	(6)
α	0.998	(8)
δ	0.9	(8)
∂	0.999	(8)
ε	0.92	(9)

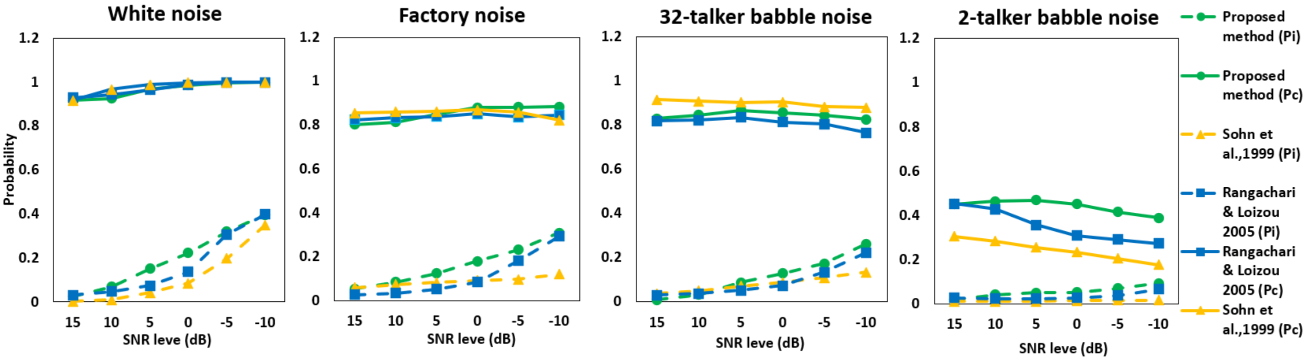

Fig. 5. Evaluated correct detection probability (Pc marked with solid lines) and incorrect detection probability (Pi marked with dash lines) of the proposed method, statistical model based method, and power-based method as a function of SNR level in white, factory,32-talker and 2-talker babble noise.

IV. Results

A. VAD Accuracy

This paper focuses on demonstrating a computationally efficient VAD method with relatively high accuracy. For comparison two algorithms having low computational resources [2],[5] were selected. Specifically, the VAD accuracy of the statistical model method [5], and power based method [2] in white , factory, 32-, and 2 talker babble noise were evaluated. The Pc and Pi of the evaluated methods are plotted as a function of the SNR level in Fig. 5. As shown, for white noise the accuracy of the proposed method is similar to that of the power and statistical model based methods with a Pc of about 0.95. All the tested methods show an increase of Pc with decreasing SNR due to the significant increase of Pi at negative SNRs. The Pi of the proposed method is about 0.04 higher than the power-based method, particularly, at SNR of 5 dB and 0 dB. For factory noise, at SNRs of -10 dB, the Pc of the proposed method is about 0.05 higher than that of the statistical model based method. For 32-talker babble noise, when compared with the power based method, the proposed method shows higher Pc particularly at negative SNRs. However, the Pi of the proposed method is slightly higher than the compared methods at 0 and 5 dB SNR. For 2-talker babble noise, the Pc of the proposed method is much higher than the compared methods over all the SNR levels. Particularly, in comparing with the statistical model based method, Pc has been improved about 0.22 and 0.11 at SNR of 0 and -10 dB respectively.

B. Computational Efficiency

The required CPU time of the proposed method running in MATLAB was compared with that of the statistical model and power based methods. The tested CPU was Intel I7 and the version of MATLAB was 2019b. The total CPU time of processing a 160 s noisy speech was measured. For each of the tested methods, 5 times repeat measurements were performed to reduce errors caused by other CPU background processes. The mean and standard derivation of the CPU time are listed in Table II. To make the results comparable, the CPU time of the filter-bank process was also measured. According to Table II, the CPU time of the proposed method is close to that of the power-based method which is only about 1/10 of the statistical model based method; although the CPU time of the filter-bank process is higher than that of the power-based approach. When implementing VAD in hearing aids, the proposed method can use the hearing aid filter-bank outputs directly. Thus, the CPU time of the filter-bank should not be considered as an extra delay in hardware implementation.

TABLE II. CPU Time WhEN Running the Algorithms in Matlab

	Proposed method	*Statistical model based method*	*Power-based method*	*Filter-bank*
CPU time (s)	0.066 ±0.008	5.609 ±0.011	0.062 ±0.006	2.1311 ±0.036

To estimate the required computational resource for hardware implementation, the number of floating-point operations (FLOPs) of the tested methods were counted. Each method was implemented as a MATLAB function with minimized computational steps. The results were obtained by scanning and parsing each line of the MATLAB codes and inferring the FLOPs. The required mathematical operations were estimated by analyzing the matrix sizes. The total number of FLOPs for processing a 160 s length noisy speech was obtained. The results are listed in Table III. The FLOPs of the proposed algorithm are only about 1/100 of the statistical model based method.

TABLE III. Number of Floating Point Operations

	Proposed method (without filter-bank)	*Statistical model based method*	*Power-based method*
FLOPs	5.815×10^6	5.197×10^8	1.121×10^6

V. Conclusion

A SpE based VAD method, which can directly use the outputs of the hearing aid filter-bank has been presented. The approach reduces computational complexity. The performance of the proposed method has been evaluated and compared with other computationally efficient VAD methods. Results have shown that the proposed method has higher VAD accuracy at SNR level < 0 dB. The required CPU time of the proposed method is close to that of the power-based method and much lower than that of the statistical model based method. The FLOPs of the proposed method are only about 1/100 of that of the statistical model based method.

REFERENCES

[1] I. Panahi, N. Kehtarnavaz, and L. Thibodeau, “Smartphone-based noise adaptive speech enhancement for hearing aid applications,” pp. 85–88, in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, 2016.

[2] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Commun.*, vol. 48, no. 2, pp. 220–231, 2006

[3] J. Junqua, B. Reves, and B. Mak, “A study of endpoint detection algorithms in adverse conditions: incidence on DTW and HMM recognizer,” in *European Conference Speech Communication*, Genova, Italy, September 1991, pp. 927–930.

[4] J. A. Haigh and J. S. Mason, “Robust voice activity detection using cepstral features,” In Proceedings of TENCon'93. IEEE Region 10 International Conference on Computers, Communications and Automation, vol. 3, pp. 321–324, 1993

[5] J. Sohn, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999,

[6] T. H. Zaw and N. War, “The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection,” *20th Int. Conf. Comput. Inf. Technol. ICCIT 2017*, vol. 2018-Janua, pp. 1–5, 2018

[7] B. Lee and D. Muhkerjee, “Spectral entropy-based voice activity detector for videoconfencing systems,” *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010*, Makuhari, Chiba, Japan, September, pp. 3106–3109, 2010.

[8] B. F. Wu and K. C. Wang, “Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 762–774, 2005

[9] J. W. Shin, J. H. Chang, and N. S. Kim, “Voice activity detection based on statistical models and machine learning approaches,” *Comput. Speech Lang.*, vol. 24, no. 3, pp. 515–530, 2009

[10] J. Wu and X.L. Zhang, “Maximum margin clustering based statistical VAD with multiple observation compound feature,” *IEEE Signal Process. Lett. Process.*, vol. 18, no. 5, pp. 283–286, 2011.

[11] R. Gil-Pita, J. García-Gomez, M. Bautista-Durán, E. Combarro, and A. Cocaña-Fernández, “Evolved frequency log-energy coefficients for voice activity detection in hearing AIDS,”In 2017 *IEEE Int. Conf. Fuzzy Syst*, Naples, Italy, pp. 1-6, 2017

[12] B. Zellner, “Pauses and the temporal structure of speech,” in *Fundamentals of speech synthesis and speech recognition. Basic concepts, state of the art and future challenges*, Chichester: John Wiley, 1994, pp. 41–62.

[13] J. Shen, J. Hung, and L. Lee, “Robust entropy-based endpoint detection for speech recognition in noisy environments,” In Fifth international conference on spoken language processing. 1998.

[14] W. Q. Ong, A. W. C. Tan, V. V. Vengadasalam, C. H. Tan, and T. H. Ooi, “Real-time robust voice activity detection using the upper envelope weighted entropy measure and the dual-rate adaptive nonlinear filter,” *Entropy*, vol. 19, no. 11, 2017

[15] F. Liu, A. Demosthenous, and I. Yasin, “Auditory filter-bank compression improves estimation of signal-to-noise ratio for speech in noise,” *J. Acoust. Soc. Am.*, vol. 147, no. 5, pp. 3197–3208, 2020

[16] T. Jürgens, N. R. Clark, W. Lecluyse, and R. Meddis, “Exploration of a physiologically-inspired hearing-aid algorithm using a computer model mimicking impaired hearing,” *Int. J. Audiol.*, vol. 55, no. 6, pp. 346–357, 2016

[17] R. Meddis, L. P. O’Mard, and E. a Lopez-Poveda, “A computational algorithm for computing nonlinear auditory frequency selectivity.,” *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 2852–2861, 2001

[18] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, 1990

[19] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *Speech Audio Process. IEEE Trans.*, vol. 9, no. 5, pp. 504–512, 2001

[20] I. Cohen and B. Baruch, “Speech enhancement for non-stationary noise environments.,” *Signal Processing*, vol. 81.11, pp. 2403–2418, 2001.

[21] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition:{ II}. {NOISEX-92}: A database and an experiment to study the effct of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[22] H. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” *ASR2000- Autom. Speech Recognit.* Challenges new Millenium ISCA Tutor. Res. Work., Paris, France, 2000.