A SPEECH ENHANCEMENT SYSTEM FOR AUTOMOTIVE SPEECH RECOGNITION WITH A HYBRID VOICE ACTIVITY DETECTION METHOD

Haikun Wang[1]*, Zhongfu Ye*[1]*, and Jingdong Chen*[2]

[1]: Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei 230027, China
Email: hkwang@iflytek.com, yezf@ustc.edu.cn
[2]: Center of Intelligent Acoustics and Immersive Communications
Northwestern Polytechnical University, Shaanxi 710072, China
Email: jingdongchen@ieee.org

ABSTRACT

This paper presents a front-end speech enhancement approach to robust speech recognition in automotive environments. It combines hybrid voice activity detection (VAD), relative transfer function (RTF) based generalized sidelobe cancelation, and single-channel post filtering to enhance the speech signal of interest, thereby improving the robustness of speech recognition. First, we choose four typical driving scenarios, which include most of the noise types in automobiles to record training data. The recorded data is then used to train deep neural network models (DNNs) for both speech and noise. The trained DNNs are subsequently used to estimate the speech presence probability on a frame-by-frame basis. This speech presence probability is then combined with the output of an energy-based VAD to form a hybrid VAD, which serves as the basis for the rest components of the speech enhancement system, including RTF estimation, adaptive beamforming, and post-filtering. Experiments are conducted in real automotive environments. The results show that the developed method can significantly improve the performance of both VAD and automatic speech recognition (ASR).

Index Terms—Speech enhancement, deep neural network, voice activity detection, microphone array, speech recognition.

1. INTRODUCTION

Speech interaction based on automatic speech recognition (ASR) in automotive systems is becoming more and more popular in recent years as it can help improve driving safety by enabling hands-free operations. However, noise in automotive environments may dramatically affect the ASR performance and, therefore, speech enhancement is needed in such applications, which has attracted a significant amount of attention over the past decade [1, 2, 3, 4, 5, 6, 7, 8, 9]. Many methods have been developed [10, 11, 12, 13, 14, 15], which have achieved a certain degree of success in either enhancing the quality of hands-free voice communication or improving the ASR performance for human-machine interaction. But dealing with noise in automotive environments remains a challenging problem. This paper studies this problem and presents a speech enhancement approach to robust ASR in automotive environments based on the use of beamforming (with two microphones) and postfiltering techniques.

In automotive environments, the major sources of noise that affect ASR performance can be divided into the following four categories.

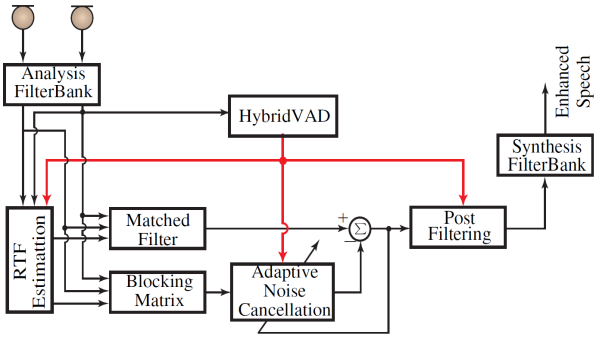

Fig. 1. A schematic diagram of the presented speech enhancement system, where $x_1(t)$ and $x_2(t)$ denote, respectively, the observation noisy signals from the first and second microphones, and $I(k,l)$ and $p(k,l)$ denote, respectively, the presence of speech and the speech presence probability of the kth frequency bin and lth frame.

- Engine noise. This noise is caused by air/fuel mixture in the engine cylinder being ignited. Its level and spectrum are mainly affected by the speed of the engine and acceleration-deceleration. Generally, most energy of engine noise concentrates at the frequencies below 500 Hz.
- Tyre noise. Tyre noise is caused by a number of different factors: tyre rolling, sound of the tread contacting the road, sound of air being compressed inside the tread grooves, etc. The level of this type of noise is principally affected by the speed of the rolling process and the roughness of the road surfaces involved with the rolling. Generally, the frequency components of tyre noise concentrates on the frequency range below 1000 Hz.
- Wind noise. Wind noise is mainly related to the speed of the vehicle and the speed and direction of the wind. This type of noise is typically broadband.
- Ventilator noise. Ventilator noise is generated by the air conditioner. The energy of this noise generally spans up to 4000 Hz.

While they are generated from different sources and have very different statistics and spectra, the aforementioned four types of noise co-exist and are non-stationary in general, which makes speech enhancement a very challenging problem. Another factor that makes speech enhancement in automotive environments a difficult task is

978-1-5386-8151-0/18/$31.00

the low signal-to-noise ratio (SNR), particularly at low and medium frequencies. At low frequencies, it is not uncommon that SNR is below 0 dB. In such challenging SNR environments, voice activity detection (VAD) [18, 19] and power spectral density (PSD) estimation of noise [16, 17], which play a paramount role in speech enhancement performance, is no longer a trivial task.

In this paper, we present a front-end speech enhancement approach to robust ASR for automotive applications. The schematic diagram of the presented method is shown in Fig. 1. Different from other supervised model-based VAD methods [22, 23, 24, 25, 26, 27, 28, 29, 30, 31], we propose to use a compact DNN architecture suitable for online embedded applications to deal with VAD in low SNR conditions. The DNN detection system acquires frame level speech presence/absence information, which is subsequently combined with the output of an energy-based VAD to obtain the frame-and-frequency-bin level speech presence probability. This probability is then served as the basic information for the rest parts of our speech enhancement system, such as RTF estimation, adaptive beamforming, and postfiltering. We evaluate the presented method in terms of speech recall/false alarm rate in VAD and word error rate (WER) of ASR in real, different driving environments and conditions. The results demonstrate the property of the presented speech enhancement method.

2. SIGNAL MODEL AND PROBLEM FORMULATION

We consider the speech enhancement problem with the use of two omnidirectional microphones, which capture a speech signal of interest in some noise field. The received signals are written as

$$\begin{aligned} x_1(t) &= s(t) + v(t), \\ x_2(t) &= a(t) * s(t) + w(t), \end{aligned} \tag{1}$$

where $x_m(t)$ is the signal observed at the mth microphone sensor, $m = 1, 2$, $*$ denotes convolution, $s(t)$ is the desired signal to be enhanced, $v(t)$ and $w(t)$ are, respectively, the noise signals at the two microphones, which may be correlated with each other, but are uncorrelated with $s(t)$, and $a(t)$ denotes the relative transfer function (RTF) between the two microphones, which carries the spatial, temporal, as well as spectral information due to the source, the microphones, and the acoustic environments.

In this work, we consider to work in the short-time-Fourier-transform (STFT) domain to make the implementation efficient. In this domain, the signal model given in (1) are expressed as [12]

$$\begin{aligned} X_1(k,l) &= S(k,l) + V(k,l), \\ X_2(k,l) &= A(k,l)\, S(k,l) + W(k,l), \end{aligned} \tag{2}$$

where k and l denote, respectively, the frequency and frame indices, $X_m(k,l)$, $S(k,l)$, $A(k,l)$, $V(k,l)$, and $W(k,l)$ are the STFTs of $x_m(t)$, $s(t)$, $a(t)$, $v(t)$, and $w(t)$, respectively.

With the signal model given in (2), the objective of speech enhancement is then to estimate the desired signal, $S(k,l)$, given the two observation signals $X_m(k,l)$, $m = 1, 2$. The method to achieve this objective is shown in Fig. 1, the details of which will be explained in the following sections.

3. DNN-BASED VAD

DNN has been studied for VAD over the recent years and showed more promising results than other supervised methods [28, 29, 30, 31]. In this paper, we propose to use a compact DNN with the following architecture: an input layer with 195 units, 2 hidden layers with, respectively, 128 and 64 rectified linear units. As illustrated in Fig. 2, the input to the DNN is a 195-dimensional feature vector [consisting 5 consecutive frames and each frame consisting of 39 perceptual linear prediction coefficients (PLPs)]. The output layer of the DNN consists of two softmax units representing, respectively, the speech and noise posterior probability. This compact DNN is suitable for online and embedded speech enhancement system and is shown to be able to obtain good detection result in automotive applications.

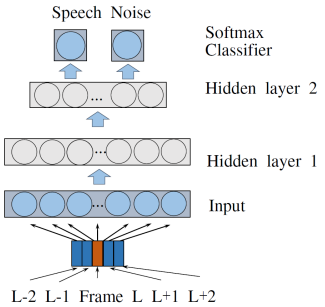

Fig. 2. The DNN architecture used in this paper.

We choose four typical driving situations to record training data, which covers most of the noise types and statistics in automotive environments according to our experimental study.

- Scenario 1: driving speed between $30 - 80$ km/h in downtown environments with all the windows closed. This is the most common driving scenario with different kinds of regular automotive noise.
- Scenario 2: driving speed between $80-120$ km/h on highway with all the windows closed. This is a high-speed driving scenario in which engine and tyre noise dominate.
- Scenario 3: driving speed between $60 - 70$ km/h with all the windows opened. Wind noise is the dominant noise component in this scenario.
- Scenario 4: driving speed $60-70$ km/h with the air conditioner turned to its maximum level, but all the windows closed. In this scenario, ventilator noise dominates.

In each of the aforementioned four scenarios, we recorded 10,000 Chinese sentences in more than 20 different vehicles. These sentences are balanced in phonetic composition. All the recorded data is then labeled to a frame level with the speech presence/absence information. These labeled data is used to train the DNN. The network is trained with backpropagation for 50 epochs (an epoch consisting of 100,000 examples) using mini-batch gradient descent with a mini-batch size of 50 and learning rate of 0.001. Training was accelerated by use of a momentum of 0.9. No pretraining was performed.

In the test process, we first obtain the speech presence probability $p_s(l)$ for every frame from the output unit of the DNN. A smoother probability is then obtained according to the following recursion:

$$p_{\text{model}}(l) = \alpha p_{\text{model}}(l-1) + (1-\alpha) p_s(l), \tag{3}$$

where $p_{\text{model}}(l)$ is the smoothed, model based speech presence probability for frame l and $\alpha \in (0,1)$ is a smoothing factor. In this work, we set $\alpha = 0.85$.

4. HYBRID VAD

Speech presence probability plays a critical role in the accuracy of noise spectrum estimation and gain function estimation [18, 20].

Traditional methods such as the energy-based or minimum statistics based ones often break down in low SNR environments as experienced in automotive applications, which in turn leads to significant speech distortion in the subsequent noise reduction process. In comparison, DNN-based VAD can yield robust estimates in low SNR conditions; but its frame level probability cannot be directly applied to speech enhancement, which operates on a frame-and-frequency-bin basis. In this work, we present a way to combine the frame level VAD from DNN with an energy-based frequency-bin level VAD, which demonstrates great improvement in terms of VAD accuracy and robustness as will become clear in the experiment section of this paper.

We obtain the energy-based frequency-bin level speech presence probability $p_{\text{energy}}(k,l)$ from one of the microphone signal using the method developed in [20]. This speech presence probability is then combined with the frame level DNN-based VAD output $p_{\text{model}}(l)$, i.e.,

$$p(k,l) = \begin{cases} p_{\text{energy}}(k,l), & \text{if } p_{\text{energy}}(k,l) > p_{\text{model}}(l) \\ p_{\text{model}}(l), & \text{otherwise} \end{cases}. \quad (4)$$

In other words, if the speech presence probability estimated from the energy-based VAD is smaller, we choose to trust the results from the DNN-based VAD; otherwise we trust the energy-based VAD. This operation was found through our study to have more accurate speech detection, which leads to less speech distortion in the subsequent speech enhancement stage.

Moreover, we propose the following approximate decision about speech presence, which is called a speech presence indicator:

$$I(k,l) = \begin{cases} 0, & \text{if } p_{\text{energy}}(k,l)\, p_{\text{model}}(l) < \beta \\ 1, & \text{otherwise} \end{cases}, \quad (5)$$

which means the speech presence indicator is zero only when both energy-based VAD and DNN-based VAD generate low speech presence probability. This speech presence indicator will be used in the RTF estimation stage. In this work, we set the value of the threshold β to 0.3.

5. RTF BASED GENERALIZED SIDELOBE CANCELLATION AND POST-FILTERING

To cancel spatially correlated noise, we adopt the RTF based generalized sidelobe canceller (GSC) in this study. In comparison with the traditional GSC using the direction-of-arrival (DOA) information to construct the blocking matrix, the RTF based GSC is robust to the amplitude and phase inconsistency between the microphone sensors as well as the uncertainty in microphones positions.

The most critical step in RFT based GSC is the estimation of RTF $A(k,l)$, which is achieved with a method similar to that in [21]. However, our RTF estimation yields improved performance since in our system the speech presence probability and signal presence indicator are estimated using a hybrid method, which is more robust than the energy-based VAD used in [21].

Now, we discuss the GSC and post-filtering processes based on the use of hybrid VAD and RTF results. The beamforming filter and the blocking matrix based on the estimated RTF are given by

$$\mathbf{w}(k,l) = \frac{1}{1+|\hat{A}(k,l)|^2} \begin{bmatrix} 1 & \hat{A}(k,l) \end{bmatrix}^T, \quad (6)$$

$$\mathbf{B}(k,l) = \begin{bmatrix} -\hat{A}^*(k,l) & 1 \end{bmatrix}^T, \quad (7)$$

where the superscript ^{T} and * denote the transpose and conjugate operators, respectively. Let us denote the outputs of the beamforming filter and blocking matrix, respectively, as $Y_{\text{MF}}(k,l)$ and $Z(k,l)$, i,e.,

$$Y_{\text{MF}}(k,l) = \mathbf{w}^H(k,l) \begin{bmatrix} X_1(k,l) & X_2(k,l) \end{bmatrix}^T, \quad (8)$$

$$Z(k,l) = \mathbf{B}^H(k,l) \begin{bmatrix} X_1(k,l) & X_2(k,l) \end{bmatrix}^T, \quad (9)$$

where the superscript ^{H} is the conjugate-transpose operator. The output of the GSC is then

$$Y_{\text{GSC}}(k,l) = Y_{\text{MF}}(k,l) - H^*(k,l)\, Z(k,l), \quad (10)$$

where the gain, $H(k,l)$, is updated during the absence of speech [i.e., when $I(k,l)=0$] as

$$H(k,l+1) = H(k,l) + \gamma \frac{Y^*_{\text{GSC}}(k,l)\, Z(k,l)}{P_{\text{est}}(k,l)}, \quad (11)$$

with γ being the step size (in this work, we set γ to 0.02),

$$P_{\text{est}}(k,l) = \rho P_{\text{est}}(k,l-1) + (1-\rho)\,|Z(k,l)|^2, \quad (12)$$

and $\rho = 0.95$ being a smoothing factor.

The output of the GSC is further enhanced by a post-filtering process as [17]

$$Y_{\text{PF}}(k,l) = [G(k,l)]^{p(k,l)}\, G_{\min}^{1-p(k,l)} Y_{\text{GSC}}(k,l), \quad (13)$$

where the log-spectral amplitude gain $G(k,l)$ is computed as [18]

$$G(k,l) = \arg\min_{G(k,l)} E\left[\log|S(k,l)| - \log|G(k,l)\, Y_{\text{GSC}}(k,l)|\right]^2, \quad (14)$$

and $G_{\min}$ is the minimal spectral gain.

6. EXPERIMENTS

In this section, we study the performance of the presented method All the test data were recorded in real driving environments. The data set consists of 20 vehicle models and 109 different speakers (59 male and 50 female). Every speaker was asked to sit in the front passenger seat and read 100 Chinese sentences including music search and point of interest (POI) search in the following four different driving scenarios.

- Scenario 1: downtown environments with a driving speed of 40 km/h and all the windows closed.
- Scenario 2: highway with a driving speed of 100 km/h and all the windows closed.
- Scenario 3: highway with a driving speed of 60 km/h and all the windows opened.
- Scenario 4: highway with a driving speed of 60 km/h, the air conditioner being turned to its maximum level, and all the windows closed.

The two-microphone array is mounted in the center of the central control panel. The spacing between the two microphones is 8 cm The distance between the speaker and the microphone array varies with the height of the speaker and the size of the car and it is generally with in the range of $40-60$ cm.

The acoustic model of the ASR system is a DNN based one with 6 hidden layers and 2048 nodes for each layer. It was trained using 15,340 hours of speech data recorded in various driving conditions. The Language model is a 5-gram model trained with textual data consisting of music song titles, artists, POI names, etc.

Table 1. Frame level speech recall rate (%) and false alarm rate (%) of the energy-based VAD, complex DNN based VAD, and the presented DNN based VAD in four driving scenarios.

Method	Speech recall rate				Speech false alarm rate			
	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Energy-based VAD	77.2	77.4	36.6	71.2	7.5	27.3	6.4	26.4
Complex DNN VAD	93.4	88.4	90.3	88.7	5.5	25.4	5.3	22.3
Proposed DNN VAD	93.1	87.9	90.4	88.5	5.9	25.7	5.6	22.0

Table 2. WERs(%) of unprocessed noisy speech, speech enhanced by the dual microphone enhancement method with the energy-based VAD, and speech enhanced by the dual microphone enhancement method with the hybrid VAD in four driving scenarios.

Condition	WERs (%)		
	Noisy speech observed at one microphone	Enhanced speech with RTF Estimation and energy-based VAD	Enhanced speech with RTF Estimation and Hybrid VAD
Scenario 1	9.3	5.6	**4.2**
Scenario 2	23.5	14.2	**9.8**
Scenario 3	13.2	9.3	**6.1**
Scenario 4	18.7	11.9	**7.9**
Average	16.2	10.3	**7.0**

In the first experiment, we examine the VAD performance of our compact DNN based method and compare it with that of the complex DNN method presented in [30], which has 3 hidden layers, each containing 512 Rectified Linear Units. For comparison, we also evaluate the VAD performance of an energy-based VAD method. The results are shown in Table 1. It is seen that the present DNN method yielded comparable result with the complex DNN based method, and the VAD performance of both DNN methods in terms of the speech recall rate and false alarm rate is much better than that of the energy-based VAD.

In the second experiment, we investigate the ASR performance of the presented system in the four different driving scenarios. The ASR performance is evaluated in terms of WER. The results are presented in Table 2. As seen, the presented dual microphone enhancement method can significantly improve the ASR performance. If the energy-based VAD method is used, the average WER is 10.3%, which translates to a relative WER reduction of approximately 36.42% as compared to the recognition with the noisy speech from one of the two microphones. If the hybrid VAD method is used, the average WER drops to 7.0%, so an additional WER reduction of 32.04% is achieved, indicating the effectiveness of the presented speech enhancement method and the hybrid VAD approach.

Figure 3 plots the spectrograms of a segment of the noisy speech signal, the corresponding enhanced speech with the energy-based VAD, and the enhanced speech signal with the hybrid VAD. It is clearly seen that both enhancement methods have significantly attenuated the noise. In comparison, the method based on the hybrid VAD has a smaller amount of speech distortion.

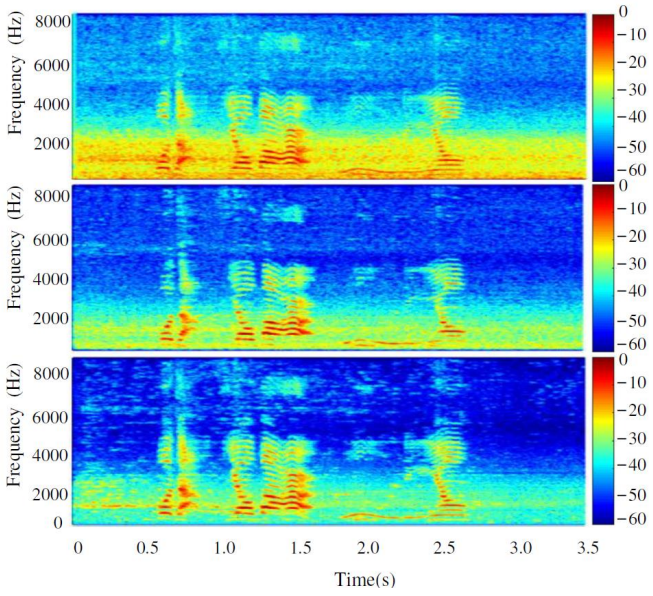

Fig. 3. Spectrograms: a segment of the noisy speech signal (top), the enhanced signal with two microphones using an energy-based VAD (middle), and the enhanced signal with two microphones using hybrid VAD (bottom).

7. CONCLUSIONS

In this paper, we presented a two-microphone speech enhancement approach to robust speech recognition in automotive environments. The major contributions are as follows. 1) A hybrid VAD method is presented, which combines a compact DNN-based VAD and an energy-based one. This hybrid method works effectively in various driving conditions and even when the SNR is below 0 dB. 2) A two-microphone speech enhancement method is developed, which consists of hybrid VAD, RTF estimation, GSC, and Post-filtering. Experiments have demonstrated the advantage of this enhancement method for ASR in real automotive applications.

8. ACKNOWLEDGEMENTS

The authors would like to thank IFLYTEK Research for sharing their ASR system and automotive database with us.

9. REFERENCES

[1] S. Nordholm, I. Claesson, and N. Grbić, "Optimal and adaptive microphone arrays for speech input in automobiles," in *Microphone Arrays*, pp. 307–329, 2001.

[2] D. Ayllón, V. Benito-Olivares, and C. Llerena-Aguilar, "Optimum microphone array for hands-free devices in a car," in *Proc. WSEAS*, pp. 88–92, 2011.

[3] J.-T. Chien and P.-Y. Lai, "Car speech enhancement using a microphone array," *International Journal Speech Tech.*, vol. 8, no. 1, pp. 79–91, 2005.

[4] J. Whittington, H. Ye, K. Kamalakannan, N. Vu, M. Mason, T. Kleinschmidt, and S. Sridharan, "Low-cost hardware speech enhancement for improved speech recognition in automotive environments," in *Proc. ARRB*, pp. 1–17, 2010.

[5] M. Buck, T. Wolff, T. Haulick, and G. Schmidt, "A compact microphone array system with spatial post-filtering for automotive applications," in *Proc. IEEE ICASSP*, pp. 221–224, 2009.

[6] R. Chen, C.-F. Chan, H.-C. So, J. S. Lee, and C. Leung, "Speech enhancement in car noise environment based on an analysis-synthesis approach using harmonic noise model," in *Proc. IEEE ICASSP*, pp. 4413–4416, 2009.

[7] N. Hanai and R. M. Stern, "Robust speech recognition in the automobile.," in *Proc. ICSLP*, 1994.

[8] M. Krini and G. Schmidt, "Model-based speech enhancement for automotive applications," in *Proc. IEEE ISPA*, pp. 632–637, 2009.

[9] M. Buck and M. Rößler, "First order differential microphone arrays for automotive applications," in *Proc. IWAENC*, 2001.

[10] J. Benesty, J. Chen, and E. A. Habets, *Speech Enhancement in the STFT domain*. Berlin, Germany: Springer-Verlag, 2011.

[11] S. Gannot, D. Burshtein, and E. Weinstein, "Analysis of the power spectral deviation of the general transfer function GSC," *IEEE Trans. Signal Process.*, vol. 52, pp. 1115–1120, Apr. 2004.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Berlin, Germany: Springer-Verlag, 2008.

[13] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[14] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "A framework for speech enhancement with ad hoc microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1038–1051, 2016.

[15] L. Huang, J. Zhang, X. Xu, and Z. Ye, "Robust adaptive beamforming with a novel interference-plus-noise covariance matrix reconstruction method," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1643–1650, 2015.

[16] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1149–1160, 2004.

[17] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 6, pp. 561–571, 2004.

[18] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[19] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

[20] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[21] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[22] J. W. Shin, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech, Lang.*, vol. 24, no. 3, pp. 515–530, 2010.

[23] P. Teng and Y. Jia, "Voice activity detection via noise reducing using non-negative sparse coding," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 475–478, 2013.

[24] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering." *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1261–1271, 2013.

[25] T. Yu, "Discriminative training for multiple observation likelihood ratio based voice activity detection,." *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 897–900, 2010.

[26] D. Ying, "Voice activity detection based on an unsupervised learning framework, " *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2624–2644, 2011.

[27] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 507–510, 2012.

[28] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE ICASSP*, 2013, pp. 7378–7382.

[29] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, 2013.

[30] N. Ryant, "Speech activity detection on youtube using deep neural networks," in *Proc. Interspeech*, 2013, pp. 728-731.

[31] Zhang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. Interspeech*, 2014, pp. 1534-1537.