AUC OPTIMIZATION FOR DEEP LEARNING BASED VOICE ACTIVITY DETECTION

Zi-Chen Fan, Zhongxin Bai, Xiao-Lei Zhang, Susanto Rahardja, and Jingdong Chen

Center for Intelligent Acoustics and Immersive Communications and
School of Marine Science and Technology, Northwestern Polytechnical University

ABSTRACT

Voice activity detection (VAD) based on deep neural networks (DNN) has demonstrated good performance in adverse acoustic environments. Current DNN based VAD optimizes a surrogate function, e.g. minimum cross-entropy or minimum squared error, at a given decision threshold. However, VAD usually works on-the-fly with a dynamic decision threshold; and ROC curve is a global evaluation metric of VAD that reflects the performance of VAD at all possible decision thresholds. In this paper, we propose to optimize the area under ROC curve (AUC) by DNN, which can maximize the performance of VAD in terms of the ROC curve. Experimental results show that optimizing AUC by DNN results in higher performance than the common method of optimizing the minimum squared error by DNN.

Index Terms— AUC, deep neural networks, voice activity detection.

1. INTRODUCTION

Voice activity detection (VAD) aims to separate target voices from background noises. How to make it effective in low signal-to-noise ratio (SNR) environments is a challenge. Early research on VAD focused on the statistics of acoustic features, including energy in the time domain, zero-crossing rate, pitch detection [1], cepstral coefficients [2], higher-order statistics [3], etc. However, a single acoustic feature reflects only part characteristics of human voice, which may be ineffective in some difficult scenarios when used alone.

Statistical signal processing based VAD, which fits signals to predefined models and updates the parameters of a prior probability distribution in an online learning mode, is another major research branch. An accurate model assumption to the real-world distribution of speech data is a crucial problem. Existing model assumptions include Gaussian [4, 5], Laplacian [6], Gamma distributions [7] and their combinations [8]. However, they use limited local data to train/update model parameters, leaving large amount of prior knowledge unexplored. Moreover, real-world data distributions may be too complicated to be modeled accurately by a predefined model assumption.

Recently, supervised learning based VAD, which regards VAD as a classification problem, has received much attention. It is flexible in incorporating prior knowledge, such as manually labeled data. It is also good at fusing multiple acoustic features. Existing supervised models include linear discriminant analysis [9], support vector machines (SVM) [10], multimodal methods [11] , sparse coding [12, 13], and deep neural networks (DNN) [14–25]. Particularly, DNN has demonstrated strong scalability in building multiple layers of nonlinear transforms on a large-scale training corpus, e.g. [18], which is important to make off-line supervised training methods practical towards real-world applications. Hence, there is a bloom on the development of DNN-based VAD methods, which has focused mainly on two respects—acoustic features, e.g. [14, 18, 23, 24], and deep models, e.g. [16, 17, 19, 21, 25].

An important missing aspect of the DNN-based VAD research is on the training target. It is known that the decision threshold of VAD is usually determined on-the-fly, and different applications may have different minimum requirements to the missing detection rate. Hence, it is needed to optimize the performance of VAD at a wide range of decision thresholds. Moreover, receiver operating characteristic (ROC) curve and the area under ROC curve (AUC) are two standard evaluation metrics to measure the global performance of VAD. However, the training costs of existing DNN based VADs are either classification-loss-based minimum cross-entropy [14] or regression-loss-based minimum square error (MSE) [18], which are both surrogate loss functions that do not optimize ROC curve or AUC directly.

Motivated by [10], this paper proposes a DNN-based VAD to optimize AUC directly. The method, named *MaxAUC-DNN*-based VAD first relaxes the AUC calculation, which is an NP-hard problem, to a polynomial-time solvable problem, then calculates the gradient of the AUC loss with respect to the parameters of the output layer of DNN, and finally back-propagates the gradient to the entire DNN. We have evaluated the proposed method in babble and factory noise scenarios at a wide range of SNR levels. Experimental results show that the MaxAUC-DNN-based VAD outperforms the MSE-DNN-

This work was supported in part by the National Natural Science Foundation of China (NSFC) funding scheme under Project No. 61671381 and in part by the Shaanxi Natural Science Basic Research Program under grant No. 2018JM6035. (*Corresponding author: Xiao-Lei Zhang*)

Zi-Chen Fan and Zhongxin Bai contribute equally to this paper as the co-first authors.

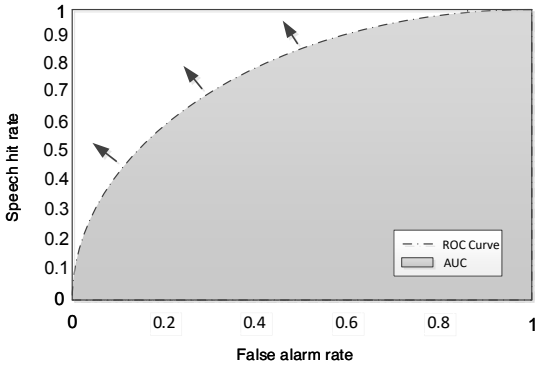

Fig. 1. Illustration of ROC curve and AUC

based VAD, given either the short-time Fourier transform (STFT) or multi-resolution cochleagram (MRCG) [18, 26] as the acoustic feature.

2. MAXAUC-DNN BASED VAD

2.1. Motivation and problem formulation

Supervised learning based VAD can be viewed as a binary classification problem—speech or nonspeech.[1] Suppose we have a training corpus $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i$ is a high-dimensional acoustic feature of the i-th frame signal, e.g. STFT or MRCG, and y_i is the ground-truth label of $\mathbf{x}_i$. If $\mathbf{x}_i$ is a speech frame, then $y_i = 1$; otherwise, $y_i = 0$. DNN based VAD aims to learn a multilayer nonlinear mapping function $f_\alpha(\cdot)$ from $\mathcal{X}$, such that when using $f_\alpha(\cdot)$ in test by the following criterion, its performance can be optimized:

$$\overline{y} = \begin{cases} 1, & \text{if } f_\alpha(\mathbf{x}) \geq \eta \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where α is the model parameter of DNN $f_\alpha(\cdot)$, and η is a decision threshold. The training objective of $f_\alpha(\cdot)$ falls into the following two classes: (i) MSE minimizes $\sum_{i=1}^n ||\mathbf{y}_i - f_\alpha(\mathbf{x}_i)||^2$ with respect to α, and (ii) minimum cross-entropy minimizes $-\sum_{i=1}^n (y_i \log(f_\alpha(\mathbf{x}_i)) + (1 - y_i)\log(1 - f_\alpha(\mathbf{x}_i)))$. These surrogate functions cannot guarantee that the VAD performs the best at any valid decision threshold. However, the decision threshold of VAD is usually decided on-the-fly. It may vary significantly in different applications. Based on the above analysis, it is needed to optimize the performance of VAD at any decision threshold, which is the motivation of this paper as shown in Fig. 1.

2.2. Optimization objective

Let X be the acoustic feature space, and $\mathcal{D}_+$ and $\mathcal{D}_-$ are the probability distributions of speech and silence on X, respectively. We aim to learn a DNN function $f_\alpha : X \to \mathbb{R}$ which maximizes the AUC.

Definition 1. *The detection probability $P_{\rm D}$ and the false alarm probability $P_{\rm FA}$ is defined respectively as*

$$P_{\rm D}(\gamma) = \mathbf{P}_{\mathbf{x}^+ \sim \mathcal{D}_+}[f_\alpha(\mathbf{x}^+) > \gamma] \tag{2}$$

$$P_{\rm FA}(\gamma) = \mathbf{P}_{\mathbf{x}^- \sim \mathcal{D}_-}[f_\alpha(\mathbf{x}^-) > \gamma] \tag{3}$$

where $\mathbf{x}^+$ and $\mathbf{x}^-$ are two random samplings form X according to $\mathcal{D}_+$ and $\mathcal{D}_-$ respectively, and $\gamma \in \mathbb{R}$ is the threshold.

Given Definition 1, the ROC curve produced by f_α is then defined as the plot of $P_{\rm FA}(\gamma)$ against $P_{\rm D}(\gamma)$ at all possible γ as show in Fig. 1. Then the AUC is calculated by

$$\text{AUC}_{f_\alpha} = \int_0^1 P_{\rm D}(P_{\rm FA}^{-1}(u))du \tag{4}$$

where $P_{\rm FA}^{-1}(u) = \inf\{\gamma \in \mathbb{R} | \mathrm{P}_{\rm FA}(\gamma) \leq \mathrm{u}\}$. In practice, because the training samples from X are limited to $\mathcal{X}$, we only estimate an approximate AUC of (4) from $\mathcal{X}$.

We define the positive and negative samples in $\mathcal{X}$ as $\mathcal{X}^+ = \{(\mathbf{x}_j^+, y_j = 1) | j = 1, 2, 3, \cdots J\}$ and $\mathcal{X}^- = \{(\mathbf{x}_k^-, y_k = 1) | k = 1, 2, 3, \cdots K\}$, respectively, with $n = J + K$. We further define the optimization objective of the MaxAUC-DNN as follows:

Theorem 1. *Given the limited training samples of $\mathcal{X}^+$ and $\mathcal{X}^-$, a loss function for optimizing the AUC* (4) *can be defined as:*

$$\ell = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \max\left[0, \delta - f_\alpha\left(\mathbf{x}_j^+\right) + f_\alpha\left(\mathbf{x}_k^-\right)\right] \tag{5}$$

where δ is a tunable hyperparameter.

Proof. The AUC (4) can be calculated by [27]:

$$\text{AUC}_{f_\alpha} = \mathbf{P}_{(\mathbf{x}^+, \mathbf{x}^-) \sim \mathcal{D}_+ \times \mathcal{D}_-}[f_\alpha(\mathbf{x}^+) > f_\alpha(\mathbf{x}^-)] \tag{6}$$

Given the limited training samples $\mathcal{X}^+$ and $\mathcal{X}^-$, the empirical AUC is calculated by

$$\widehat{\text{AUC}}_{f_\alpha} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbb{I}\left[f_\alpha\left(\mathbf{x}_j^+\right) > f_\alpha\left(\mathbf{x}_k^-\right)\right] \tag{7}$$

where $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the statement is true, and 0 otherwise. However, maximizing (7) directly is NP-hard. To deal with this problem, we relax (7) by replacing the indicator function by a hinge loss function:

$$\ell_{hinge}(z < 0) = \max(0, \delta - z) \tag{8}$$

where $z = f_\alpha\left(\mathbf{x}_j^+\right) - f_\alpha\left(\mathbf{x}_k^-\right)$, and $\delta > 0$ is a tunable hyperparameter controlling the distance margin between $f_\alpha(\mathbf{x}_j^+)$ and $f_\alpha(\mathbf{x}_k^-)$. Substituting (8) into (7) transforms the maximization problem of (7) into a minimization problem of (5). Theorem 1 is proved.

□

[1] Nonspeech contains many noise scenarios. Hence, VAD is a problem of discriminating one class to the rest classes rigorously.

2.3. Optimization algorithm

In this paper, we employ the mini-batch stochastic gradient descent algorithm to solve (5). Because the gradient $\nabla f_\alpha(\mathbf{x}_i)$ with respect to $\mathbf{x}_i$ can be easily backpropagated throughout the network in a standard procedure, we only need to derive the gradient at the output layer.

Theorem 2. *The gradient of* (5) *at the output layer of DNN is:*

$$\nabla \ell = -\sum_{i=1}^{n} y_i' \times \omega_i \times \nabla f_\alpha(\mathbf{x}_i) \tag{9}$$

where $y_i' = 2y_i - 1$ *is the ground-truth label of* $\mathbf{x}_i$, $\{\omega_i | i = 1 \cdots n\} = \{\omega_j | j = 1 \cdots J\} \cup \{\omega_k | k = 1 \cdots K\}$ *is the weight of* $\mathbf{x}_i$ *with* $\omega_j = \frac{1}{JK}\sum_{k=1}^{K} \Pi(j,k)$ *and* $\omega_k = \frac{1}{JK}\sum_{j=1}^{J} \Pi(j,k)$, *where* $\Pi \in \{0,1\}^{J \times K}$ *is an index matrix defined as*

$$\Pi(j,k) = \begin{cases} 1, & \textit{if } f_\alpha(\mathbf{x}_j^+) < \delta + f_\alpha(\mathbf{x}_k^-) \\ 0, & \textit{otherwise} \end{cases}. \tag{10}$$

Proof. The gradient of (5) is

$$\nabla \ell = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} \nabla \ell_{hinge} \tag{11}$$

where the gradients $\nabla \ell_{hinge}$ with respect to $\mathbf{x}_j^+$ and $\mathbf{x}_k^-$ are given respectively by

$$\nabla \ell_{hinge}(\mathbf{x}_j^+) = \begin{cases} -\nabla f_\alpha(\mathbf{x}_j^+), & \text{if } f_\alpha(\mathbf{x}_j^+) < \delta + f_\alpha(\mathbf{x}_k^-) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$\nabla \ell_{hinge}(\mathbf{x}_k^-) = \begin{cases} \nabla f_\alpha(\mathbf{x}_k^-), & \text{if } f_\alpha(\mathbf{x}_j^+) < \delta + f_\alpha(\mathbf{x}_k^-) \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

Taking (12) and (13) into (11) obtains:

$$\nabla \ell = -\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} \Pi(j,k)\left[\nabla f_\alpha(\mathbf{x}_j^+) - \nabla f_\alpha(\mathbf{x}_k^-)\right] \tag{14}$$

For simplicity, (14) can be rewritten as:

$$\begin{aligned} \nabla \ell &= -\Big(\sum_{j=1}^{J} \omega_j \nabla f_\alpha(\mathbf{x}_j^+) - \sum_{k=1}^{K} \omega_k \nabla f_\alpha(\mathbf{x}_k^-)\Big) \\ &= -\sum_{i=1}^{n} y_i' \times \omega_i \times \nabla f_\alpha(\mathbf{x}_i). \end{aligned} \tag{15}$$

Theorem 2 is proved. □

In this paper, the activation function of the output layer of DNN is sigmoid function. Because $y_i' \in \{1, -1\}$, we map the output of the sigmoid function from $[0, 1]$ to $[-1, 1]$.

3. EXPERIMENTS

3.1. Datasets

We conducted an experimental comparison on the CHiME-4 challenge. The audio data are 16 bit stereo WAV files sampled at 16 kHz. We used the "tr05_org" corpus of CHiME-4 as clean speech. We selected 6,340 sentences for training, and 798 sentences for testing. We constructed a number of noisy development sets by adding PED noise in CHiME-4 to the clean speech at SNR levels $[-10, -5, 0, 5]$ dB respectively for the hyperparameter δ selection problem. Then, we constructed noisy training and test sets by adding babble and factory noise respectively from the NOISEX-92 database to the clean speech at the same SNR levels as the development sets for a formal comparison. The SNR levels and noise types of training and test were matching in all experiments. For each pair of training and test corpora, we split the noise source into two segments, one added to the training data and the other to the test data.

Because CHiME-4 does not have sample-level ground-truth labels, we used the prediction result of the Sohn VAD [4] on the clean speech as the ground-truth labels. This labeling method has been shown to be reliable in [18].

3.2. Experimental settings

We set the frame length to 30 milliseconds and frame shift to 10 milliseconds. To verify the effectiveness of the proposed algorithm with different acoustic features, we took STFT and MRCG features respectively as the input of a DNN model. STFT is a basic acoustic feature, while MRCG is an advanced one. Based on the experimental conclusion in Section 3.4, we set the hyperparameter δ of the MaxAUC-based VAD to 0.8 when MRCG was used, and 0.1 when STFT was used.

We compared the MaxAUC-DNN-based VAD with the MSE-based VAD. Both of their DNN models adopted the same hyperparameter setting as follows. Each DNN model contains two hidden layers. The numbers of hidden units were set to 512 for the two hidden layers. The activation functions of the hidden units and output units were set to rectified linear unit and sigmoid function, respectively. The number of epoches was set to 30. The batch size was set to 512. The scaling factor for the adaptive stochastic gradient descent was set to 0.0015, and the learning rate decreased linearly from 0.08 to 0.001. The momentum of the first 5 epochs was set to 0.5, and the momentum of other epochs was set to 0.9. The dropout rate of the hidden units was set to 0.2. A contextual window was used to expand each input frame to its context along the time axis. The window size was set to 5.

We adopted the ROC curve and AUC as the evaluation metrics.

Table 2. Effect of hyperparameter δ on performance in terms of AUC in PED noise.

Feature	SNR	MaxAUC										MSE
		δ =0.1	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.5$	$\delta = 0.6$	$\delta = 0.7$	δ =0.8	$\delta = 0.9$	$\delta = 1.0$	
MRCG	5 dB	0.8977	0.9024	0.9051	0.9063	0.9097	0.9092	0.9087	**0.9104**	0.9057	0.8995	0.9033
	0 dB	0.8612	0.8651	0.8679	0.8719	0.8756	0.8751	0.8760	0.8766	**0.8769**	0.8728	0.8663
	−5 dB	0.7842	0.7921	0.7908	0.7997	0.7990	0.7971	0.8047	0.8119	**0.8142**	0.8135	0.7983
	−10 dB	0.6980	0.6983	0.7036	0.7098	0.7128	0.7141	0.7217	0.7239	0.7308	**0.7382**	0.7074
STFT	5 dB	**0.8405**	0.8232	0.8046	0.7813	0.6858	0.7341	0.7534	0.6803	0.6224	0.6024	0.7555
	0 dB	**0.7867**	0.7658	0.7404	0.6617	0.6128	0.6322	0.6418	0.6139	0.5938	0.5914	0.6528
	−5 dB	**0.7304**	0.6766	0.6204	0.5960	0.5922	0.5928	0.5930	0.5907	0.5893	0.5889	0.5954
	−10 dB	**0.6356**	0.5845	0.5839	0.5882	0.5889	0.5886	0.5876	0.5889	0.5889	0.5889	0.5882

Table 1. AUC comparison in babble and factory noises.

Noise	Feature	SNR	MaxAUC	MSE
Babble	MRCG	5 dB	**0.9208**	0.9183
		0 dB	**0.8922**	0.8888
		−5 dB	**0.8455**	0.8410
		−10 dB	**0.7742**	0.7601
	STFT	5 dB	**0.8699**	0.8507
		0 dB	**0.8259**	0.7346
		−5 dB	**0.7640**	0.6424
		−10 dB	**0.6954**	0.5902
Factory	MRCG	5 dB	**0.9065**	0.9059
		0 dB	**0.8815**	0.8809
		−5 dB	**0.8409**	0.8353
		−10 dB	**0.7743**	0.7675
	STFT	5 dB	**0.8144**	0.8131
		0 dB	**0.7509**	0.7207
		−5 dB	**0.6671**	0.6039
		−10 dB	**0.5996**	0.5893

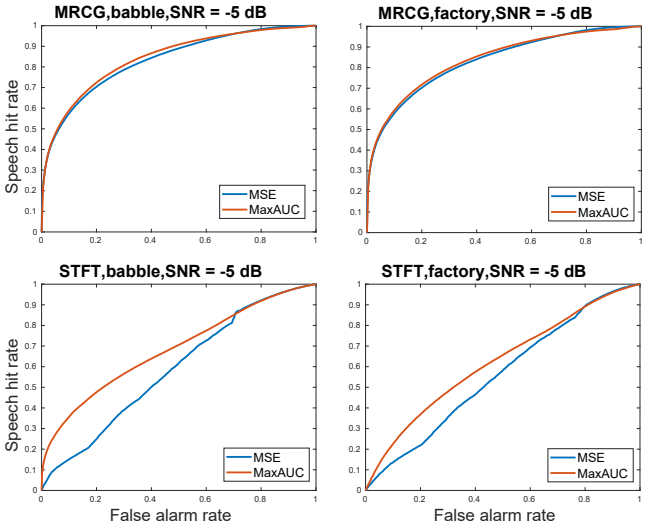

Fig. 2. ROC curve comparison at an SNR of −5 dB.

3.3. Main results

Table 1 and Fig. 2 list the comparison results. From the table and figure, we see that the MaxAUC-DNN-based VAD clearly outperforms the MSE-DNN-based VAD at all SNR levels. The relative improvement with STFT is much larger than that with MRCG. For example, the MaxAUC-DNN based VAD achieves more than 8% improvement over the MSE-DNN-based VAD in the babble noise. Moreover, the MaxAUC-DNN-based VAD behaves more robust across different acoustic features than the MSE-DNN-based VAD.

3.4. Effect of hyperparameter δ

We tuned hyperparameter δ in grid from 0.1 to 1 with a step size 0.1 in the PED noise environment. Table 2 lists the experimental results. From the table, we observe that δ is insensitive to SNR, and behaves robustly in a wide range. However, it is sensitive to different acoustic features. Specifically, for the MRCG feature, the MaxAUC-DNN-based VAD with δ selected from $[0.4, 1]$ outperforms the MSE-DNN-based VAD; the best δ appears around 0.8 at all SNR levels. For the STFT feature, when δ is selected from $[0.1, 0.4]$, the MaxAUC-DNN-based VAD outperforms the MSE-DNN-based VAD; and the best δ appears around 0.1. Therefore, we chose $\delta = 0.8$ for MRCG, and $\delta = 0.1$ for STFT in Section 3.3 where the experimental results further recognize our conclusion here.

4. CONCLUSIONS

In this paper, we have proposed a DNN based VAD for maximizing AUC directly, so as to improve the performance of DNN based VAD at any decision threshold. Specifically, we first relax the AUC calculation to a polynomial-time solvable problem, then compute the gradient of the AUC loss with respect to the parameters of the output layer of DNN, and finally back-propagate the gradient to its hidden layers. We have evaluated the proposed method in babble and factory noise scenarios at a wide range of SNR levels. Experimental results show that the proposed method outperforms the MSE-DNN-based VAD at all SNR levels, given either STFT or MRCG as the acoustic feature; moreover, it is insensitive to δ.

5. REFERENCES

[1] R Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, no. 4, pp. 377–380, 1992.

[2] J-C Junqua and Hisashi Wakita, "A comparative study of cepstral lifters and distance measures for all pole models of speech in noise," in *Proc. ICASSP*, 1989, pp. 476–479.

[3] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans., Speech, Audio Process.*, vol. 9, no. 3, pp. 217–231, 2001.

[4] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[5] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, 2005.

[6] Joon-Hyuk Chang and Nam Soo Kim, "Voice activity detection based on complex laplacian model," *Electron. Lett.*, vol. 39, no. 7, pp. 632–634, 2003.

[7] Jong Won Shin, Joon-Hyuk Chang, and Nam Soo Kim, "Statistical modeling of speech signals based on generalized gamma distribution," *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 258–261, 2005.

[8] Joon Hyuk Chang, Nam Soo Kim, and Sanjit K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, 2006.

[9] Jaume Padrell, Dusan Macho, and Climent Nadeu, "Robust speech activity detection using lda applied to ff parameters," in *Proc. ICASSP*, 2005, vol. 1, pp. I–557.

[10] Ji Wu and Xiao-Lei Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 466–499, 2011.

[11] David Dov, Ronen Talmon, and Israel Cohen, "Multimodal kernel method for activity detection of sound sources," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1322–1334, 2017.

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

[13] Shi-Wen Deng and Ji-Qing Han, "Statistical voice activity detection based on sparse representation over learned dictionary," *Digital Signal Process.*, vol. 23, no. 4, pp. 1228–1232, 2013.

[14] Xiao-Lei Zhang and Ji Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, 2013.

[15] Xiao-Lei Zhang and Ji Wu, "Denoising deep neural networks based voice activity detection," in *Proc. ICASSP*, 2013, pp. 853–857.

[16] Thad Hughes and Keir Mierle, "Recurrent neural networks for voice activity detection," in *Proc. ICASSP*, 2013, pp. 7378–7382.

[17] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Proc. ICASSP*, 2013, pp. 483–487.

[18] Xiao-Lei Zhang and DeLiang Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 2, pp. 252–264, 2016.

[19] Inyoung Hwang, Hyung-Min Park, and Joon-Hyuk Chang, "Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection," *Computer Speech & Lang.*, vol. 38, pp. 1–12, 2016.

[20] Qing Wang, Jun Du, Xiao Bao, Zi-Rui Wang, Li-Rong Dai, and Chin-Hui Lee, "A universal vad based on jointly trained deep neural networks," in *Proc. Interspeech*, 2015, pp. 1210–1214.

[21] Longbiao Wang, Khomdet Phapatanaburi, Zeyan Go, Seiichi Nakagawa, Masahiro Iwahashi, and Jianwu Dang, "Phase aware deep neural network for noise robust voice activity detection," in *Proc. ICME*, 2017, pp. 1087–1092.

[22] Jong Hwan Ko, Josh Fromm, Matthai Philipose, Ivan Tashev, and Shuayb Zarar, "Limiting numerical precision of neural networks to achieve real-time voice activity detection," in *Proc. ICASSP*, 2018, pp. 2236–2240.

[23] Yuuki Tachioka, "DNN-based voice activity detection using auxiliary speech models in noisy environments," in *Proc. ICASSP*, 2018, pp. 5529–5533.

[24] Wissam A. Jassim and Naomi Harte, "Voice activity detection using neurograms," in *Proc. ICASSP*, 2018, pp. 5524–5528.

[25] Youngmoon Jung, Younggwan Kim, Yeunju Choi, and Hoirin Kim, "Joint learning using denoising variational autoencoders for voice activity detection," *Proc. Interspeech*, 2018, pp. 1210–1214.

[26] Jitong Chen, Yuxuan Wang, and DeLiang Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, 2014.

[27] Narasimhan, Harikrishna and Agarwal, Shivani, "A structural SVM based approach for optimizing partial AUC," in *Proc. ICML*, 2013, pp. 516–524.