EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

AUC optimization for deep learning-based voice activity detection

Xiao-Lei Zhang[1,2*] and Menglong Xu[1,2]

Abstract

Voice activity detection (VAD) based on deep neural networks (DNN) have 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 the receiver operating characteristic (ROC) curve is a global evaluation metric for VAD at all possible decision thresholds. In this paper, we propose to maximize the area under the ROC curve (MaxAUC) by DNN, which can maximize the performance of VAD in terms of the entire ROC curve. However, the objective of the AUC maximization is nondifferentiable. To overcome this difficulty, we relax the nondifferentiable loss function to two differentiable approximation functions—sigmoid loss and hinge loss. To study the effectiveness of the proposed MaxAUC-DNN VAD, we take either a standard feedforward neural network or a bidirectional long short-term memory network as the DNN model with either the state-of-the-art multi-resolution cochleagram or short-term Fourier transform as the acoustic feature. We conducted noise-independent training to all comparison methods. Experimental results show that taking AUC as the optimization objective results in higher performance than the common objectives of the minimum squared error and minimum cross-entropy. The experimental conclusion is consistent across different DNN structures, acoustic features, noise scenarios, training sets, and languages.

Keywords: Area under the ROC curve, Deep learning-based voice activity detection, MaxAUC

1 Introduction

Voice activity detection (VAD) aims to detect target voices from background noises. It has demonstrated its effectiveness in many speech processing tasks, such as speech communications, speech recognition, speaker recognition, keyword spotting, and acoustic event detection. A major challenge of VAD is how to deal with low signal-to-noise ratio (SNR) environments. To address this issue, many methods in the early research stage of VAD focused on extracting the statistics of acoustic features. Typical features include energy in the time domain, zero-crossing rate, pitch detection [1], cepstral coefficients [2], and higher-order statistics [3]. Later on, the focus shifted to building statistical models from acoustic features. It fits signals to predefined models and learns the parameters of a prior probability distribution on-the-fly. A crucial problem of the statistical VAD is how to make an accurate model assumption for the real-world distribution of speech data. Existing model assumptions include Gaussian [4, 5], Laplacian [6], Gamma distributions [7], and their combinations [8]. A substantial difficulty that hinders the statistical VAD from adverse environments is that the model parameters are updated using limited local data, leaving a large amount of prior knowledge unexplored. Moreover, real-world data distributions may be too complicated to be modeled accurately by a predefined model assumption.

Machine learning-based VAD has received much attention recently. It regards VAD as a classification problem. It is flexible in incorporating prior knowledge, such as manually labeled data. It is also good at fusing

*Correspondence: xiaolei.zhang@nwpu.edu.cn

[2] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
Full list of author information is available at the end of the article

multiple acoustic features. Existing supervised models include linear discriminant analysis [9], support vector machines (SVM) [10], multi-modal methods [11], sparse coding [12, 13], and deep neural networks (DNN) [14–25]. Particularly, DNN has demonstrated a 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 otwo rgeinpects—acoustic features, e.g. [14, 18, 23, 24], and deep models, e.g. [16, 17, 19, 21, 25].

Recently, some new forms of deep learning based VAD have been studied as well. For example, VAD has been jointly studied with speech enhancement. Some work uses advanced speech enhancement models, like denoising variational autoencoders [25], convolutional-recurrent-network-based speech enhancement, residual-convolutional neural-network [26] and U-Net [27], to extract denoised features for VAD. In [28, 29], the works optimizes VAD and speech enhancement jointly in the framework of multitask learning. In [30], information about the speaker was exploited for the VAD, which makes VAD able to extract speaker-dependent speech segments.

Although deep learning-based VAD has been extensively studied, a fundamental missing research aspect is the training target. To our knowledge, the training targets of VAD are limited to either classification-loss-based minimum cross-entropy (MCE) [14] or regression-loss-based minimum mean square error (MMSE) [18]. 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, the 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, MCE and MMSE are both surrogate loss functions that do not optimize the ROC curve or AUC directly.

Motivated by the above issue, this paper proposes *MaxAUC-DNN* VAD, which optimizes the AUC directly by DNN. Specifically, the AUC optimization is originally formulated as an NP-hard integer programming problem. We first relaxes this nondifferentiable problem to a polynomial-time solvable convex optimization problem by two approximation functions—a sigmoid-loss function and a hinge-loss function, and then calculates the gradient of the relaxed AUC loss. Finally, we take the relaxed AUC loss as the training target of DNN, and back-propagate the gradient to the entire DNN. To benefit from both the relaxed AUC loss and other loss functions, we also propose a *hybrid loss* to optimize the loss functions jointly.

To demonstrate the strong generalization ability of the MaxAUC-DNN VAD systematically, we test the MaxAUC-DNN VAD with two conventional DNN models, which are a standard feedforward neural network and a bidirectional long short-term memory (BLSTM) network. We also adopt two kinds of acoustic features, which are the short-term Fourier transform (STFT) and multi-resolution cochleagram (MRCG). The above settings amount to six MaxAUC-DNN VADs. To evaluate their generalization ability to unknown test scenarios, we train them with large-scale noise-independent training, and evaluate their performance extensively in both noise-mismatching and language-mismatching test scenarios. We compared MaxAUC-DNN VAD with the other two common DNN-based VADs—MMSE-DNN VAD and MCE-DNN VAD, using the same types of the basic deep model and acoustic feature. Experimental results show that MaxAUC-DNN VAD yields significantly higher performance than the MMSE-DNN VAD and MCE-DNN VAD. The experimental conclusion is consistent across different DNN structures, acoustic features, noise scenarios, training sets, and languages.

This paper differs from our preliminary work [31] in several major aspects, which include the use of two relaxation functions in this paper (but not in [31]), several MaxAUC-DNN VADs with BLSTM in this paper (but not in [31]), noise-independent training in this paper (but not in [31]), different parameter settings, training and evaluation datasets, and experiments for evaluating the generalization ability of DNN models (but not in [31]). Particularly, the proposed MaxAUC in [31] is only a special case of the proposed $\text{MaxAUC}_{\text{hinge}}$ in this paper. Consequently, experimental results in this paper show that the relative improvement of MaxAUC over the comparison methods in the mismatched environments is at least as good as that in the matched environments, which has not be observed in [31].

The paper is organized as follows. In Section 3, we present the motivation and problem formulation of the proposed algorithm. In Section 4, we present the MaxAUC-DNN VAD algorithm. In Section 6, we present results with noise-independent training. Finally, we conclude in Section 7.

2 Notations

We first introduce some notations here. Regular small letters, e.g. s, t, and γ, indicate scalars. Bold small letters, e.g. $\mathbf{y}$ and $\boldsymbol{\alpha}$, indicate vectors. Bold capital letters, e.g. $\mathbf{P}$ and $\boldsymbol{\Phi}$, indicate matrices. Letters in calligraphic fonts, e.g.

$\mathcal{X}$, indicate sets. $\mathbf{0}$ ($\mathbf{1}$) is a vector with all entries being 1 (0). The operator ^{T} denotes the transpose. The operator $\circ$ denotes the element-wise product.

3 Motivation

Supervised learning based VAD aims to detect speech from nonspeech, which can be viewed as a typical binary classification problem. More precisely speaking, because nonspeech contains a lot of noise scenarios, VAD is essentially a problem of discriminating one class (i.e. noisy speech) to the rest classes (i.e. various kinds of noises). Here we formulate supervised learning-based VAD problem as follows.

Given a training corpus $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$ where $\mathbf{x}_i$ is a high-dimensional acoustic feature of the i-th frame and y_i is the ground-truth label of $\mathbf{x}_i$. If $\mathbf{x}_i$ is labeled as a speech frame, then $y_i = 1$; otherwise, $y_i = 0$. In the modal training stage, supervised VAD learns a mapping function $f_\alpha(\cdot)$ given $\mathcal{X}$, where α is the model parameter. In the test stage, VAD conducts:

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

where η is a decision threshold. $f_\alpha(\cdot)$ can be various supervised models. In this paper, we set $f_\alpha(\cdot)$ to a deep neural network. To restrict the output of $f_\alpha(\cdot)$ to a range of [0, 1], we set the output units of $f_\alpha(\cdot)$ to sigmoid functions or softmax units.

For DNN-based VAD, there are mainly two training objectives, i.e., MMSE and MCE. MMSE minimizes $\sum_{i=1}^{n} ||\mathbf{y}_i - f_\alpha(\mathbf{x}_i)||^2$. MCE minimizes $-\sum_{i=1}^{n}(y_i \log(f_\alpha(\mathbf{x}_i)) + (1-y_i)\log(1-f_\alpha(\mathbf{x}_i)))$. However, both of them were not carefully designed for VAD. In real-world applications, η is usually determined on-the-fly. For example, it is set close to zero in relatively clean environments, and far away from zero in noisy environments. η also varies in different applications. It is tuned for high speech detection rates in speech communications, and tuned for low false alarm rates in speaker recognition. Hence, the ROC curve and its corresponding AUC, which are unrelated to η, are used as the global evaluation metrics of VAD instead of classification accuracy. Because the mean squared error and cross entropy do not have direct connections to AUC, traditional DNN-based VADs yield suboptimal performance in terms of AUC.

4 MaxAUC-DNN-based VAD

In this section, we first present how to calculate AUC in Section 4.1, then present the optimization objective—MaxAUC in Section 4.2, and finally present the optimization algorithm of the MaxAUC-DNN VAD in Section 4.3.

4.1 AUC calculation

The ROC curve of $f_\alpha(\cdot)$ is defined as a curve of the speech detection rate P_{D} against false alarm rate P_{FA} at all possible decision thresholds η:

$$\begin{cases} P_{\mathrm{D}}(\eta) = P\big(f_\alpha(\mathbf{x}^+) > \eta\big) \\ P_{\mathrm{FA}}(\eta) = P(f_\alpha(\mathbf{x}^-) > \eta) \end{cases}, \quad \forall \eta \in \mathbb{R} \tag{2}$$

where $\mathbb{R}$ denotes the set of real numbers and $P(\cdot)$ denotes probability. The AUC is calculated by:

$$\mathrm{AUC} = \int_0^1 P_{\mathrm{D}}(\eta) dP_{\mathrm{FA}}(\eta) \tag{3}$$

$$= \int_{-\infty}^{+\infty} P_{\mathrm{D}}(\eta) p_{\mathrm{FA}}(\eta) d\eta \tag{4}$$

$$= P(f_\alpha(\mathbf{x}^+) > f_\alpha(\mathbf{x}^-)) \tag{5}$$

where $p_{\mathrm{FA}}(\eta)$ is the probability density function $f_\alpha(\mathbf{x}^-)$ of the random variable at point η.

When the number of the training samples M is finite, i.e., $M < +\infty$, the AUC of $f_\alpha(\mathbf{x})$ on the training data is calculated as follows. We denote the subsets of $\mathcal{X}$ containing the speech and nonspeech frames as $\mathcal{X}^+ = \{(\mathbf{x}_i^+, y_i^+)\}_{i=1}^{P}$ and $\mathcal{X}^- = \{(\mathbf{x}_j^-, y_j^-)\}_{j=1}^{N}$, respectively, with $M = P + N$. The AUC on the finite training set equals to the normalized Wilcoxon-Mann-Whitney statistic [32] of $f_\alpha(\mathbf{x})$ in the following form:

$$\mathrm{AUC} = \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} g(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-)) \tag{6}$$

where $g(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-))$ is defined as:

$$g(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-)) = \begin{cases} 1, & \text{if } f_\alpha(\mathbf{x}_i^+) > f_\alpha(\mathbf{x}_j^-) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Note that if we merge $\{f_\alpha(\mathbf{x}_i^+)\}_{i=1}^{P}$ and $\{f_\alpha(\mathbf{x}_j^-)\}_{j=1}^{N}$, and sort the merged data in an ascending order, (6) can be calculated efficiently by:

$$\mathrm{AUC} = \frac{1}{PN} \left(\sum_{i=1}^{P} r_i - \frac{P(P+1)}{2} \right) \tag{8}$$

where $r_i \in \{1, 2, \ldots, N\}$, $i = 1, \ldots, P$ is the ranking list of the scores $f_\alpha(\mathbf{x}_i^+)$ in the merged data. We try to maximize (6) for MaxAUC-DNN VAD, and use (8) as the calculation method of AUC in the evaluation stage (Fig. 1).

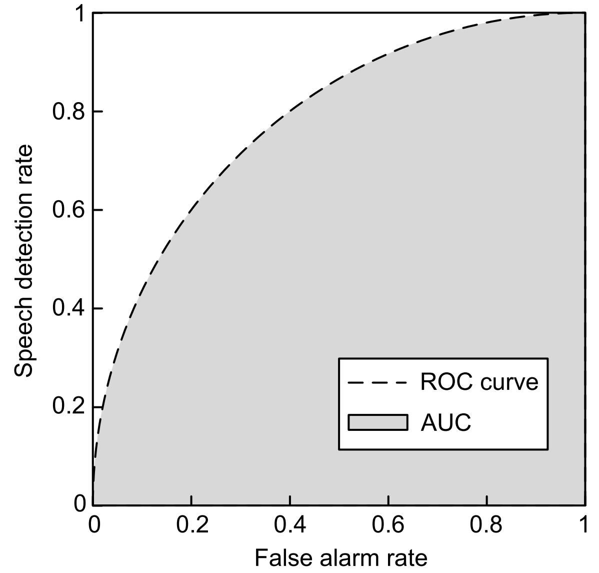

Fig. 1 Illustration of ROC curve and AUC

4.2 Objective formulation: MaxAUC

The ideal objective of maximizing AUC is to maximize (6). However, $g(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-))$ in (6) is nondifferentiable. To overcome this problem, we have to replace it with a differentiable approximation function. This paper considers the sigmoid function:

$$g_{\text{sigm}}(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-)) = \frac{1}{1 + e^{-\beta(f_\alpha(\mathbf{x}_i^+) - f_\alpha(\mathbf{x}_j^-))}} \tag{9}$$

as an approximation function, which results in the following *optimization objective* of the MaxAUC-DNN VAD:

$$\begin{aligned} \text{MaxAUC}_{\text{sigm}} &= \max_\alpha \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} g_{\text{sigm}}(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-)) \\ &= 1 - \min_\alpha \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} g_{\text{sigm}}(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-)) \end{aligned} \tag{10}$$

where $\beta > 0$ is a free-parameter. When $\beta < 1$, (9) is too smooth to approximate to (7). The larger β is, the better function (9) approximates to (7). However, when β is too large, the gradient of (10) will encounter numerical problems.

Another approximation function is the p-order hinge-loss function:

$$g_{\text{hinge}}(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-)) = \begin{cases} \left(-(f_\alpha(\mathbf{x}_i^+) - f_\alpha(\mathbf{x}_j^-) - \gamma)\right)^p, & \text{if } f_\alpha(\mathbf{x}_i^+) - f_\alpha(\mathbf{x}_j^-) < \gamma \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $0 < \gamma \leq 1$ is a predefined discriminative margin indicating that, if $f_\alpha(\mathbf{x}_i^+) < \gamma + f_\alpha(\mathbf{x}_j^-)$, the speech-nonspeech pair $(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-))$ is regarded as a wrong pair produced by $f_\alpha(\cdot)$, and $p > 1$ is a predefined parameter that enforces different loss to different wrong pairs according to their distances from the margin. The optimization objective of the MaxAUC-DNN VAD with (11) is:

$$\text{MaxAUC}_{\text{hinge}} = \min_\alpha \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} g_{\text{hinge}}(f_\alpha(\mathbf{x}_i^+), f_\alpha(\mathbf{x}_j^-)) \tag{12}$$

Note that the optimization objective in our conference version [31] is a special case of (12) with $p = 1$.

4.3 Optimization algorithm

In this paper, we employ the mini-batch stochastic gradient descent algorithm to solve (10) and (12). 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.

We can easily derive the gradient of (10) as:

$$\begin{aligned} \Delta_{\text{sigm}} = &\frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} \left(\beta \left(1 + e^{-\beta\left(f_\alpha(\mathbf{x}_i^+) - f_\alpha(\mathbf{x}_j^-)\right)} \right) \right. \\ &\left. e^{-\beta\left(f_\alpha(\mathbf{x}_i^+) - f_\alpha(\mathbf{x}_j^-)\right)} \left(\Delta f_\alpha(\mathbf{x}_i^+) - \Delta f_\alpha(\mathbf{x}_j^-) \right) \right) \end{aligned} \tag{13}$$

and the gradient of (12) as:

$$\Delta_{\text{hinge}} = \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} -p\Pi(i,j)\left(\Delta f_\alpha(\mathbf{x}_i^+) - \Delta f_\alpha(\mathbf{x}_j^-)\right) \tag{14}$$

where $\Delta f_\alpha(\mathbf{x})$ denotes the gradient of $f_\alpha(\mathbf{x})$ at the output layer, and $\Pi(i,j)$ is defined as:

$$\Pi(i,j) = \begin{cases} 1, & \text{if } f_\alpha(\mathbf{x}_i^+) - f_\alpha(\mathbf{x}_j^-) < \gamma \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

5 Hybrid loss

Although the original purpose of the proposed method is to optimize the evaluation metric of VAD directly, MaxAUC, which takes hinge loss or sigmoid loss to relax the 0–1 loss, is actually a surrogate function of the

AUC maximization. Therefore, there is no guarantee that MaxAUC will outperform other loss functions in all cases.

To combine the advantage of multiple loss functions, here we propose a *hybrid loss*:

$$\begin{aligned} &\min \sum_{i}^{C} \lambda_i \ell_i \\ &\text{subject to:} \quad \sum_{i}^{C} \lambda_i = 1, \\ &\qquad\qquad 0 \leq \lambda_i \leq 1, \quad \forall i = 1, \ldots, C \end{aligned} \tag{16}$$

where ℓ_i is a base loss function that can be AUC, cross entropy, squared error, etc., C is the number of the base loss functions in the hybrid loss, and $\{\lambda_i\}_{i=1}^{C}$ are learnable parameters that balance the loss functions. Note that $\{\lambda_i\}_{i=1}^{C}$ are jointly optimized with the parameters of the deep model by backpropagation. In this paper, we jointly optimize MaxAUC$_{\text{hinge}}$ and MCE as a special case of the hybrid loss.

6 Experiments

In this section, we first present the datasets and experimental settings in Sections 6.1 and 6.2, respectively, then present the main results in Section 6.3, and finally discuss the effects of the hyperparameters of the MaxAUC-DNN VAD on performance in Section 6.4.

6.1 Datasets

We used the LibriSpeech ASR database,[1] CHiME-4 challenge,[2] and THCHS-30[3] corpora as the source of clean speech. We used a large-scale sound effect library[4] and the NOISEX-92 database as the source of additive noise. All audio files were sampled at 16 KHz. The LibriSpeech ASR corpus is a large-scale corpus of 1000 h of read English speech. The single-channel clean speech data of CHiME-4, named the "tr05_org" subset, are a read Engish speech corpus based on the original WSJ0 training data, which contains 7138 utterances. THCHS-30 is an open Chinese speech database consisting of 35 h of clean speech signals. The sound effect library contains over 20,000 sound effects. NOISEX-92 is a widely used noise database containing 9 noise scenarios, each of which is about 5 minutes long.

6.1.1 Construction of the training and test sets

We constructed a noisy training set by mixing the "train-clean-100" subset of LibriSpeech ASR with the sound effect library at a SNR range from −10 to 20 dB, which generated over 200 h noisy English speech. We also constructed a development set by adding the babble and factory noise in NOISEX-92 to part of the clean speech of THCHS-30 at −5 dB for the hyperparameter selection problem. We denote the training data as the "Noisy-LibriSpeech."

We constructed two noisy test sets, one in English and the other in Chinese, by mixing the "tr05-org" subset of CHiME-4 and THCHS-30 with all 9 noise scenarios of NOISEX-92 at SNR levels of $\{-10, -5, 0, 5, 10, 15, 20\}$ dB, respectively. The three datasets do not have sample-level ground-truth labels. To address this issue, we applied Sohn VAD to the clean speech of the data sets, and used the prediction of the clean speech as the ground-truth labels, which has been proven to be reliable in [33]. We denote the English test data as "Noisy-CHiME-4" and the Chinese test data as "Noisy-THCHS-30."

All DNN models were trained with the Noisy-LibriSpeech dataset unless otherwise stated. All evaluations were conducted in mismatching conditions, including the mismatches of noise types, SNR levels, and languages.

Note that, to save the space of the paper, we only report the results in the babble, factory, and volvo noise scenarios of NOISEX-92, leaving the results in all 9 noise scenarios listed in the Supplementary material.

6.2 Experimental settings

To verify the effectiveness of the proposed algorithm with different acoustic features and different models, we took STFT and MRCG features respectively as its input. We set the frame length to 30 and 20 ms for the STFT and MRCG features, respectively, and set the frame shift to 10 ms for both features. The hyperparameter β of the MaxAUC$_{\text{sigm}}$-DNN VAD was set to 25 when MRCG was used, and set to 45 when STFT was adopted. The hyperparameters γ and p of the MaxAUC$_{\text{hinge}}$ were set to 0.2 and 1, respectively. We took feedfoward neural network and BLSTM as two basic deep models. Their hyperparameter settings are as follows.

For the feedfoward neural network, it contains two hidden layers. The number of the hidden units per hidden layer was set to 256. The activation functions of the hidden units and output units were set to the rectified linear units and sigmoid functions, respectively. The dropout rate was set to 0.2. We used stochastic gradient descent as the optimizer with an initial learning rate of 0.01 and a decay coefficient of 0.05. The number of training epochs was set to 30. The momentum of the first 3 epochs was

[1] http://www.openslr.org/12/

[2] http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/

[3] https://www.openslr.org/18/

[4] http://www.sound-ideas.com/sound-effects/series-6000-combo-sound-effects.html

Table 1 Computational resources (in terms of number of parameters) required by the proposed approach and the baselines

	STFT	MRCG
DNN	0.25M	0.36M
BLSTM	0.63M	0.85M

Table 2 AUC results of the comparison VADs with the feedforward neural network and STFT acoustic feature on the English Noisy-CHiME-4 test dataset. We use the names of the objectives of the VADs to represent the VADs for short

Noise type	SNR	MCE	MMSE	$\text{MaxAUC}_{\text{sigm}}$	$\text{MaxAUC}_{\text{hinge}}$
Babble	− 10 dB	0.5319	0.5381	**0.5631**	0.5561
	− 5 dB	0.6006	0.6097	**0.6450**	0.6359
	0 dB	0.7092	0.7109	**0.7431**	0.7363
	5 dB	0.8036	0.8046	**0.8226**	0.8187
	10 dB	0.8652	0.8673	**0.8762**	0.8726
	15 dB	0.9028	0.9021	**0.9071**	0.9044
	20 dB	0.9208	0.9191	**0.9214**	0.9204
Factory	− 10 dB	0.6321	0.6303	0.6399	**0.6400**
	− 5 dB	0.7275	0.7260	0.7314	**0.7341**
	0 dB	0.8078	0.8072	0.8071	**0.8114**
	5 dB	0.8616	0.8611	0.8587	**0.8628**
	10 dB	0.8967	0.8955	0.8936	**0.8968**
	15 dB	**0.9162**	0.9139	0.9132	0.9151
	20 dB	**0.9263**	0.9235	0.9236	0.9247
Volvo	− 10 dB	0.8910	0.8793	**0.9002**	0.8968
	− 5 dB	0.9109	0.9042	**0.9136**	0.9132
	0 dB	0.9217	0.9177	0.9214	**0.9218**
	5 dB	**0.9276**	0.9242	0.9260	0.9260
	10 dB	**0.9311**	0.9275	0.9285	0.9280
	15 dB	**0.9329**	0.9292	0.9299	0.9292
	20 dB	**0.9338**	0.9302	0.9306	0.9301

set to 0.5, and the momentum of other epochs was set to 0.9. The batch size was set to 4096. A contextual window was used to expand each input frame to its context along the time axis. The window size was set to 3.

For the BLSTM model, it comprises of a fully connected layer with the rectified linear units as the activation functions, followed by a BLSTM layer with the tangent functions as the activation functions, and an output layer with the sigmoid functions. The numbers of the hidden units for the fully connected layer and BLSTM layer were set to 512 and 256, respectively. The dropout rate was set to 0.2. The batch size was set to 4096. The network was randomly initialized, and optimized by the stochastic gradient descent with the Adam optimizer. The number of training epochs was set to 30. The learning rate was initialized to 1 and decreased with a decay coefficient of 0.05. The BLSTM model adopted the same contextual window as the DNN model for expanding the input. Note that for specific features and networks, the computational resources required by the proposed AUC loss-based method and the baselines are the same. We list the computational resources required by these approaches in Table 1.

We compared the MaxAUC-DNN VAD with the MMSE-DNN VAD and MCE-DNN VAD. To do the comparison fairly, we compared the loss functions, i.e., MMSE and MCE, with the two variants of MaxAUC only, leaving the other parts of the comparison methods the same. All experiments were conducted in a non-reverberant environment. We adopted the ROC curve and AUC as the evaluation metrics.

6.3 Main results

We first compared the VAD methods with the feedforward neural network and STFT feature on the Noisy-CHiME-4 dataset. From the comparison results in Table 2, we see that, when the SNR levels are below 10 dB, the $\text{MaxAUC}_{\text{hinge}}$-DNN VAD outperforms the MCE-DNN VAD and the MMSE-DNN VAD by relatively 2.21% and 6.90%, respectively, meanwhile, the $\text{MaxAUC}_{\text{sigm}}$-DNN VAD outperforms the two competitive VADs by relatively 1.36% and 6.07%, respectively. The MaxAUC-DNN VADs perform similarly with the MCE-DNN VAD in the other scenarios, both of which outperform the MMSE-DNN VAD significantly.

6.3.1 Robustness to different DNN models

To evaluate how different types of DNN models affect the performance, we replaced the feedforward neural network by BLSTM. Table 3 lists the comparison results on Noisy-CHiME-4. From the table, we see that the experimental phenomenon is consistent with that in Table 2. Moreover, the $\text{MaxAUC}_{\text{hinge}}$-DNN VAD outperforms the MCE-DNN VAD by relatively 5.66% when the SNR levels are greater than or equal to 10 dB, which is an interesting phenomenon unobserved in Table 2.

6.3.2 Robustness to mismatched test languages

To further evaluate the generalization ability of the proposed method on different languages, we compared the VAD methods that adopted the BLSTM model and STFT feature on the Chinese Noisy-THCHS-30 test corpus. Table 4 lists the comparison results. From the table, we see that the experimental phenomenon is similar to that in Table 3, though all methods suffer some performance degradation due to the mismatch between the training and test languages.

Table 3 AUC results of the comparison VADs with the BLSTM model and STFT acoustic feature on the English Noisy-CHiME-4 test dataset

Noise type	SNR	MCE	MMSE	MaxAUC$_{sigm}$	MaxAUC$_{hinge}$
Babble	− 10 dB	0.5163	0.5270	**0.5428**	0.5383
	− 5 dB	0.5636	0.5761	**0.6010**	0.5940
	0 dB	0.6491	0.6567	**0.6867**	0.6787
	5 dB	0.7466	0.7499	**0.7716**	0.7641
	10 dB	0.8227	0.8241	**0.8362**	0.8283
	15 dB	0.8703	0.8696	**0.8765**	0.8699
	20 dB	0.8977	0.8974	**0.9003**	0.8978
Factory	− 10 dB	0.6024	0.6031	0.6066	**0.6089**
	− 5 dB	0.6864	0.6830	0.6898	**0.6923**
	0 dB	0.7659	0.7610	0.7653	**0.7685**
	5 dB	**0.8243**	0.8196	0.8204	0.8240
	10 dB	**0.8617**	0.8580	0.8573	0.8599
	15 dB	**0.8862**	0.8826	0.8811	0.8824
	20 dB	**0.9033**	0.8995	0.8977	0.8984
Volvo	− 10 dB	0.8562	0.8432	**0.8752**	0.8702
	− 5 dB	0.8871	0.8780	**0.8996**	0.8961
	0 dB	0.9062	0.9010	**0.9137**	0.9107
	5 dB	0.9166	0.9136	**0.9223**	0.9182
	10 dB	0.9220	0.9194	**0.9261**	0.9214
	15 dB	0.9248	0.9227	**0.9277**	0.9229
	20 dB	0.9264	0.9248	**0.9287**	0.9241

Table 4 AUC results of the comparison VADs with the BLSTM model and STFT acoustic feature on the Chinese Noisy-THCHS-30 test dataset

Noise type	SNR	MCE	MMSE	MaxAUC$_{sigm}$	MaxAUC$_{hinge}$
Babble	− 10 dB	0.5226	0.5268	**0.5315**	0.5308
	− 5 dB	0.5826	0.5918	**0.5944**	**0.5944**
	0 dB	0.6800	0.6901	0.6897	**0.6943**
	5 dB	0.7787	0.7834	0.7853	**0.7915**
	10 dB	0.8484	0.8481	0.8520	**0.8563**
	15 dB	0.8870	0.8864	0.8893	**0.8928**
	20 dB	0.9096	0.9099	0.9116	**0.9140**
Factory	− 10 dB	0.6247	0.6238	0.6300	**0.6420**
	− 5 dB	0.7177	0.7168	0.7216	**0.7314**
	0 dB	0.7962	0.7948	0.7976	**0.8030**
	5 dB	0.8483	0.8471	0.8483	**0.8511**
	10 dB	0.8805	0.8798	0.8810	**0.8828**
	15 dB	0.9031	0.9020	0.9033	**0.9053**
	20 dB	0.9198	0.9175	0.9188	**0.9223**
Volvo	− 10 dB	**0.8851**	0.8753	0.8848	0.8845
	− 5 dB	0.9077	0.8984	0.9095	**0.9101**
	0 dB	0.9208	0.9145	0.9234	**0.9252**
	5 dB	0.9292	0.9257	0.9313	**0.9332**
	10 dB	0.9352	0.9328	0.9353	**0.9382**
	15 dB	0.9384	0.9361	0.9368	**0.9412**
	20 dB	0.9398	0.9374	0.9375	**0.9425**

6.3.3 Robustness to different acoustic features

To study how different acoustic features affect the performance, we replaced STFT with MRCG as the acoustic feature for all comparison methods, and conducted the experiment on the Chinese Noisy-THCHS-30 test corpus. Table 5 lists the comparison results. More results with the STFT feature can be found in the Supplementary materials. From the table, we see that the proposed methods outperform the competitive VADs in most cases at low SNR levels. Comparing Table 5 with Table 4, we also see that the performance of the comparison methods with MRCG is better than that that with STFT.

6.3.4 Robustness to different training sets

To further investigate how different training sets affect the effectiveness of the proposed methods, we conducted a comparison on the Noisy-CHiME-4 test dataset, with the Chinese Noisy-THCHS-30 dataset as the clean speech source of the training set. Note that the generation process of the noisy Noisy-THCHS-30 training set, which was mixed from the training subset of THCHS-30 and the large-scale sound effect library at a SNR range of −10 to 20 dB, was similar to the generation process of the Noisy-LibriSpeech training set. Table 6 lists the comparison results, which again demonstrate the superiority of the proposed methods.

6.3.5 Summary and analysis of the comparison results

Figure 2 summarizes the relative improvement of the MaxAUC VADs over the competitive VADs when the STFT feature is used as the acoustic feature, where we summarize not only the results in Tables 2, 3, and 4 but also the result in the Supplementary materials. From the figure, we see that the relative improvement reaches the maximum around 0 dB. The MaxAUC$_{sigm}$-DNN VAD performs better than the MaxAUC$_{hinge}$-DNN VAD when the basic deep model is the feedforward neural network. Although the curves in Figs. 2c and d tend to decrease along with the increase of the SNR level, the relative improvement of the MaxAUC$_{hinge}$-DNN VAD drops slower than that of the MaxAUC$_{sigm}$-DNN VAD. Importantly, we find that the mismatch of the test languages does not affect the relative improvement of the MaxAUC-DNN VADs over their comparison methods.

Figure 3 shows the relative AUC improvement when the MRCG feature is used as the acoustic feature from the figure. From the figure, we see that the MaxAUC-DNN VADs outperform the competitive VAD methods in most test scenarios, except the scenario in Fig. 3a

Table 5 AUC results of the comparison VADs with the BLSTM model and MRCG acoustic feature on the Chinese Noisy-THCHS-30 test dataset

Noise type	SNR	MCE	MMSE	MaxAUC$_{\text{sigm}}$	MaxAUC$_{\text{hinge}}$
Babble	− 10 dB	0.6276	0.6209	0.6324	**0.6370**
	− 5 dB	0.7238	0.7073	0.7278	**0.7362**
	0 dB	0.8165	0.7947	0.8184	**0.8269**
	5 dB	0.8763	0.8586	0.8774	**0.8826**
	10 dB	0.9061	0.8974	0.9080	**0.9110**
	15 dB	0.9223	0.9197	0.9246	**0.9280**
	20 dB	0.9358	0.9345	0.9369	**0.9414**
Factory	− 10 dB	0.7542	0.7479	0.7618	**0.7658**
	− 5 dB	0.8355	0.8284	0.8414	**0.8457**
	0 dB	0.8813	0.8761	0.8846	**0.8873**
	5 dB	0.9053	0.9017	0.9075	**0.9089**
	10 dB	0.9201	0.9176	0.9219	**0.9231**
	15 dB	0.9314	0.9296	0.9330	**0.9349**
	20 dB	0.9410	0.9390	0.9421	**0.9445**
Volvo	− 10 dB	0.9359	0.9352	0.9373	**0.9376**
	− 5 dB	0.9472	0.9459	0.9473	**0.9483**
	0 dB	0.9549	0.9529	0.9543	**0.9556**
	5 dB	0.9594	0.9570	0.9584	**0.9599**
	10 dB	0.9615	0.9588	0.9604	**0.9621**
	15 dB	0.9620	0.9590	0.9607	**0.9628**
	20 dB	0.9616	0.9584	0.9601	**0.9628**

Table 6 AUC results of the comparison VADs with the BLSTM model and STFT acoustic feature on the Noisy-CHiME-4 test dataset, when the Chinese Noisy-THCHS-30 dataset was used as the training set

Noise type	SNR	MCE	MMSE	MaxAUC$_{\text{sigm}}$	MaxAUC$_{\text{hinge}}$
Babble	− 10 dB	0.5752	0.5664	**0.5833**	0.5799
	− 5 dB	0.6442	0.6391	**0.6588**	0.6565
	0 dB	0.7272	0.7222	**0.7462**	0.7441
	5 dB	0.7900	0.7867	**0.8076**	0.8057
	10 dB	0.8246	0.8289	0.8387	**0.8390**
	15 dB	0.8420	0.8430	**0.8529**	0.8467
	20 dB	0.8487	0.8624	0.8579	**0.8628**
Factory	− 10 dB	0.5992	0.5938	**0.6115**	0.6011
	− 5 dB	0.6743	0.6694	**0.6897**	0.6822
	0 dB	0.7340	0.7294	**0.7536**	0.7474
	5 dB	0.7791	0.7769	**0.7994**	0.7929
	10 dB	0.8142	0.8166	**0.8299**	0.8285
	15 dB	0.8373	0.8458	0.8485	**0.8503**
	20 dB	0.8474	0.8603	0.8569	**0.8646**
Volvo	− 10 dB	0.7571	0.7551	0.7790	**0.7858**
	− 5 dB	0.7933	0.7905	0.8195	**0.8270**
	0 dB	0.8244	0.8229	0.8443	**0.8534**
	5 dB	0.8350	0.8343	0.8530	**0.8602**
	10 dB	0.8374	0.8295	0.8560	**0.8602**
	15 dB	0.8423	0.8518	**0.8600**	0.8589
	20 dB	0.8476	0.8595	**0.8645**	0.8593

at 20 dB. We also find interestingly that, although the relative improvement of the two MaxAUC-DNN VADs over the competitive methods are similar in the low SNR levels, the relative improvement of the MaxAUC$_{\text{hinge}}$ VAD over the comparison methods tends to be enlarged when the SNR is increased in Fig. 3b, c, and d, while the relative improvement of the MaxAUC$_{\text{sigm}}$ VAD over the comparison methods is reduced on the contrary. Moreover, we find that the relative improvement of the MaxAUC-DNN VADs over their comparison methods on the mismatched test language is higher than that on the matched test language.

Comparing Figs. 2 and 3, we summarize that the MaxAUC$_{\text{hinge}}$-DNN VAD has a slightly stronger generalization ability than the MaxAUC$_{\text{sigm}}$-DNN VAD in most cases. The better the basic deep model and acoustic feature are, the larger the superiority of the MaxAUC$_{\text{hinge}}$-DNN VAD achieves.

At last, we exemplify the ROC curves of the comparison methods with the BLSTM model and MRCG feature on the Chinese Noisy-THCHS-30 at −5 dB in Fig. 4. From the figure, it is clear that both of the proposed VADs outperform the competitive VADs, and the MaxAUC$_{\text{hinge}}$-DNN VAD performs the best in most cases except the machinegun scenario. The above phenomena are observed in most other evaluations too.

6.4 Effects of hyperparameters on performance

In this subsection, we evaluated the hyperparameters of the MaxAUC-DNN VAD with the BLSTM model and MRCG feature on the babble and factory noise scenarios of the Chinese development dataset at −5 dB, and applied the optimal hyperparameters to all other test scenarios in this paper. The hyperparameter β of the MaxAUC$_{\text{sigm}}$-DNN VAD was selected from a range of [2, 50]. Figure 5 lists the experimental result. From the figure, we observe that β behaves robustly in a wide range of [20, 50]. The hyperparameters γ and p of the MaxAUC$_{\text{hinge}}$-DNN VAD were selected from [0.1:0.1:0.9], and [1:1:9], respectively, where the symbol [a:b:c] denotes a serial numbers starting from a and ending at c with a step size of c. We searched (γ, p) jointly in a mesh grid. Figure 6 lists the experimental result. From the figure, it seems that the two hyperparameters have a strong correlation. If one of the hyperparameters was enlarged, and if the other one was enlarged accordingly, then the performance is stable across the two evaluation scenarios. The best performance appears around $\gamma = 0.2$ and $p = 1$.

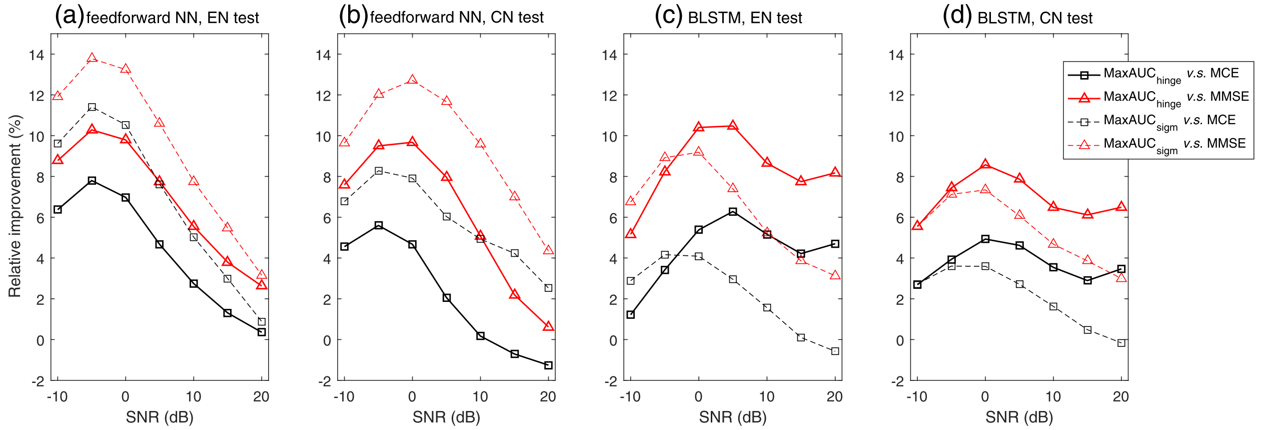

Fig. 2 Relative AUC improvement of the proposed methods over the competitive methods, when STFT is used as the acoustic feature. **a** Feedforward neural network is used as the basic deep model; the evaluation is conducted on the English Noisy-CHiME-4 dataset. **b** Feedforward neural network is used; the evaluation is conducted on the Chinese Noisy-THCHS-30 dataset. **c** BLSTM is used; the evaluation is conducted on the English Noisy-CHiME-4 dataset. **d** BLSTM is used; the evaluation is conducted on the Chinese Noisy-THCHS-30 dataset

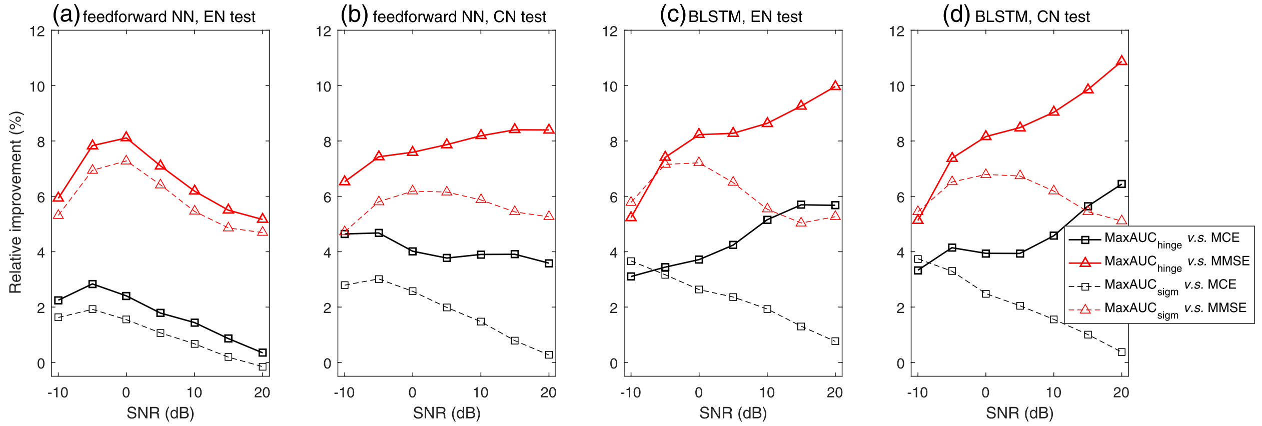

Fig. 3 Relative AUC improvement of the proposed methods over the competitive methods, when MRCG is used as the acoustic feature. The terms "EN" and "CH" are short for English and Chinese respectively. The term "NN" is short for neural networks. **a** Feedforward neural network is used as the basic deep model; the evaluation is conducted on the English Noisy-CHiME-4 dataset. **b** Feedforward neural network is used; the evaluation is conducted on the Chinese Noisy-THCHS-30 dataset. **c** BLSTM is used; the evaluation is conducted on the English Noisy-CHiME-4 dataset. **d** BLSTM is used; the evaluation is conducted on the Chinese Noisy-THCHS-30 dataset

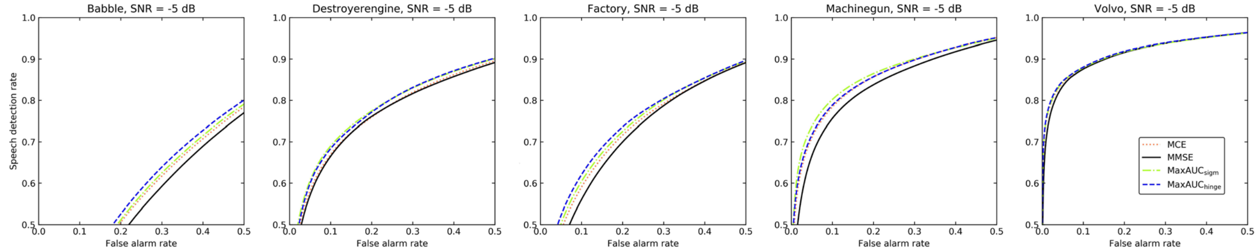

Fig. 4 ROC curves of the comparison methods with the BLSTM model and MRCG feature on the Chinese Noisy-THCHS-30 test dataset at −5 dB

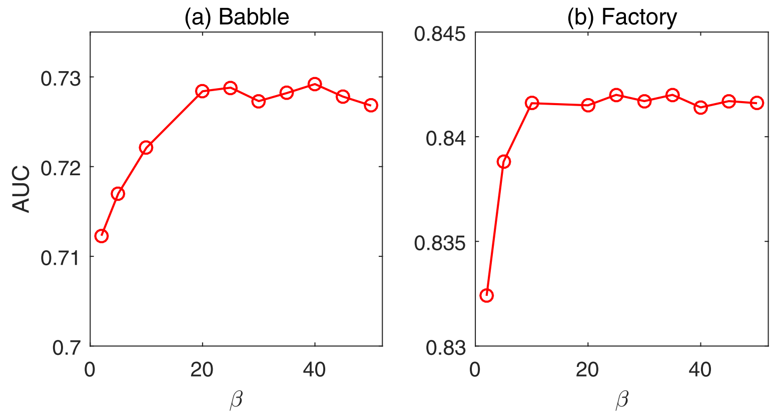

Fig. 5 Effect of the hyperparameter β of MaxAUC$_{\text{sigm}}$ VAD on the Chinese development dataset at −5 dB

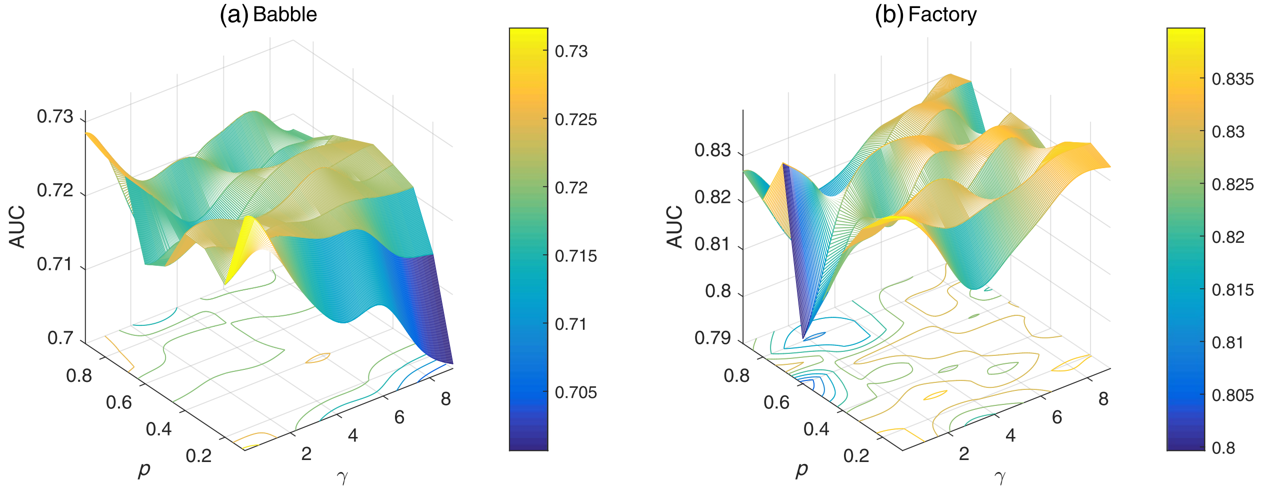

Fig. 6 Effect of the hyperparameters γ and p of MaxAUC$_{\text{hinge}}$ VAD on the Chinese development dataset at −5 dB

6.5 Effects of the hybrid loss on performance

In this subsection, we evaluated the hybrid loss of MaxAUC and MCE with the BLSTM model and STFT feature in the challenging babble and factory noise scenarios of the Noisy-CHiME-4 test dataset, where the Chinese Noisy-THCHS-30 dataset was used as the training set. The weights λ of MaxAUC and MCE, which were obtained automatically by optimizing (16) on the Chinese Noisy-THCHS-30 training data, are 0.7764 and 0.2236, respectively. This manifests that MaxAUC is a more effective training loss than MCE in generating a good local minimum of the BLSTM model. Table 7 lists the comparison results. From the table, we see that the hybrid-loss-based VAD outperforms the MaxAUC-DNN VAD and MCE-DNN VAD in the babble and volvo noise scenarios. However, it does not outperform the MaxAUC-DNN VAD in the factory noise scenario, which needs further investigation.

7 Conclusions

In this paper, we have proposed the MaxAUC-DNN VAD for improving the performance of the DNN-based VAD at any decision threshold. Specifically, we first relax the AUC calculation, which is an integer optimization problem, to a polynomial-time solvable problem by a differentiable function, then compute the gradient of the relaxed AUC loss with respect to the parameters of the output layer of DNN, and finally back-propagate the gradient to its hidden layers. We proposed two approximation functions—a sigmoid loss approximation and a hinge loss approximation. To integrate the advantage of the proposed loss with existing VAD loss functions, we propose a hybrid loss framework that jointly optimizes the loss functions. We evaluated the effectiveness of the MaxAUC-DNN VAD in a wide range of test scenarios from the respects of different

Table 7 AUC results of the comparison VADs with the MaxAUC $_{\text{hinge}}$, MCE and hybrid losses on the Noisy-CHiME-4 test dataset, where the Chinese Noisy-THCHS-30 dataset was used as the training set

Noise type	SNR	MCE	MaxAUC$_{\text{hinge}}$	Hybrid loss
Babble	− 10 dB	0.5752	0.5799	**0.5808**
	− 5 dB	0.6442	0.6565	**0.6579**
	0 dB	0.7272	**0.7441**	0.7422
	5 dB	0.7900	**0.8057**	0.8030
	10 dB	0.8246	0.8390	**0.8403**
	15 dB	0.8420	0.8467	**0.8616**
	20 dB	0.8487	0.8628	**0.8715**
Factory	− 10 dB	0.5992	0.6011	**0.6041**
	− 5 dB	0.6743	**0.6822**	0.6799
	0 dB	0.7340	**0.7474**	0.7350
	5 dB	0.7791	**0.7929**	0.7806
	10 dB	0.8142	**0.8285**	0.8175
	15 dB	0.8373	**0.8503**	0.8488
	20 dB	0.8474	0.8646	**0.8663**
Volvo	− 10 dB	0.7571	0.7858	**0.7862**
	− 5 dB	0.7933	0.8270	**0.8305**
	0 dB	0.8244	0.8534	**0.8572**
	5 dB	0.8350	**0.8602**	0.8595
	10 dB	0.8374	0.8602	**0.8613**
	15 dB	0.8423	0.8589	**0.8620**
	20 dB	0.8476	0.8593	**0.8638**

DNN models, acoustic features, training sets, and the noise mismatching and language mismatching scenarios. Empirical results show that the MaxAUC-DNN VAD outperforms the MMSE-DNN VAD and MCE-DNN VAD in most test scenarios, and the relative improvement over the comparison methods tends to be enlarged when the training and test conditions are mismatched; it is also insensitive to the hyperparameter selection. Finally, the hybrid loss has shown its potential in outperforming its components.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13636-022-00260-9.

Additional file 1: Supplementary materials.

Acknowledgements
The authors would like to thank the editors and anonymous reviewers for their volunteer endeavor on this paper which greatly improved the quality of the paper.

Authors' contributions

Model development: XZ and MX. Experimental testing: XZ and MX. Writing paper: XZ and MX. The authors read and approved the final manuscript.

Authors' information

Not applicable

Funding
This work was supported in part by National Science Foundation of China under Grant No. 62176211, in part by Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality under grant No. JSGG20210802152546026 and JCYJ20210324143006016.

Availability of data and materials

Not applicable.

Declarations

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen, China. [2]School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China.

Received: 26 September 2021 Accepted: 8 October 2022
Published online: 22 October 2022

References

1. R. Tucker, Tucker, Voice activity detection using a periodicity measure. IEE Proc. I (Commun. Speech Vis.). **139**(4), 377–380 (1992)
2. J.-C. Junqua, H. Wakita, in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference On*. A comparative study of cepstral lifters and distance measures for all pole models of speech in noise (IEEE, 1989), pp. 476–479
3. E. Nemer, R. Goubran, S. Mahmoud, Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Trans. Speech Audio Process. **9**(3), 217–231 (2001)
4. J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection. IEEE Signal Process. Lett. **6**(1), 1–3 (1999)
5. J. Ramírez, J.C. Segura, C. Benítez, L. García, A. Rubio, Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Process. Lett. **12**(10), 689–692 (2005)
6. J.-H. Chang, N.S. Kim, Voice activity detection based on complex laplacian model. Electron. Lett. **39**(7), 632–634 (2003)
7. J.W. Shin, J.-H. Chang, N.S. Kim, Statistical modeling of speech signals based on generalized gamma distribution. IEEE Signal Process. Lett. **12**(3), 258–261 (2005)
8. J.H. Chang, N.S. Kim, S.K. Mitra, Voice activity detection based on multiple statistical models. IEEE Trans. Signal Process. **54**(6), 1965–1976 (2006)
9. J. Padrell, D. Macho, C. Nadeu, in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference On*. Robust speech activity detection using lda applied to ff parameters, vol. 1 (IEEE, 2005), p. 557
10. J. Wu, X.L. Zhang, Efficient multiple kernel support vector machine based voice activity detection. IEEE Signal Process. Lett. **18**(8), 466–499 (2011)
11. D. Dov, R. Talmon, I. Cohen, Multimodal kernel method for activity detection of sound sources. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(6), 1322–1334 (2017)
12. P. Teng, Y. Jia, Voice activity detection via noise reducing using non-negative sparse coding. IEEE Signal Process. Lett. **20**(5), 475–478 (2013)
13. S.-W. Deng, J.-Q. Han, Statistical voice activity detection based on sparse representation over learned dictionary. Digit. Signal Process. **23**(4), 1228–1232 (2013)
14. X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection. IEEE Trans. Audio Speech Lang. Process. **21**(4), 697–710 (2013)

15. X.-L. Zhang, J. Wu, in *the 38th IEEE International Conference on Acoustic, Speech, and Signal Processing*. Denoising deep neural networks based voice activity detection (2013), pp. 853–857
16. T. Hughes, K. Mierle, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*. Recurrent neural networks for voice activity detection (2013). pp. 7378–7382
17. F. Eyben, F. Weninger, S. Squartini, B. Schuller, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies (IEEE, 2013) pp. 483–487
18. X.-L. Zhang, D. Wang, Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(2), 252–264 (2016)
19. I. Hwang, H.-M. Park, J.-H. Chang, Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection. Comput. Speech Lang. **38**, 1–12 (2016)
20. Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai, C.-H. Lee, In: *Sixteenth Annual Conference of the International Speech Communication Association*. A universal vad based on jointly trained deep neural networks (2015)
21. L. Wang, K. Phapatanaburi, Z. Go, S. Nakagawa, M. Iwahashi, J. Dang, in *Proceedings of ICME*. Limiting numerical precision of neural networks to achieve real-time voice activity detection (2018), pp. 1087–1092
22. Y. Tachioka, in *Proceedings of ICASSP*. Limiting numerical precision of neural networks to achieve real-time voice activity detection (2018), pp. 2236–2240
23. Y. Tachioka, in *Proceedings of ICASSP*. Dnn-based voice activity detection using auxiliary speech models in noisy environments (2018). pp. 5529–5533
24. W.A. Jassim, N. Harte, in *Proceedings of ICASSP*. Voice activity detection using neurograms (2018), pp. 5524–5528
25. Y. Jung, Y. Kim, Y. Choi, H. Kim, in *Interspeech*. Joint learning using denoising variational autoencoders for voice activity detection (2018), pp. 1210–1214
26. T. Xu, H. Zhang, X. Zhang, in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Joint training rescnn-based voice activity detection with speech enhancement (IEEE, 2019), pp. 1157–1162
27. G.W. Lee, H.K. Kim, Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection. Appl. Sci. **10**(9), 3230 (2020)
28. Y. Zhuang, S. Tong, M. Yin, Y. Qian, K. Yu, in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Multi-task joint-learning for robust voice activity detection (IEEE, 2016), pp. 1–5
29. X. Tan, X.-L. Zhang, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech enhancement aided end-to-end multi-task learning for voice activity detection (IEEE, 2021), pp. 6823–6827
30. Y. Chen, S. Wang, Y. Qian, K. Yu, End-to-end speaker-dependent voice activity detection. arXiv preprint arXiv:2009.09906 (2020)
31. Z.-C. Fan, Z. Bai, X.-L. Zhang, S. Rahardja, J. Chen, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Auc optimization for deep learning based voice activity detection (IEEE, 2019), pp. 6760–6764
32. H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 50–60 (1947)
33. X.-L. Zhang, D. Wang, Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Trans Audio Speech Lang. Process. **24**(2), 252–264 (2015)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.