ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746261

COMPLEX IRM-AWARE TRAINING FOR VOICE ACTIVITY DETECTION USING ATTENTION MODEL

Yifei Zhao[⋆] *Yazid Attabi*[⋆] *Benoit Champagne*[⋆] *Wei-Ping Zhu*[†]

[⋆]Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
[†]Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

ABSTRACT

Although many state-of-the-art approaches for improving the accuracy of Voice Activity Detection (VAD) have been proposed, their performance under adverse noise conditions with low Signal-to-Noise Ratio (SNR) remains limited. In this paper, we introduce a novel attention model-based deep neural network (DNN) architecture for VAD which takes advantage of complex Ideal Ratio Mask (cIRM). The proposed model, named AM-cIRM, consists of three sequential modules: extraction of cIRM features from the noisy speech using a DNN-based architecture; combination of cIRM with log-Mel spectrogram features along with temporal contextual extension; and VAD using an attention model that exploits the spectro-temporal information in the transformed features. Experimental results show that the proposed AM-cIRM achieves improved VAD performance when compared to state-of-the-art methods under different noise conditions.

Index Terms— Voice activity detection, deep neural network, attention mechanisms, complex Ideal Ratio Mask.

1. INTRODUCTION

Voice Activity Detection (VAD) refers to a family of methods that classify frames of audio signals into speech and non-speech. It serves as an important preprocessor for many speech-related applications including speaker identification, automatic speech recognition, and hearing aids [1, 2]. Early VAD methods were mainly based on average magnitude and power calculations in the time domain [3, 4], under the assumption that the power of speech is greater than the noise power. Other methods were subsequently developed that rely on the use of various features of speech signals, such as zero crossing rate [5], spectral or cepstral features [6, 7], higher order statistics [8] and pitch detection [9]. Several VAD methods have been developed based on the Likelihood Ratio Test (LRT) [10], assuming *a priori* knowledge of the speech signal and noise statistical distributions.

In recent years, Machine Learning (ML) techniques have demonstrated good classification results on VAD tasks. For instance, linear discriminant analysis [11], Support Vector Machines (SVM) [12], sparse coding [13], and especially Deep Neural Networks (DNNs) have shown superior performance over traditional (i.e., non ML-based) approaches. In particular, many studies reveal that the choice of acoustic feature plays an important role in DNN-based VAD. Inspired by [14], where auxiliary features (e.g., phoneme information) are used to improve speech enhancement performance, [15] shows that the performance of DNN-based VAD can be notably improved by using output features from two types of auxiliary speech models. Ref. [16] proposes a boosted DNN (bDNN) architecture by combining Multi-Resolution Stacking (MRS) and Multi-Resolution CochleaGram (MRCG) features. A combined VAD system is introduced in [17] which utilizes Wavenet-based network [18] for acoustic feature extraction and a deep residual network for video feature extraction. There is currently growing interest in the use of attention mechanisms for VAD. The Adaptive Context Attention Model (ACAM) [19] adopts an attention mechanism to exploit temporal information. However, its reinforcement loss function is sensitive to hyperparameters and subject to instability during training. By contrast, the Spectro-Temporal Attention-based Model (STAM) for VAD [20] improves performance by applying attention mechanisms to both contextual and spectral information.

To further improve the robustness of VAD in noisy environments, we propose a novel attention model-based DNN architecture which takes advantage of the complex Ideal Ratio Mask (cIRM). The proposed method, called AM-cIRM, consists of three sequential modules performing the following tasks: extraction of cIRM features from the noisy speech signal; combination of the cIRM with log-Mel spectrogram features, along with temporal contextual extension; and finally, VAD using a STAM-based model that exploits the spectro-temporal information contained in the transformed features. Experimental results in terms of the F1-score and detection cost function show that the proposed AM-cIRM method achieves improved VAD performance compared to the state-of-the-art methods under different noise types and SNR conditions.

2. BACKGROUND

In this section, we briefly describe the acoustic features used in our work and review the STAM approach [20].

2.1. Acoustic Features

The input noisy speech signal is modeled as $x[n] = s[n] + w[n]$, where $x[n]$, $s[n]$ and $w[n]$ denote the noisy speech, clean speech, and noise signals, respectively, while $n \in \mathbb{Z}$ is the discrete-time index. The Short-Time-Fourier-Transform (STFT) of $x[n]$ is represented by matrix $\mathbf{X}_{\mathrm{FT}} \in \mathbb{C}^{F\times T}$, where T is the number of frames and F is the number of frequency bins. The log-Mel spectrogram of $x[n]$ is represented by $\mathbf{X}_{\mathrm{MC}} \in \mathbb{R}^{T\times D}$, where D is the number of Mel coefficients. It is obtained by applying a bank of Mel-scaled

Support for this work was provided by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada, with industrial sponsor Microchip Technology (Ottawa, Canada).

triangular energy filters to each column of $|\mathbf{X}_{\mathrm{FT}}|$ and taking the logarithm of each filter output [21].

2.2. Spectro-Temporal Attention Model

The STAM [20] includes 4 modules, i.e.: spectral attention, pipe-net, temporal attention and post-net, as shown Fig. 1.

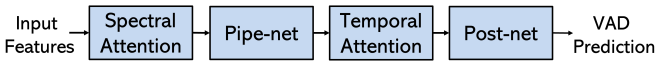

Fig. 1. Model architecture of STAM.

• **Spectral Attention:** This module includes several blocks, each one comprising a pair of convolutional layers, with one of them followed by a sigmoid activation function. An additional 1-D max pooling layer is applied along the frequency axis after the convolutional layers in each block.

• **Pipe-Net:** The pipe-net contains two Fully Connected Networks (FCN)s with hidden dimension N_p. Its output is represented by $\mathbf{G} \in \mathbb{R}^{N_p \times L}$, where L is the contextual dimension.

• **Temporal Attention:** STAM adopts multi-headed self-attention, allowing the model to simultaneously attend to information at different positions. The query $\mathbf{q} = \sigma(\mathbf{W_q g}) \in \mathbb{R}^{N_d}$, key $\mathbf{K} = \sigma(\mathbf{W_K G}) \in \mathbb{R}^{N_d \times L}$, and value $\mathbf{V} = \sigma(\mathbf{W_V G}) \in \mathbb{R}^{N_d \times L}$ are obtained using the pipe-net output $\mathbf{G}$, where σ is an activation function, $\mathbf{g} \in \mathbb{R}^{N_p}$ is obtained by averaging $\mathbf{G}$ along the frame dimension, and $\mathbf{W_q}, \mathbf{W_K}, \mathbf{W_V} \in \mathbb{R}^{N_d \times N_p}$ are affine transformation matrices with hidden dimension N_d. The multi-headed attention operation is employed, i.e.:

$$\mathrm{MultiHead}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_H) \quad (1)$$

$$\mathrm{head}_h = \mathrm{Attention}(\mathbf{q}_h, \mathbf{K}_h, \mathbf{V}_h) \quad (2)$$

where H is the number of parallel attention layers, or heads, $\mathbf{q}_h$, $\mathbf{K}_h$ and $\mathbf{V}_h$ are the h^{th} slice of $\mathbf{q}$, $\mathbf{K}$ and $\mathbf{V}$, respectively, and $h \in \{1, ..., H\}$ is the head index. The attention function can be calculated as follows:

$$\mathrm{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\mathbf{1}\mathbf{q}^T\mathbf{K}/\sqrt{N_d}\right) \cdot \mathbf{V} \quad (3)$$

where $\cdot$ is the element-wise product, $\mathbf{1} = \{1, 1, ..., 1\}^T$.

• **Post-Net:** The post-net includes two FCNs followed by a sigmoid activation function to make the VAD predictions.

3. PROPOSED METHOD

The proposed AM-cIRM model, whose block diagram is shown in Fig. 2, consists of three modules: cIRM extractor, feature transformation, and attention-based VAD. The cIRM extractor estimates the complex spectrogram of the clean speech (which conveys both magnitude and phase information), and outputs the cIRM. In addition to combining the cIRM with the log-Mel spectrogram, the feature transformation module acts as a preprocessor for the VAD module by incorporating contextual information from neighboring frames. Finally, the VAD module outputs speech/non-speech predictions by applying attention mechanisms to the spectro-temporal information contained in the transformed features.

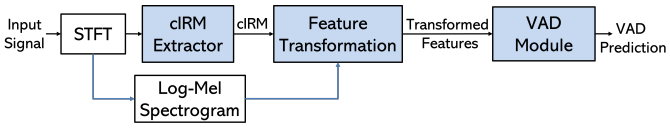

Fig. 2. Block diagram of the proposed model.

3.1. cIRM Feature Extractor

Similar to VAD, Speech Enhancement (SE) has been widely used as a prepossessing step in speech applications where one of the goals is to remove background noise from a noisy signal. Besides the classical SE methods based on statistical modeling, e.g. [22], many recent studies have focused on DNN-based SE methods. Among the later, the Deep Complex-valued U-net (DCUnet) [23] has been proposed and shown superior SE performance compared to the earlier Ideal Binary Mask (IBM) and Ideal Ratio Mask (IRM)-based [24] methods. In effect, DCUnet employs the U-Net-based architecture [25] to estimate the magnitude and phase of clean speech simultaneously. Inspired by the effectiveness of the DCUnet in extracting important speech information from noisy signals, it is chosen as the cIRM extractor for the proposed AM-cIRM model.

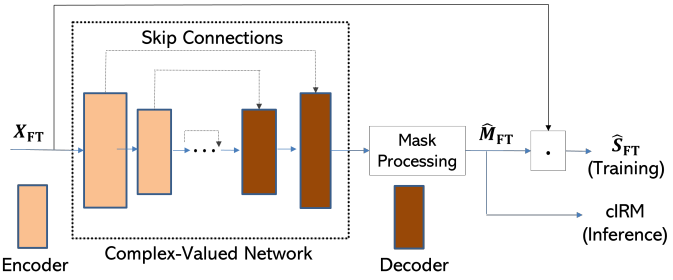

Fig. 3. Illustration of DCUnet-based cIRM extractor

The block diagram of the DCUnet-based cIRM feature extractor is shown in Fig. 3, where each encoder block consists of a complex convolutional layer, complex batch normalization layer, and leaky rectified linear unit (Leaky ReLU). In the decoding phase, skip connections are implemented by concatenating the outputs from the last decoder and corresponding encoder. The decoder is similar to the encoder except that the complex convolutional layer is replaced by a complex transposed convolutional layer. Finally, mask processing is implemented to bound the magnitude of the estimated cIRM as proposed in [23]. The input features of the cIRM extractor are the STFT coefficients of the noisy speech signal $x[n]$, represented by $\mathbf{X}_{\mathrm{FT}} \in \mathbb{C}^{F \times T}$. The main output of the extractor is the corresponding set of estimated cIRM coefficients, represented by $\hat{\mathbf{M}}_{\mathrm{FT}} \in \mathbb{C}^{F \times T}$. We note that the enhanced speech signal samples, $\{\hat{s}[n]\}$, are needed to train the DCUnet. These are obtained in two steps as follows: calculation of the enhanced speech STFT matrix $\hat{\mathbf{S}}_{\mathrm{FT}} \in \mathbb{C}^{F \times T}$ by multiplication of the noisy speech STFT with the cIRM, i.e.,

$$\hat{\mathbf{S}}_{\mathrm{FT}} = \hat{\mathbf{M}}_{\mathrm{FT}} \cdot \mathbf{X}_{\mathrm{FT}} \in \mathbb{C}^{F \times T} \quad (4)$$

followed by application of the Inverse Short-Time-Fourier-Transform (ISTFT) with overlap-add technique.

3.2. Feature Transformation

As shown in Fig. 4, the feature transformation module involves three steps. First, the estimated cIRM $\hat{\mathbf{M}}_{\mathrm{FT}} \in \mathbb{C}^{F\times T}$ is aggregated by applying a 1-D convolutional layer to compress its information content. The resulting output is denoted by $\bar{\mathbf{M}} \in \mathbb{R}^{T\times D}$, where D is the feature dimension. Secondly, the aggregated mask $\bar{\mathbf{M}}$ is concatenated with the log-Mel spectrogram features $\mathbf{X}_{\mathrm{MC}} \in \mathbb{R}^{T\times D}$, to form a new feature matrix $\boldsymbol{\chi}' \in \mathbb{R}^{T\times 2D}$. Finally, the combined features are contextually expanded by considering $L = \lfloor 2((R-1)/U)+3 \rfloor$ neighboring frames indexed by set $\mathcal{T} = \{-R, -R+U, -R+2U, ..., -1, 0, 1, ..., R-2U, R-U, R\}$, where integers R and U are user-defined parameters described in [19].

These expanded frames are used to form a super feature tensor $\boldsymbol{\chi} \in \mathbb{R}^{(T-2R)\times L\times 2D}$ for the VAD module. The expanded data set can also be represented as $\{\boldsymbol{\chi}_t, \mathbf{y}_t^{\mathrm{truth}}\}_{t=1}^{T-2R}$ with:

$$\boldsymbol{\chi}_t = \{\boldsymbol{\chi}'_{t+l}\}_{l\in\mathcal{T}} \in \mathbb{R}^{L\times 2D}, \quad \mathbf{y}_t^{\mathrm{truth}} = \{y_{t+l}^{\mathrm{truth}}\}_{l\in\mathcal{T}} \in \mathbb{R}^{L} \tag{5}$$

where $\boldsymbol{\chi}'_{t+l} \in \mathbb{R}^{2D}$ contains the $(t+l)^{th}$ row of $\boldsymbol{\chi}'$, and $y_{t+l}^{\mathrm{truth}} \in \{0,1\}$ is the ground truth label for $(t+l)^{th}$ frame.

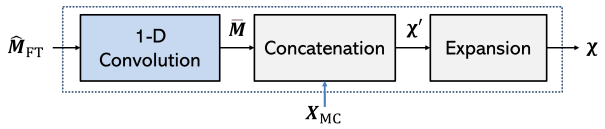

Fig. 4. Block diagram of the feature transformation module.

3.3. VAD Module

Considering the effectiveness of STAM [20], it is chosen as the VAD module in the proposed AM-cIRM. Referring to the STAM architecture in Fig. 1, since the dimension of the acoustic feature matrix $\boldsymbol{\chi}_t \in \mathbb{R}^{L\times 2D}$ is doubled in this work (due to concatenation of the log-Mel spectrogram $\mathbf{X}_{\mathrm{MC}}$ and transformed mask $\bar{\mathbf{M}}$), the number of input channels of the first pair of convolution filters in the spectral attention module is also doubled, while the number of output channels remains the same. The remaining parts of the STAM-based VAD module use the same parameter settings as the original STAM.

Let the output of the VAD module for the t^{th} frame be denoted as $\mathbf{y}_t = \{y_{t+l}\}_{l\in\mathcal{T}} \in \mathbb{R}^L$, where y_{t+l} is the soft prediction for the $(t+l)^{th}$ neighboring frame. The predicted t^{th} frame label, $\hat{y}_t$, is computed by averaging all the soft predictions relative to the current frame t across l, i.e., $\hat{y}_t = \frac{1}{L}\sum_{l\in\mathcal{T}} y_{t+l}$. The final decision label $\bar{y}_t$ is obtained by comparing $\hat{y}_t$ with a positive threshold θ_{VAD}:

$$\bar{y}_t = \begin{cases} 1, & \text{if } \hat{y}_t \geq \theta_{\mathrm{VAD}} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

3.4. Loss Functions

To prevent vanishing gradients and accelerate convergence, two different loss functions are implemented, one being calculated from the cIRM extractor and the other one from the VAD module. Let $\mathbf{x} = \{x[n]\}$, $\mathbf{s} = \{s[n]\}$ and $\hat{\mathbf{s}}[n] = \{\hat{s}[n]\}$ denote the vectors of time-domain samples of the noisy speech, clean speech and enhanced speech signals, respectively. The vectors of time-domain samples of the true noise and estimated noise samples can be obtained as $\mathbf{w} = \mathbf{x} - \mathbf{s}$ and $\hat{\mathbf{w}} = \mathbf{x} - \hat{\mathbf{s}}$, respectively. For the cIRM extractor, the weighted-Source-to-Distortion Ratio loss (wSDR) in [24] is calculated as follows:

$$\mathcal{L}_{\mathrm{wSDR}}(\mathbf{x}, \mathbf{s}, \hat{\mathbf{s}}) = \alpha\mathcal{L}_{\mathrm{SDR}}(\mathbf{s}, \hat{\mathbf{s}}) + (1-\alpha)\mathcal{L}_{\mathrm{SDR}}(\mathbf{w}, \hat{\mathbf{w}}) \tag{7}$$

$$\mathcal{L}_{\mathrm{SDR}}(\mathbf{s}, \hat{\mathbf{s}}) = -\frac{\langle \mathbf{s}, \hat{\mathbf{s}} \rangle}{\|\mathbf{s}\|\|\hat{\mathbf{s}}\|}, \quad \mathcal{L}_{\mathrm{SDR}}(\mathbf{w}, \hat{\mathbf{w}}) = -\frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle}{\|\mathbf{w}\|\|\hat{\mathbf{w}}\|} \tag{8}$$

where $\alpha = \frac{\|\mathbf{s}\|^2}{\|\mathbf{s}\|^2+\|\mathbf{w}\|^2}$ is an energy ratio, $\langle \cdot, \cdot \rangle$ is the inner product operator, and $\|\cdot\|$ is the norm operator.

For the VAD module, the cross-entropy loss is calculated for each one of the pipe-net, temporal attention, and post-net modules in Fig. 1, as proposed in [20]:

$$\mathcal{L}_{\eta} = -\sum_{t=R}^{T-R-1}\sum_{l\in\mathcal{T}}\Big(y_{t+l}^{\mathrm{truth}} \log y_{t+l}^{\eta} + (1-y_{t+l}^{\mathrm{truth}}) \log(1-y_{t+l}^{\eta})\Big) \tag{9}$$

where y_{t+l}^{η} is the $(t+l)^{th}$ component of the soft prediction vector $\mathbf{y}_t \in \mathbb{R}^L$ at the output of the corresponding module, as indicated by symbol $\eta \in \{\mathrm{pipe}, \mathrm{att}, \mathrm{post}\}$. Then the total loss for the proposed AM-cIRM model is defined as:

$$\mathcal{L}_{\mathrm{total}} = \lambda_1\mathcal{L}_{\mathrm{wSDR}} + \lambda_2\mathcal{L}_{\mathrm{pipe}} + \lambda_3\mathcal{L}_{\mathrm{att}} + \lambda_4\mathcal{L}_{\mathrm{post}} \tag{10}$$

where parameters λ_1, λ_2, λ_3 and λ_4 are the weights given to the loss functions of the different modules.

4. RESULTS

We first briefly describe our experimental setup, and then compare and discuss the performance of different methods.

4.1. Experimental Methodology

Dataset: The TIMIT corpus [26] is used to train the proposed and baseline models. In our experiments, 95% of speech utterances from the training dataset are used for training and 5% are used for model validation. A 1-second silence segment is added before and after each utterance to alleviate the class imbalance problem [20]. The training and validation sets are augmented by adding eight types of noises (babble, F16, destroyer, M109, Volvo, white, and two types of factory noises) from the NOISEX-92 dataset [27] with SNR levels at -10, -5, 0, 5, 10 dB. In the test phase, the TIMIT test dataset and the subset 'clean_test' from LibriSpeech test dataset [28] are used. All 8 types of unseen noises from the AURORA noise dataset [29] are used to corrupt the clean signals. The SNRs are set to -5, 0, 5 and 10 dB. The TIMIT dataset has ground truth labels, but the LibriSpeech dataset does not have, thus rVAD [9] is applied to the clean utterances to generate pseudo ground truth labels.

Parameter Setting: All training and test utterances are sampled at 16 kHz and framed by applying a 25 ms Hanning window with 10 ms window shifts, followed by a STFT with 1024 points. The log-Mel filter banks with feature dimension $D = 80$ and cIRM of the same dimension are processed by the VAD module. R, U and L are set to 19, 9 and 7, respectively, to form expanded feature vectors. The parameter settings of the cIRM extractor and VAD module follow the original architecture of DCUnet-10 in [23] and STAM in [20], except for the first convolution pair in the spectral attention module of the VAD module as discussed in Section 3.3. The 1-D convolution in the feature transformation module uses a stride of 1, kernel size 2 and output channels 80.

For training, the mini-batch approach with batch size of $T = 550$ is applied. The Adam optimizer is employed, with learning rate starting from 10^{-3} and exponentially decaying with rate 0.8 until reaching 10^{-5}. Parameters λ_1, λ_2, λ_3, λ_4 in Eq. (10) and θ_{VAD} in Eq. (6) are set to 0.5, 1, 0.2, 1 and 0.5, respectively. The dropout rate is set to 0.5.

Baseline Methods: The proposed method is compared with the following baseline approaches:

• **rVAD [9]:** Traditional VAD method exploiting pitch information via *a posteriori* SNR weighted energy difference.

• **DCU-10 [23]:** DNN-based SE model including 10 complex layers, extended to predict VAD labels. The cIRM $\hat{\mathbf{M}}_{\mathrm{FT}}$ is averaged along the frequency axis and the magnitude of the resulting average is compared with a threshold.

• **ACAM [19]:** DNN-based attention model for VAD exploiting only temporal attention.

• **STAM [20]:** DNN-based attention model for VAD exploiting both spectral and temporal attention.

The default parameter setting provided with the original source code is applied to rVAD, while the DNN-based models, i.e. DCU-10, ACAM, and STAM, are trained using the same approach as proposed in the above references.

Evaluation Metrics: The F1-score and detection cost function (DCF) [9] are used as metrics for comparison. The F1-score, which takes both accuracy and recall metrics into account, is commonly used as evaluation index of binary classification problems. It is calculated as $\mathrm{F1} = 2\,\mathrm{TP}\,/(2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, where TP, FP, FN represent the number of true positive, false positive, and false negative cases, respectively. The DCF, which reflects the wrong performance of the model, is defined as $\mathrm{DCF} = (1-\beta)\,\mathrm{P_{FN}} + \beta\,\mathrm{P_{FP}}$, where $\mathrm{P_{FP}}$ is the rate of FP, $\mathrm{P_{FN}}$ is the rate of FN, and β is weight herein set to 0.25 in order to penalize missed speech frames more heavily. Higher/lower values of the F1-score/DCF metrics indicate better performance.

4.2. Results and Discussion

Table 1 presents the comparative results of F1-score and DCF (both in percent), averaged over different SNRs and noise types. Clearly, all attention-based methods (ACAM, STAM and AM-cIRM) achieve better results than the non-attention-based ones (rVAD and DCU-10). For the TIMIT test dataset, STAM greatly improves the performance compared to ACAM by exploiting both spectral and temporal attention. The proposed AM-cIRM model, which exploits both the magnitude and phase information through the cIRM features, further improves the performance compared to STAM, i.e.: increase of 0.5% in F1-score and reduction of 0.7% in DCF. For the Librispeech test dataset, similar trends are observed but the improvements with AM-cIRM are even more significant.

Table 2 shows the detailed results of F1-score and DCF on TIMIT dataset with different SNR levels ranging from -5 dB to 10 dB. It can be observed that AM-cIRM outperforms all baseline methods across all SNR levels. It is also noteworthy that DCU-10 and ACAM achieve similar F1-score at low SNRs, which supports our presupposition that the cIRM contains useful information for the VAD task.

Table 3 demonstrates the influence of neighboring frames on the performance of the proposed AM-cIRM. Specifically, it shows the average F1-score and DCF values on TIMIT dataset for different choices of R and U, and fixed $L = 7$. It can be seen that at the cost of a slight decrease in performance, the latency of AM-cIRM (specified by parameter R) can be significantly reduced to facilitate real-time implementation.

Table 4 shows the number of model parameters and the averaged run time for processing a 10-second utterance. Experiments were conducted on a platform equipped with Intel Core i7-10700F CPU and NVIDIA GeForce RTX 2070 SUPER GPU. The results indicate that the merits of AM-cIRM come at the cost of additional computation resources.

Table 1. Comparison of Averaged F1-Score and DCF (in percent)

Dataset	Metric	rVAD	DCU-10	ACAM	STAM	AM-cIRM
TIMIT	F1	87.3	90.7	91.2	98.1	**98.6**
	DCF	5.4	5.1	3.7	1.3	**0.6**
Libri-Speech	F1	NA	82.5	87.5	88.3	**90.1**
	DCF	NA	15.3	11.7	13.4	**10.3**

Table 2. Comparison of F1-Score and DCF on TIMIT versus SNR

SNR	Metric	rVAD	DCU-10	ACAM	STAM	AM-cIRM
-5 dB	F1	79.5	86.4	85.9	97.7	**98.0**
	DCF	8.3	7.8	6.2	1.5	**1.0**
0 dB	F1	86.0	89.8	90.7	98.0	**98.5**
	DCF	5.8	5.7	3.7	1.3	**0.7**
5 dB	F1	92.4	92.3	95.4	98.3	**98.9**
	DCF	3.9	4.0	2.6	1.2	**0.5**
10 dB	F1	94.0	94.2	96.0	98.4	**99.1**
	DCF	3.4	2.8	2.3	1.1	**0.4**

Table 3. Influence of Neighboring Frames on Proposed AM-cIRM

Metric	R=19, U=9	R=13, U=6	R=9, U=4	R=7, U=3
Avg. F1-Score	**98.6**	98.5	98.3	98.3
Avg. DCF	**0.6**	**0.6**	**0.6**	0.7

Table 4. Number of Parameters and Averaged Running Time

Methods	rVAD	DCU-10	ACAM	STAM	AM-cIRM
Parameters	NA	2808K	957K	559K	3613K
Run Time (ms)	86	251	1263	132	269

5. CONCLUSION

In this paper, we proposed a novel VAD model, callled AM-cIRM, which firstly extracts cIRM features from noisy speech, then combines the cIRM and log-Mel spectrogram features before the temporal contextual extension, and finally applies attention model to predict the presence/absence of speech. Experimental results show that the proposed AM-cIRM achieves improved VAD performance when compared to state-of-the-art methods under different noise conditions.

6. REFERENCES

[1] J. Ramirez, J. M. Girriz and J. C. Segura, "Voice activity detection. Fundamentals and speech recognition system robustness," in M. Grimm and K. Kroschel (Eds), *Robust Speech Recognition and Understanding*, I-Tech: Vienna, 2007, pp. 1-22.

[2] X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, "Comparative study on voice activity detection algorithm," in *Proc. Int. Conf. Electric. Control Eng.*, pp. 599–602, Wuhan, China, Jun. 2010.

[3] J. C. Junqua, B. Reaves and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognizer," in *Proc. EUROSPEECH*, pp. 1371–1374, Genova, Italy, 1991.

[4] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. IEEE Speech Coding Workshop*, pp. 85–86, Quebec, Canada, Oct. 1993.

[5] A. Benyassine, E. Shlomot, H.Y. Su, D. Massaloux, C. Lamblin and J.P. Petit, "ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," in *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64-73, Sept. 1997.

[6] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE Region 10 Int. Conf. on Computers, Communications and Automation*, vol. 59, pp. 321–324, Beijing, China, Oct. 1993.

[7] J. Shen, J. Hung and L. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. Int. Conf. on Spoken Language Processing*, pp. 232–235, Sydney, Australia, Nov. 1998.

[8] 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, Mar. 2001.

[9] Z. H. Tan, A. Sarkar and N. Dehak, "rVAD: An Unsupervised Segment-Based Robust Voice Activity Detection Method," *Computer Speech and Language*, vol. 59, pp. 1-21, Jan. 2019.

[10] 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, Jan. 1999.

[11] J. Padrell, D. Macho and C. Nadeu, "Robust speech activity detection using LDA applied to FF parameters," in *Proc. ICASSP*, vol. 1, pp. I–557, Philadelphia, U.S, Mar. 2005.

[12] J. Wu and X. L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Process. Lett*., vol. 18, no. 8, pp. 466–499, Jun. 2011.

[13] 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, Mar. 2013.

[14] K. Kinoshita, M. Delcroix, A. Ogawa and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proc. INTERSPEECH*, pp. 1760–1764, 2015.

[15] Y. Tachioka, "Dnn-Based Voice Activity Detection Using Auxiliary Speech Models in Noisy Environments," in *Proc. ICASSP*, pp. 5529-5533, Calgary, CA, Apr. 2018.

[16] X. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 252–264, Dec. 2016.

[17] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using WaveNet encoder and residual networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 265–274, May 2019.

[18] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXic preprint arXiv:1609.03499*, Sep. 2016.

[19] J. Kim and M. Hahn, "Voice Activity Detection Using an Adaptive Context Attention Model," *IEEE Signal Processing Lett.*, vol. 25, no. 8, pp. 1181-1185, Aug. 2018.

[20] Y. Lee, J. Min, D. K. Han and H. Ko, "Spectro-Temporal Attention-Based Voice Activity Detection," *IEEE Signal Processing Lett.*, vol. 27, pp. 131-135, Dec. 2020.

[21] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, USA, NJ, Upper Saddle River:Prentice-Hall, 2001.

[22] M. Parchami, W. Zhu, B. Champagne and E. Plourde, "Recent Developments in Speech Enhancement in the Short-Time Fourier Transform Domain," *IEEE Circuits and Systems Magazine*, vol. 16, no. 3, pp. 45-77, Aug. 2016.

[23] H.S. Choi, J.H. Kim, J. Huh, A. Kim, J.W. Ha and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, Sep. 2018.

[24] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, pp. 7092-7096, Vancouver, CA, May 2013.

[25] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer Assisted Intervention*, pp. 234–241, Springer, Cham, Oct. 2015.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon*, USA, Tech. Rep. NISTIR 4930, vol. 93, Feb. 1993.

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

[28] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, pp. 5206–5210, South Brisbane, Australia, Apr. 2015.

[29] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Autom. Speech Recognit.: Challenges Millenium ISCA Tut. Res. Workshop*, pp. 181–188, Paris, France, Sep. 2000.