IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 February 2023, accepted 23 March 2023, date of publication 28 March 2023, date of current version 31 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262518

RESEARCH ARTICLE

Voice Activity Detection Optimized by Adaptive Attention Span Transformer

WENPENG MU AND BINGSHAN LIU
Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China
Corresponding author: Wenpeng Mu (1341494317@qq.com)

ABSTRACT Voice Activity Detection (VAD) is a widely used technique for separating vocal regions from audio signals, with applications in voice language coding, noise reduction, and other domains. While various strategies have been proposed to improve VAD performance, such as ACAM, DCU-10, and Tr-VAD, these approaches often suffer from common limitations, including being unsuitable for long audio and being time-consuming. To address these issues, a new method called AAT-VAD is proposed, which integrates an adaptive width attention learning mechanism into the classic transformer framework. The approach involves extracting Mel-scale Frequency Cepstral Coefficients (MFCC) from the Mel scale frequency domain, adding a masking function to each transformer attention head, and inputting the features processed by the transformer encoder layer into the classifier. Experimental results indicate that a 12.8% higher F1-score is achieved by the method compared to DCU-10, and a 0.6% higher F1-score is achieved compared to Tr-VAD under different noise interferences. Furthermore, the average detection cost function (DCF) value of the method is only 14.3% of DCU-10 and 92.4% of Tr-VAD, and the test time of AAT-VAD is only 37.4% of that of Tr-VAD for the same noisy vocal mixed audio.

INDEX TERMS Voice activity detection, adaptive attention span transformer, voice biometrics, voice command recognition.

I. INTRODUCTION

In recent years, biometric identification technology has become prevalent in our daily lives, particularly with the emergence of non-contact authentication methods like mobile phone face unlocking and fingerprint payment. However, fingerprint recognition requires contact and is vulnerable to interference from light and fingerprint collectors, while facial recognition is susceptible to occlusion by foreign objects. As a result, voice recognition has emerged as a promising non-contact, occlusion-free biometric identification method.

Within the field of voice recognition, Voice Activity Detection (VAD) [1], [2] plays a critical role in the extraction of human and non-human audio components, allowing for more efficient voiceprint identity authentication and channel information transmission. Machine learning and deep learning methods such as Support Vector Machine (SVM) [3], Recurrent Neural Network (RNN) [4], and Residual Neural Network (ResNet) [5] have significantly improved the accuracy of voiceprint feature recognition using VAD.

Recently, researchers have begun to apply the Transformer model [6], [7] to VAD, achieving even better performance than traditional methods. However, the quadratic cost of Transformer's calculation with input sequence size restricts its applicability for longer speech sequences with greater noise interference.

To address this limitation, this letter proposes a novel VAD detection method based on the Adaptive Attention Span Transformer (AAT-VAD). Unlike the basic Transformer model, AAT-VAD employs dynamic attention learning heads that learn the optimal attention correlation and obtain coherent audio information, allowing for the expansion of input sequence length without sacrificing performance or incurring excessive computational costs. The performance of AAT-VAD was evaluated based on F1-score, detection cost function (DCF), and average test time. Experimental results demonstrate that AAT-VAD achieves optimal performance under various conditions.

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani.

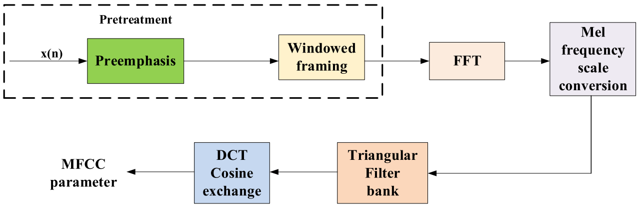

FIGURE 1. Pre-processing framework.

The contributions of this paper can be summarized in the following several aspects:

- transformer model performs well in various tasks in the computer field. Based on such an idea, we innovatively propose a new voiceprint speech recognition algorithm, AAT-VAD.
- The proposed AAT-VAD method shows its strong ability in the processing efficiency and accuracy of long audio.
- We hope it will help draw more attention to this seriously under-explored field and inspire more great works.

This letter is divided into four sections. Section II provides a brief introduction to acoustic characteristics. Section III describes the framework of AAT-VAD. Section IV outlines the experimental details, while Section V presents the conclusion.

II. FEATURE EXTRACTION

This section presents the preprocessing techniques employed to extract acoustic information. The specific framework is illustrated in Figure 1.

We model the input audio signal as $x[n]$.

$$\mathrm{x}\,[\mathrm{n}] = \mathrm{clean}\,[\mathrm{n}] + \mathrm{noise}\,[\mathrm{n}] \tag{1}$$

where clean[n] represents the clean audio signal, noise[n] denotes the background noise added to the audio, and n represents the discrete-time segment. To restrict sample amplitudes to the range of $[-1, 1]$ and diminish size discrepancies between various sound samples, we employed the z-score normalization method.

$$x^* = \frac{x-\mu}{\sigma} \tag{2}$$

The symbols μ and σ represent the mean and variance of the discrete signal x[n], respectively. Here, x∈{x[0],x[1],…,x[n]} denotes the original audio signal's discrete values.

A. PREEMPHASIS

The average power spectrum of the speech signal, $x^*[n]$, is strongly influenced by glottal excitation and oronasal radiation, with attenuation of 6 dB/octave above 800 Hz. As a result, the higher frequency components are fewer in number. Consequently, prior to the analysis of $x^*[n]$, it is necessary to enhance the high frequency portion

$$\tilde{x}^*\,[n] = x^*\,[n] - ax^*\,[n-1] \tag{3}$$

where a is the pre-emphasis factor, taken as a size of 0.9285.

B. WINDOWED FRAMING

The audio signal is a non-linear time-varying signal, but it exhibits a smooth feature in short time, rendering it easy to extract short-time features through framing. To ensure minimal change in the features between adjacent frames, a frame shift of 2/5 of the frame length is selected, leading to 3/5 overlap between adjacent frames. Furthermore, the signal is windowed using a Hamming window function, enhancing its periodic characteristics for Fourier transform processing in the time domain.

$$w\,(n) = a_0 - (1-a_0)\cdot cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \tag{4}$$

where $a_0 = 12/23$,

To determine the start and end points of audio signals, an automatic endpoint detection method was employed. This was achieved by utilizing the two-threshold comparison method, which is characterized by the short-time energy E and short-time average zero-crossing rate Z.

$$E_n = \sum_{-\infty}^{\infty} [x\,(m)\,w\,(n-m)]^2 \tag{5}$$

$$Z_n = \sum\nolimits_{m=-\infty}^{\infty} |sgn\,[x\,(m)] - sgn\,[x\,(m-1)]|\,w\,(n-m) \tag{6}$$

Combining the advantages of short-time energy Z and short-time average zero-crossing rate E, the detection is more accurate and the system processing time is reduced.

C. FFT

The speech feature parameter MFCC is used to extract the feature. It is first processed by Fast Fourier Transform (FFT).

$$X\,[k] = \sum_{n=0}^{N-1} x\,[n]\,e^{-j\frac{2\pi}{N}nk} k = 0, 1, 2, \ldots, N-1 \tag{7}$$

where $x[n](n = 0, 1, 2, \ldots, N-1)$ is the discrete frame obtained after sampling the speech sequence, and N is the total frame length. $x[k]$ is a complex sequence of N points. Then the signal amplitude spectrum $|X[k]|$ is obtained by modulo $X[k]$.

D. MEL FREQUENCY SCALE CONVERSION

Converting the actual frequency scale to the Mel frequency scale.

$$Mel\,(f) = 2597lg\left(1 + \frac{f}{700}\right) \tag{8}$$

where $Mel(f)$ is the Mel frequency and f is the actual frequency.

E. TRIANGULAR FILTER BANK

Configure the triangular filter banks, calculate the output of each triangular filter after signal amplitude filtering.

$$F\,(l) = \sum_{k=f_a(l)}^{f_k(l)} w_l\,(k)\,|X\,[k]| l = 1, 2, \ldots, L \tag{9}$$

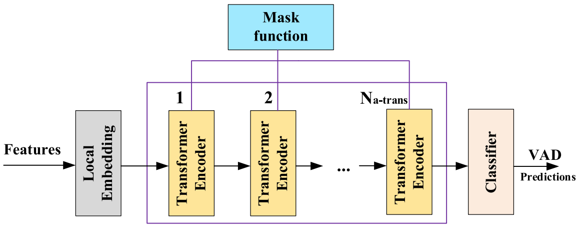

FIGURE 2. AAT-VAD model framework.

$$w_l\left(k\right)=\begin{cases}\dfrac{k-f_a\left(l\right)}{f_c\left(l\right)-f_0\left(l\right)}f_0\left(l\right)\leq k\leq f_c\left(l\right),\\ \dfrac{f_h\left(l\right)-k}{f_h\left(l\right)-f_c\left(l\right)}f_c\left(l\right)\leq k\leq f_h\left(l\right)\end{cases}\tag{10}$$

$$f_0\left(l\right)=\frac{o\left(l\right)}{\left[\frac{f_s}{N}\right]}f_h,\left(l\right)=\frac{h\left(l\right)}{\left[\frac{f_s}{N}\right]},f_c\left(l\right)=\frac{c\left(l\right)}{\left[\frac{f_s}{N}\right]}\tag{11}$$

The filter coefficient of the corresponding filter is denoted as $w_l\left(k\right)$, where $o\left(l\right)$, $c\left(l\right)$, and $h\left(l\right)$ represent the lower, center, and higher frequencies of the filter under the actual frequency, respectively. The sampling rate is represented by f_s and the number of filters by L. The output of the filter is represented as $F\left(l\right)$

F. DCT COSINE EXCHANGE

When the output of all filters is obtained, the last discrete cosine transformation (DTC) step is made to obtain the MFCC.

$$M\left(i\right)=\sqrt{\frac{2}{N}}\sum_{l=1}^{L}logF\left(l\right)cos\left[\left(l-\frac{1}{2}\right)\frac{i\pi}{L}\right]i=1,2,\ldots,Q\tag{12}$$

where Q is the order of the MFCC parameters, we take 13. $M(i)$ is the resulting MFCC parameter.

III. THE FRAMEWORK OF ADAPTIVE ATTENTION SPAN TRANSFORMER MODEL

In this section, we present our AAT-VAD model, which is designed to accurately determine the presence or absence of speech by segmenting acoustic information into smaller units and applying deep convolution. To achieve this, we employ adaptive attention learning and dynamic attention mechanisms, as illustrated in Figure 2.

The dataset used to train the model is denoted by $\{X_i, y_i^{true}\}_{i=0}^{I-1}$, where $X_i \in \mathbb{R}^D$ is the acoustic feature vector at frame i, $y_i^{truth} \in \{0, 1\}$ is the label of VAD and Tol is the total number of frames. Expand each frame's acoustic data to $L = 2k + 1$, and its relative index is $l \in$ Tol $= \{-ku, -(k-1)u, \ldots, -u, 0, u, \ldots, (k-1)u, ku\}$, where u is the step size and k is the number of adjacent frames. Expanded data can be expressed as.

$$X_i' = X_{i+l} : l\epsilon T\epsilon\mathbb{R}^{L\times D}, y_i^{truth} = y_{i+l}^{truth} : l \in T \in \mathbb{R}^L \tag{13}$$

We use the extended feature vector as input to the first embedding module, which consists of a fully connected neural network and a one-dimensional convolutional layer. Adding a convolution layer helps extract relative location information and learn useful short-term spectral-temporal patterns [9]. The output of the embedding layer is denoted as $\bar{X}_i^1 \in \mathbb{R}^{\bar{T}\times\bar{D}}$, where $\bar{T}$ and $\bar{D}$ denote the temporal and feature dimensions.

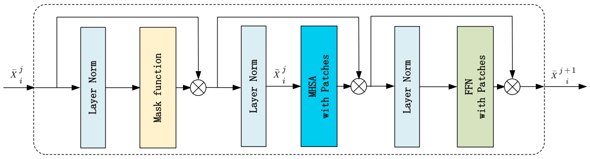

FIGURE 3. The internal structure of the *i*th transformer encoder.

Each $N_{a-trans}$ deep neural Transformer block consists of three modules: a Mask function module, a Multi-Headed Self-Attention (MHSA) module, and a Feedforward Neural Network (FNN) module, as depicted in Figure 3. Assuming that the input features of the jth Transformer are denoted as $\bar{X}_i^j \in \mathbb{R}^{\bar{T}\times\bar{D}}$, where $j \in \{1, 2, 3, \ldots, N_{a-trans}\}$.

Figure 3 shows that the normalization layer generates a normalized matrix $\tilde{X}_i^j \in \mathbb{R}^{\bar{T}\times\bar{D}}$ before passing the findings to the subsequent multi-head self-attention module.

In the Attention module, each head's attention period is learned independently. Specifically, we add a mask function to the head of each attention to control the attention span. The mask is a non-increasing function that maps distances between [0,1]. We take the lower bound of the mask function —m_θ as the true value with the parameter $\theta \in [0, S]$.

$$m_\theta\left(x\right)=min\left[max\left[\frac{1}{R}\left(R+\theta-x\right),0\right],1\right]\tag{14}$$

Here, R controls the softness of the mask function [10]. The context information and current distance are used as inputs to the mask function. θ is the parameter that needs to be learned.

$$attention\left(ir\right)=\frac{m_\theta\left(i-r\right)exp\left(s_{ir}\right)}{\sum_{q=i-S}^{i-1}m_\theta\left(i-r\right)exp\left(s_{iq}\right)}\tag{15}$$

In loss function, we add an L penalty term.

$$L=-logP\left(w_1,\ldots,w_T\right)+\frac{\lambda}{M}\sum\nolimits_k z_k\tag{16}$$

The regularization parameter λ (>0) and the number of attention heads M are critical hyperparameters in our model. These differentiable parameters are learned in conjunction with the rest of the model to optimize its performance. The regularization parameter λ controls the model's capacity and helps prevent overfitting, while the number of attention heads M determines the model's attention granularity and influences its ability to capture complex relationships among input features. By fine-tuning these parameters during the training process, we aim to achieve optimal VAD accuracy on our dataset.

The construction of the Transformer is shown in Figure 4. The output $\bar{X}_i^j$ from the mask function is partitioned into non-overlapping segments based on the temporal dimension $\bar{T}$ and the feature dimension $\bar{D}$, with N_1 segments in the temporal dimension and N_2 segments in the feature dimension.

$$\bar{X}_{i,S}^j=Sp\left(\bar{X}_i^j\right)\in\mathbb{R}^{D_s\times N_1\times N_2}\tag{17}$$

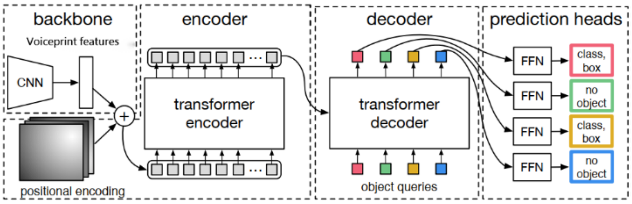

FIGURE 4. Transformer network structure in AAT-VAD.

The splitting factors N_1 and N_2, denoted by Sp(·), are used to divide the $\bar{X}^j_{i,S}$ processed by the mask function into non-overlapping segments based on the temporal dimension $\bar{T}$ and the feature dimension $\bar{D}$. Specifically, each segment has a dimension of $D_s = \frac{\bar{T}}{N_1} \times \frac{\bar{D}}{N_2}$. By extending MSHA from a single dimension to multiple dimensions, the model can be highly generalized and can handle various scenarios.

To address the limitations of Transformer's attention mechanism in capturing local information, we incorporate a highly divisible convolutional block (DW) in our model. The DW convolutional layer is used to capture local information, while the 2 × 2 convolutional layer facilitates global and partial linkage. Additionally, each convolutional block includes a batch normalization layer for processing. In convolution, the step size is set to 4, and we use DW(·) to map $\bar{X}^j_{i,S}$ to $\bar{X}^j_{i,DW}$.

$$\bar{X}^j_{i,DW} = map\left(DW\left(\bar{X}^j_{i,S}\right)\right) \in \mathbb{R}^{\frac{\bar{T}}{N_1}\times\frac{\bar{D}}{N_2}\times\frac{N_1N_2}{16}} \tag{18}$$

The presented equation involves a mapping operation, denoted as map(·), and assumes that Q, K, and V represent queries, keys, and values, respectively. The Softmax operation is applied to these components, which allows for the computation of the attention weights.

$$\bar{X}^j_{i,attention} = Softmax\left(\frac{Q^TK}{\sqrt{N_d}} + \beta\right) \cdot V \in \mathbb{R}^{\frac{\bar{T}}{N_1}\times\frac{\bar{D}}{N_2}\times\frac{N_1N_2}{16}} \tag{19}$$

The construction of the Multi-Head Self-Attention (MHSA) is depicted in Figure 4. The input tensor $\bar{X}^j_i$, which is processed by the mask function, is divided into non-overlapping segments of size $N_1 \times N_2$. These segments are based on the temporal dimension $\bar{T}$ and the feature dimension $\bar{D}$. Here, N_1 and N_2 denote the specified splitting factors, Sp(·) represents the splitting action, and D_s represents $\frac{\bar{T}}{N_1} \times \frac{\bar{D}}{N_2}$. By extending MHSA from a single dimension to multiple dimensions, the model can be highly generalized when encountering various scenario conditions.

The attention matrix is obtained by employing a highly divisible convolutional block (DW). The DW convolutional layer provides a local focus on information that the Transformer lacks, thus compensating for its shortcomings and achieving global and partial connectivity. Each convolutional block comprises a DW convolutional layer, a batch normalization layer for processing, and a 2 × 2 convolutional layer. In convolution, the step size is 4, and we use DW(·) to manipulate $\bar{X}^j_{i,S}$, which maps to $\bar{X}^j_{i,DW}$.

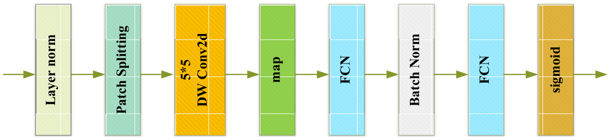

FIGURE 5. Classifier internal framework diagram.

Here, $\beta \in \mathbb{R}^{\frac{\bar{T}}{N_1}\times\frac{\bar{D}}{N_2}\times\frac{\bar{D}}{N_2}}$ represents a learning-related offset. The previous mapping employed a step size of 4, and both the time and feature dimensions have been scaled down by a factor of 4. Consequently, the self-attention cost is reduced by a factor of 16^3, which greatly reduces the cost of processing large audio signals and compresses the running time of the model.

As shown in Figure 5, the Transformer eventually outputs to the classifier. The internal structure of the classifier is plotted similarly, and the feature matrix is processed through the DW convolution block. The output feature matrix $X^j_{i,out} \in \mathbb{R}^{D_s\times\frac{N_1}{2}\times\frac{N_2}{2}}$ is mapped to $\tilde{X}^j_{i,out} \in \mathbb{R}^{\frac{N_1}{2}\times\frac{N_2D_s}{2}}$. FCNs are used to detect hidden information, and the last feature dimension is compressed to 2 dimensions. The sigmoid activation function is used to predict the probability of a human voice in audio signals, and vector $y_i \in \mathbb{R}^{\frac{N_1}{2}} = \mathbb{R}^L$ represents it. For the prediction of the ith frame, we combine the predictions of all y_i-related frames with $l \in I$. By comparing $\hat{y}_i$ with the threshold θ_T, we obtain the decision label value y_i.

$$\hat{y}_i = \frac{1}{L}\sum_{l\in I} y_{i+l} \tag{20}$$

$$y_i = \begin{cases} 1, if\ \hat{y}_i \geq \theta_T \\ 0,\ otherwise \end{cases} \tag{21}$$

where y_{i+l} is the $(i+l)$th component of y_i.

Loss function is defined as follows.

$$cost = -\sum\nolimits_{i=k}^{T-k-1}\sum\nolimits_{l'\in I}(y^{truth}_{i+l} log y_{i+l} + \left(1 - y^{truth}_{i+l}\right) log(1 - y_{i+l})) \tag{22}$$

Here y^{truth}_{i+l} is the lth component of the y^{truth}_i label.

IV. EXPERIMENTS

In this section, we present the essential details and findings of our experiment, which evaluates the performance of our proposed AAT-VAD model in multiple audio datasets and demonstrates its superiority over existing technologies.

A. DATABASE INTRODUCTION

To begin with, the widely used TIMIT corpus serves as the primary dataset for training and evaluating our model, with 95% of the speech data allocated for training and 5% for validation. To generate the noise database, we concatenated about 20,000 sound effects [11] to create a long sound wave, and then randomly added -10∼12db long sound waves to the TIMIT corpus. This process is repeated until the end of the long sound wave is reached.

For dataset D1, we combined over 1300 TIMIT corpora with silent conversations and added seven noises from the NOISERX-92 database. Additionally, we addressed the issue of signal-to-noise ratio mismatch.

For dataset D2, we utilized the ST-CMDS Chinese dataset, which contains more than 100,000 speech files from over 800 different speakers, with durations over 100 hours, and includes ambient noise.

The extended dataset requires the parameters K, u, and L to be set to 4, 4, and 9, respectively. We employed the 512 small batch method for training, with the Cosine Decay Learning Rate Scheduler AdamW [12] and 5000 linear warm-up iterations. The initial learning rate was set at 0.001, with a weight decay rate of 0.05, and after 4 × 105 iterations, the final learning rate was set at 5 × 10-6. We employed the Gaussian Error Linear Unit function (GELU) [13] as the activation function, and the model parameters were adjusted to 80, 54, 162, 18, 18, 27, 0.5, and 6, with an exit rate of 0.1. The overall parameter size and a comparison with other models are presented in Table 1.

TABLE 1. Comparison of the parameters of each model.

Method	Tr-VAD	bDNN	DNN	LSTM	ACAM	AAT-VAD
Parameters	~864K	~3010K	~3008K	~2100K	~953K	~327K

B. MODEL COMPARISON AND EVALUATION INDEX

We conducted separate comparisons between the AAT-VAD model and the following approaches:

- rVAD [14]: An unsupervised learning VAD method that utilizes the underlying audio information by computing a posteriori signal-to-noise ratio-weighted energy difference.
- Adaptive Contextual Attention Model (ACAM) [15]: A fundamental VAD model based on an attention mechanism that primarily employs spectral and temporal information.
- DCU-10 [16]: A DNN-based speech enhancement model with 10 complex layers that we extended to predict VAD labels. The method involves taking the average value of the ideal ratio mask along the frequency axis and comparing it with the threshold value.
- Tr-VAD [8]: The original Transformer-based strategy that primarily utilizes self-awareness mechanisms and multiple attention heads.

Our experiments demonstrate that the AAT-VAD model outperforms these existing methods in terms of accuracy and efficiency, indicating its potential for use in real-world applications.

In this section, we present the experimental details for evaluating the performance of the proposed AAT-VAD model and compare it with state-of-the-art methods. While rVAD employs a default parameter design, all other methods are trained using the approach detailed in the aforementioned reference. To assess the effectiveness of the proposed method, we utilize commonly used evaluation metrics for binary classification problems, namely, F1-score and DCF. These metrics have been widely adopted to measure the precision and robustness of VAD models.

$$F1 = 2TP/\left(2TP + FP + FN\right) \tag{23}$$

TABLE 2. F1-scores and DCF for each model.

SNR	Metric	rVAD	DCU-10	Tr-VAD	AAT-VAD
-5db	F1	79.5	86.5	98.6	99.3
	DCF	8.3	7.7	0.9	0.8
0db	F1	86.0	89.9	98.7	99.0
	DCF	5.8	5.6	0.7	0.7
5db	F1	92.4	92.2	99.0	99.3
	DCF	3.8	4.0	0.6	0.6
10db	F1	94.1	94.2	99.1	99.3
	DCF	3.4	2.8	0.6	0.5

In this context, TP, FP, and FN denote the quantities of true positives, false positives, and false negatives, while DCF is utilized as a gauge of the model's error performance. It is commonly employed in binary classification problems and computed as the weighted sum of false negative and false positive errors, where the weights depend on the relative costs of these errors. The higher the value of DCF, the worse the performance of the model.

$$DCT = (1 - \beta)\, P_{FN} + \beta P_{FP} \tag{24}$$

In this context, the rate of false positives is denoted by P_{FP}, while the rate of false negatives is represented by P_{FN}. Additionally, β is a weight parameter set to 1/4 to penalize lost speech frames. As per definition, a higher F1-score and a smaller DCF coefficient indicate superior model performance.

C. RESULT ANALYSIS

The exact performance findings are provided in Table 2, which shows the average results across all datasets. Based on these data, our proposed AAT-VAD method has demonstrated superior performance compared to the current state-of-the-art Tr-VAD method under different signal-to-noise ratios. Specifically, AAT-VAD showed an average improvement of 0.6 percentage points in the F1 index and nearly 15% optimization in the DCF index. Additionally, Figure 1 illustrates that the AAT-VAD method significantly reduces the number of parameters compared to several commonly used networks, resulting in a significant reduction in training time. These findings highlight the effectiveness and efficiency of the AAT-VAD method for voice activity detection in noisy audio signals.

D. EXPERIMENTAL ENVIRONMENT

The hardware environment utilized for this study included a high-performance computing cluster and a cloud computing platform for training and evaluating deep learning models. The computing cluster comprised multiple processors and accelerators such as Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), which provided efficient parallel computing capabilities.

In addition to the computing cluster, we utilized specialized hardware devices, such as microphone arrays and audio signal processors, designed specifically for voiceprint detection.

For software, we employed TensorFlow, PyTorch, and Librosa, which are widely used deep learning frameworks and audio signal processing libraries. The programming language used was Python.

The dataset utilized in this study was TIMIT and ST-CMDS, two widely used datasets comprising a large number of voice samples and related metadata. These datasets include voice data collected from diverse angles, distances, and environments to simulate voiceprint detection tasks in real-world scenarios. Additionally, we applied various preprocessing techniques, such as data augmentation and normalization, to enhance the robustness and generalization of the model.

This study's hardware and software environment and dataset were carefully selected to ensure that the experimental setup was adequate to explore the research question effectively. The utilization of these advanced hardware and software tools and datasets was critical to the success of the study, and the results obtained validate our approach.

V. CONCLUSION

In summary, this paper introduces the AAT-VAD model, which leverages adaptive attention span transformer and a mask function layer to effectively process long audio segments and reduce computational costs. The proposed model is composed of four main components and outperforms existing methods in terms of F1-score and DCF across various noise environments. This work contributes to the field of voiceprint detection by providing a new and effective approach to audio processing. The mask function layer and attention span transformer are novel features that enable efficient and accurate voice detection, which is crucial in real-world applications. Overall, the AAT-VAD model presents a significant innovation and improvement over existing methods for voiceprint detection.

The present study demonstrates the effectiveness of the proposed AAT-VAD method in voice activity detection. However, there is still room for improvement in voiceprint detection technology. Future research could focus on enhancing the precision and reliability of voiceprint recognition, supporting multiple languages and diverse application scenarios, and addressing privacy protection and ethical issues. As such, the future of voiceprint detection technology is promising, with the potential for further advancements and widespread application.

REFERENCES

[1] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2026–2038, Sep. 2011.

[2] H. Ghaemmaghami, B. J. Baker, R. J. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 3118–3121.

[3] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class SVM for learning in image retrieval," in *Proc. Int. Conf. Image Process.*, 2001, pp. 34–37.

[4] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[5] Y. Hu, H. Tang, and G. Pan, "Spiking deep residual networks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 1, 2021, doi: 10.1109/TNNLS.2021.3119238.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[7] Y. Zhao and B. Champagne, "An efficient transformer-based model for voice activity detection," in *Proc. IEEE 32nd Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Aug. 2022, pp. 1–6.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[9] B. S. Shawel, D. H. Woldegebreal, and S. Pollin, "Convolutional LSTM-based long-term spectrum prediction for dynamic spectrum access," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[10] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1562–1566.

[11] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 373–376.

[12] R. Llugsi, S. E. Yacoubi, A. Fontaine, and P. Lupera, "Comparison between Adam, AdaMax and Adam W optimizers to implement a weather forecast based on neural networks for the Andean city of Quito," in *Proc. IEEE 5th Ecuador Tech. Chapters Meeting (ETCM)*, Oct. 2021, pp. 1–6.

[13] A. Nguyen, K. Pham, D. Ngo, T. Ngo, and L. Pham, "An analysis of state-of-the-art activation functions for supervised deep neural network," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Aug. 2021, pp. 215–220.

[14] Z.-H. Tan, A. K. Sarkar, and N. Dehak, "RVAD: An unsupervised segment-based robust voice activity detection method," *Comput. Speech Lang.*, vol. 59, pp. 1–21, Jan. 2020.

[15] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.

[16] 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 *Proc. Int. Conf. Learn. Represent.*, Sep. 2018, pp. 1–20.

WENPENG MU was born in Jiangsu, China, in 2002. He is currently pursuing the bachelor's degree with the Nanjing University of Information Science and Technology.

From 2020 to 2023, he was a Research Associate with the Engineering Research Center, Ministry of Digital Forensics. He has one software copyright and one patent pending. His research interests include tamper detection of voicing fingerprints, deep model watermarking, and countering sample attacks.

Mr. Mu has won the National Gold Medal of the China International "Internet Plus" College Student Innovation and Entrepreneurship Competition and the third prize of the National Advanced Mathematics Competition.

BINGSHAN LIU was born in Jiangsu, China, in 2003. He is currently pursuing the degree in computer science and technology with the Nanjing University of Information and Science Technology.

In 2021 winter, he participated in the Winter Social Practice Project to experience the rice processing in a rice factory. He learned about the modern mechanical operation of rice processing. In May 2022, his article titled "Research on the Impact of Original Families on Individuals' Social Intercourse" (*Literature World*, a Chinese provincial magazine). He is the author of two inventions, such as Computer Software Copyright: Supermarket Cashier System and Utility Model Patent: The Mobile Hard Disk Drive for Information Security. His research interests include data analysis and digging as well as monitor of vocal print. He is excellent at probability statistics and programming.

Mr. Liu was a member of the 8th National College Students' Innovation and Entrepreneurship Training Program. His team has won Gold Award around China, in December 2022.