Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(3): 3860–3875.
DOI: 10.3934/mbe.2024172
Received: 29 November 2023
Revised: 15 January 2024
Accepted: 22 January 2024
Published: 21 February 2024

Research article

SASEGAN-TCN: Speech enhancement algorithm based on self-attention generative adversarial network and temporal convolutional network

Rongchuang Lv[1], Niansheng Chen[1], Songlin Cheng[1,*], Guangyu Fan[1], Lei Rao[1], Xiaoyong Song[1], Wenjing Lv[2] and Dingyu Yang[3]

[1] School of Electronic Information Engineering, Shanghai Dianji University, Shanghai 201306, China

[2] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

[3] Alibaba Group, Shanghai 201203, China

* **Correspondence:** Email: chengsl@sdju.edu.cn.

Abstract: Traditional unsupervised speech enhancement models often have problems such as non-aggregation of input feature information, which will introduce additional noise during training, thereby reducing the quality of the speech signal. In order to solve the above problems, this paper analyzed the impact of problems such as non-aggregation of input speech feature information on its performance. Moreover, this article introduced a temporal convolutional neural network and proposed a SASEGAN-TCN speech enhancement model, which captured local features information and aggregated global feature information to improve model effect and training stability. The simulation experiment results showed that the model can achieve 2.1636 and 92.78% in perceptual evaluation of speech quality (PESQ) score and short-time objective intelligibility (STOI) on the Valentini dataset, and can accordingly reach 1.8077 and 83.54% on the THCHS30 dataset. In addition, this article used the enhanced speech data for the acoustic model to verify the recognition accuracy. The speech recognition error rate was reduced by 17.4%, which was a significant improvement compared to the baseline model experimental results.

Keywords: speech enhancement; deep learning; generative adversarial network; autoencoder

1. Introduction

As a new technology of the internet of things, speech recognition plays an important role in various electronic products such as smart homes and vehicle-mounted equipment. However, the interference of surrounding environmental noise can seriously affect the quality and intelligibility of the speech signal.

In response to the above problems, speech enhancement technology aimed at improving the quality of the speech signal, reducing noise, and enhancing speech information has emerged [1, 2].

In the last century, by reason of limited resources and immature advanced technologies, people were able to rely more on traditional methods and techniques. Boll et al. [3] tried to obtain clear speech noise by subtracting the noise part from the spectrum, but spectral subtraction does not work well for nonstationary noise. To address the aforementioned issues, Ephraim and et al. [4] reduced the impact of noise on the speech signal by calculating the average value of samples within the window, and the experimental results show that the quality and intelligibility of speech signals has been improved significantly compared to other models.With the aim of further deepening the effect of the model in the face of nonstationary noise, some researchers used the median value of the value in the window to replace the value of the sampling point, which further improved the model's denoising effect on nonstationary noise and sudden noise [5, 6]. With a view to solve the limitations of the median filtering method, Widrowand et al. [7] used adaptive filtering, which can automatically adjust parameters according to the signal and noise, improve signal quality, effectively suppress various noises, and it is suitable for complex noise environments and real-time signal processing. Although traditional methods have made many achievements in the field of speech enhancement, their scope of use is still limited, such as the detailed parts of the speech signal and the use environment. However, deep learning methods are able to compensate for these deficiencies through data-driven feature learning, thereby achieving better noise suppression and speech enhancement [8, 9].

Up to now, speech enhancement technology has completed the transformation from traditional signal processing methods to deep learning methods [10, 11]. Among them, Grais et al. [12, 13] used a deep neural network (DNN) to process speech signals, and it completes the modeling of the spectrum or time domain characteristics of the speech signal and finds out the nonlinearity between the speech signal and the noise. Subsequently, as the complexity of speech enhancement tasks became higher and higher, Strake et al. [14, 15] introduced the convolutional neural network (CNN) into speech enhancement technology to solve complex speech enhancement problems. CNN is deeply loved by researchers due to its efficient feature extraction capability and small number of parameters. Nonetheless, CNN still cannot learn features directly from the original signal when processing speech signals, which means that it has limitations in modeling time series data [16]. To address the issues mentioned above, Choi et al. [17, 18] began to introduce recurrent neural network (RNN) into the speech enhancement model to improve the modeling ability of speech signals and noise. At the same time, Hsieh et al. [19, 20] combined CNN and RNN to not only improve the model's ability for time series data, but also speed up the model's training and prediction speed. In recent years, under the concept of data-driven models, autoencoders (AED) [21] and generative adversarial neural networks (GAN) [22] have begun to emerge, among which the AED model can realize unsupervised learning of low-dimensional representations of data and reduce the need for labels, making model training more flexible. The GAN model consisting of a generator and a discriminator is also an unsupervised learning method, which achieves data enhancement through adversarial training. Pascual et al. [23, 24] demonstrated for the first time that its performance in the field of speech enhancement has significantly improved compared to other models. However, there are many problems with the GAN model in practical applications [25, 26]. In order to further improve model performance, Hao et al. [27] began to introduce deep learning technologies such as attention mechanisms into the GAN model, and relative experimental results showed that the model can effectively capture local feature information and establish a long sequence dependency relationship

with the data. With the aim of further enhancing the feature extraction and data generation enhancement capabilities of the model, Pandey et al. [28] combined the AED and GAN models to implement a more flexible enhancement strategy.

This type of model has good performance in processing speech signals, for example, the generator of GAN can generate synthetic samples similar to real speech, and improve the generation effect through adversarial training. Additionally, GAN is able to learn and process complex speech features, including speech speed, pitch, and noise, thereby making the model more able to approximate the performance of real speech. Moreover, GAN is an unsupervised learning method that does not require a large amount of clearly labeled speech data and can reduce the difficulty of data acquisition. Last but not least, the generator of GAN can simulate multiple types of noise and makes the model highly robust in different environments, thereby improving the effectiveness of speech enhancement. These features make GAN a powerful tool for processing speech enhancement tasks. Nevertheless, these models possess certain drawbacks, such as the absence of aggregated feature information. The specific reasons why the structure design of the network may lead to discrete and non-aggregated feature information, include mismatched hierarchical structures between encoders and decoders as well as a lack of effective information transmission mechanisms in hierarchical structure design. The overly simple design of the network structure is the main factor that cannot fully capture and transmit the correlation of complex data, which results in the loss of continuity and integrity of feature information during the transmission process. However, the above models still cannot obtain the best speech enhancement effect. Through investigation and research, this article found that the above models have ignored the impact of feature information aggregation between the encoder and decoder on the performance. Therefore, this article will focus on the problem of non-aggregation of generator feature information in the GAN network.

Considering the above factors, this paper fully exploits the network advantages of the temporal convolutional network (TCN) [29]. By introducing modules such as multilayer convolutional layers, dilate causal convolutions, and residual connections in the TCN network to aggregate and interact feature information effectively, the goal is to capture the feature information between the encoder and decoder to improve feature expression ability of the overall network. The main contributions of this article are summarized as follows:

- A novel speech enhancement model is proposed. We have made some extension work on the basis of the Self-Attention Generative Adversarial Network for Speech Enhancement (SASEGAN) model [30]. By integrating the TCN network with the generator, this model can capture the local feature information and long-distance feature information to solve the problem of non-aggregation of feature information. Moreover, our model obviously improves speech signal quality and intelligibility.

- This article uses Chinese and English datasets to conduct experimental verification analysis based on SEGAN and SASEGAN models, respectively. The experimental results perform well, which validates the effectiveness and generalization of the model. During the training phase, the model has a relatively smooth and stable loss curve, which verifies that the model is more stable and has a good fitting ability compared to other models.

The remainder of this paper is organized as follows. We introduce the two baseline models of SEGAN and SASEGAN in Section 2. In Section 3, the SASEGAN-TCN model is proposed. In Section 4, we

introduce the relevant configuration of the experiment, and the results of multiple sets of experimental data are analyzed and discussed in depth.

2. Baseline model: SEGAN and SASEGAN

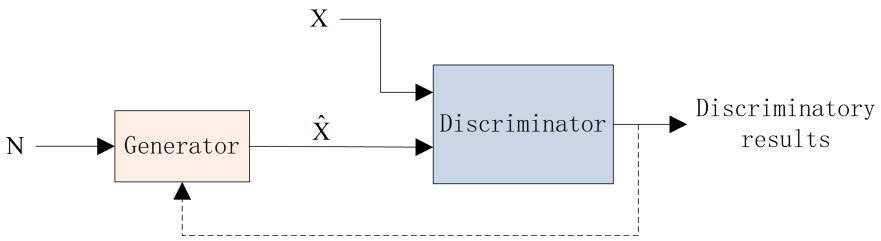

Figure 1. The structure of SEGAN model.

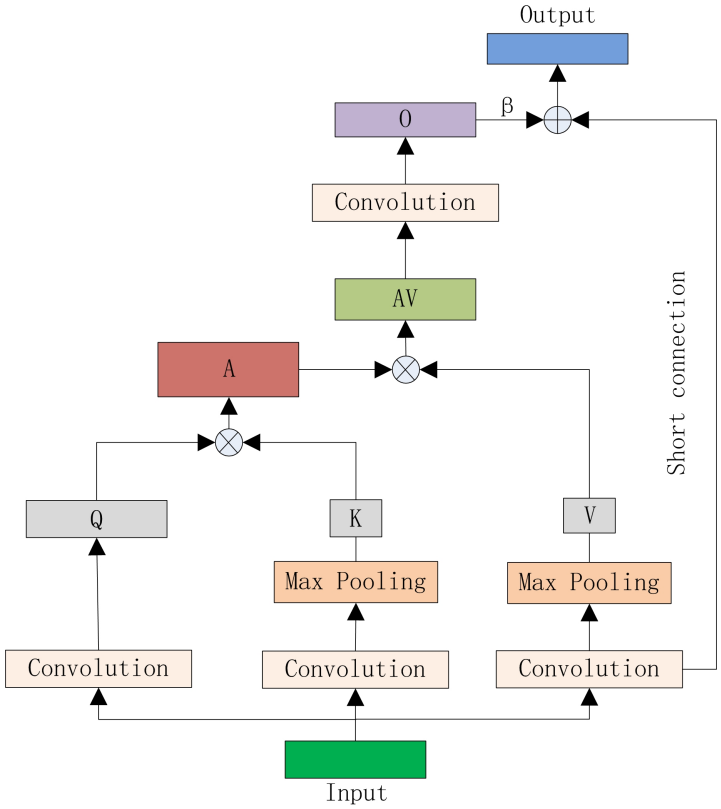

Figure 2. The structure of self-attention network.

Assume that the speech signal input to the GAN model is $\tilde{X} = X + N$, where X and N represent the intermediate variables of input data, noise, respectively. As shown in Figure 1, the goal of speech enhancement is to recover a clean signal X from a noisy signal $\tilde{X}$. The SEGAN method generates enhanced data $\hat{X} = G(\tilde{X}, Z)$ by using a generator G, where Z represent the data of encoder input value decoder. The task of the discriminator D is to distinguish between the enhanced data and the real clean

signal and learn to classify as true or false. At the same time, the generator G learns and generates an enhanced signal in order for the discriminator D to classify data as true. SEGAN is trained through this adversarial method and the least squares loss function. The least squares target loss function calculation formula of D and G can be expressed as:

$$\min_D L_{LS}(D) = \frac{1}{2}E_{X,\tilde{X}\sim p_{\text{data}}(X,\tilde{X})}(D(X,\tilde{X})-1)^2 + \frac{1}{2}E_{Z\sim p_Z(Z),\tilde{X}\sim p_{\text{data}}(\tilde{X})}D(G(Z,\tilde{X}),\tilde{X})^2, \tag{2.1}$$

$$\min_G L_{LS}(G) = \frac{1}{2}E_{Z\sim p_Z(Z),\tilde{X}\sim p_{\text{data}(\tilde{X})}}(D(G(Z,\tilde{X}),\tilde{X})-1)^2 + \lambda\|G(Z,\tilde{X})-X\|_1, \tag{2.2}$$

where $p_{data}(X)$ and Z represent the distribution probability density function of real data and latent variables, respectively. X, N, and E represent the clean speech signal, additive background noise and the expected value with respect to the distribution specified in the subscript, respectively.

When traditional GANs perform speech enhancement, they often rely entirely on the convolution operations of each layer of the CNN in the model, which may blur the event correlation of the entire sequence and provide a way to capture the correlation between long-distance speech data. The SASEGAN model combines a self-attention layer that can adapt to nonlocal features and the convolutional layer in the SEGAN model, and the effect is significantly improved.

The structure diagram of the self-attention layer is shown in Figure 2. The conv and pooling in the figure represent the convolutional layer and the max pooling layer, respectively. Assume that the input speech feature data is $F \in R^{L\times C}$, and choose to use a one-dimensional convolution to calculate one dimensional feature data. Query vector (Q), key vector (K), and value vector (V) are derived as follows:

$$Q = FW_Q, \quad K = FW_K, \quad V = FW_V, \tag{2.3}$$

where L and C represent the time dimension and the number of channels, respectively. $W_Q \in R^{C\times\frac{C}{k}}$, $W_K \in R^{C\times\frac{C}{k}}$, and $W_V \in R^{C\times\frac{C}{k}}$ represent weight matrices. Their values are determined by the convolution layer with the number of channels as $\frac{C}{k}$ and the convolution kernel size as (1×1), respectively. The optimization of the feature dimension is achieved by setting the variable k. At the same time, K and V of appropriate dimensions are selected by introducing the variable p, then the relative lower complexity O, A, and O are as follows:

$$A = \text{softmax}\left(Q\overline{K^T}\right), \quad A \in \mathbb{R}^{L\times\frac{L}{p}}, \tag{2.4}$$

$$O = (AV)W_O, \quad W_O \in \mathbb{R}^{\frac{C}{k}\times C}, \tag{2.5}$$

where k=2, p=3, C=4, and L=6 by introducing the variable β. The convolution and other nonlinear operations are used to obtain the output result O_{out}, which can be expressed as:

$$O_{\text{out}} = \beta O + F. \tag{2.6}$$

3. SASEGAN-TCN model

In the generator, in an effort to enhance the feature representation ability between the encoder and the decoder, the existing technology often ignores the aggregation of feature information between the encoder and the decoder, and the model cannot obtain long-distance feature dependencies. To this

end, this paper proposes a SASEGAN-TCN model, whose generator structure diagram is presented in Figure 3:

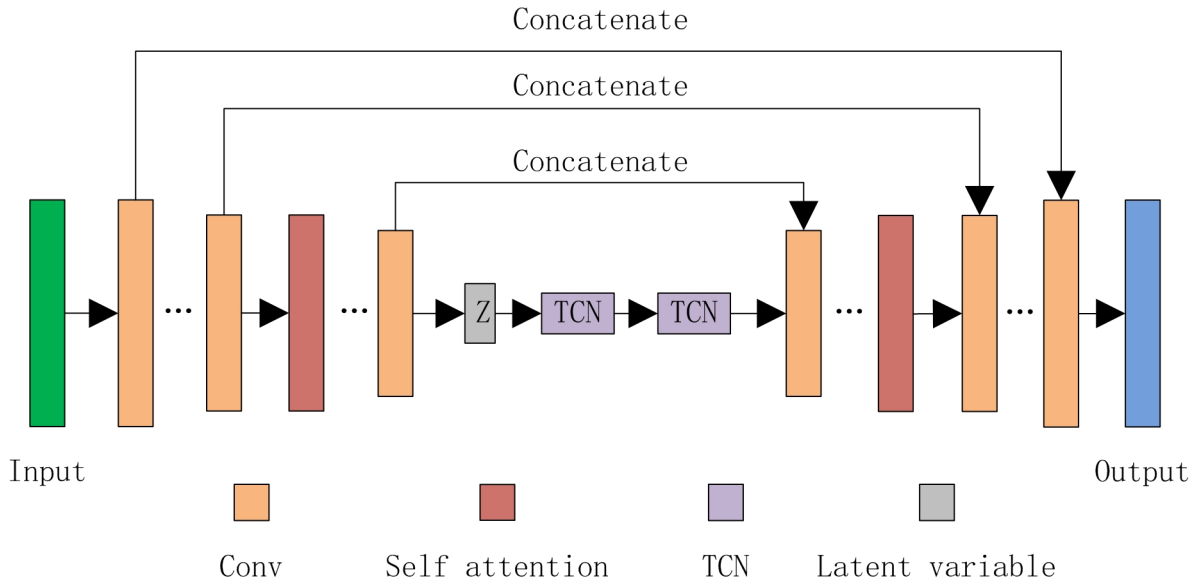

Figure 3. The generator structure of SASEGAN-TCN model.

In Figure 3, the speech signal is first extracted into matrix data with a dimension of (8192×16) through feature extraction. Second, a downsampling operation is performed through a multilayer CNN to compress the feature information, then the self-attention layer is used to obtain the dependencies of long-distance feature information until the latent variable Z between the encoder and the decoder is extracted. Finally, the obtained feature information is aggregated again through the TCN network layer. By virtue of the hole causal convolution and sum in the TCN network, the residual connection module not only avoids problems such as gradient disappearance and long-term dependence in traditional CNNs, but also it achieves the effect of aggregating feature information between the encoder and the decoder.

3.1. Dilate causal convolution

Although the SASEGAN model generates some feature vectors at each time step in the encoder, these features can only describe the local information of the input sequence, and the output of each time step is only related to the previous input in the decoder. The above situation will lead to the problem of non-aggregation of feature data in variable Z. We will choose the SASEGAN model based on the self-attention mechanism at the 10th layer for research and analysis. When processing time series data, the traditional CNN has some limitations. For example, when using a convolution kernel with a fixed kernel size for operation, the receptive field of the model is limited, which makes it impossible for the model to capture time dependencies within a limited range. In consideration of the foregoing challenges, dilated causal convolution combines the characteristics of dilated convolution and causal convolution to achieve an increase in the receptive field and an improvement in parameter efficiency and parallel operation efficiency. It can well handle long-term trend and periodic pattern data, and achieve an effect of feature information aggregation. Its structure is shown as:

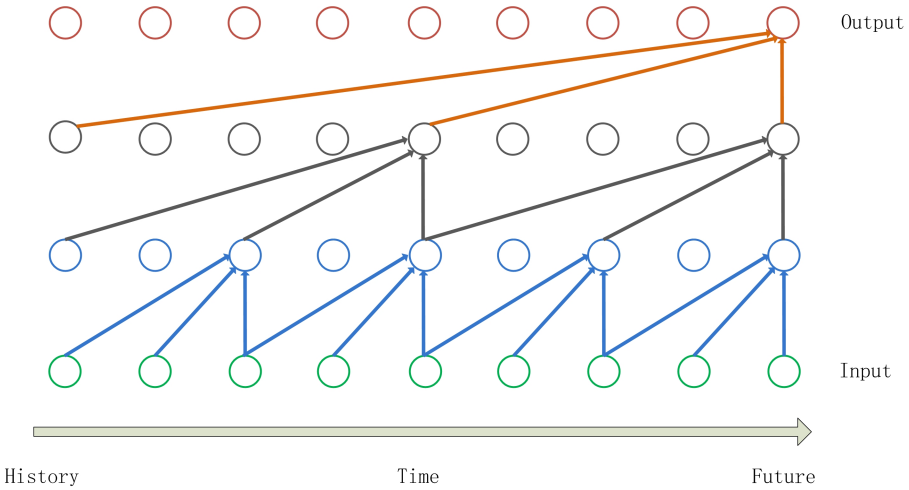

Figure 4. The structure diagram of dilated causal convolution.

In Figure 4, assuming that the input time series data is $z = [z[0], z[1], z[2], z[3], z[4], z[5], ...z[i]]$, the calculation formula of the dilate causal convolution output result is shown as:

$$l[t] = \sum_{c} z[t - d \cdot c] \cdot w[c], \tag{3.1}$$

where i, d, k, $l[t]$, $z[t - d \cdot c]$, and $w[c]$ represent the time step, dilatation rate, index of the convolution kernel, the output of the t time step, the input data at the time step, and the weight of the convolution kernel at the convolution kernel index c, respectively.

3.2. Residual module

This paper takes into account the problems of gradient disappearance and gradient explosion when traditional recursive neural networks process time series data. Therefore, the TCN network uses residual connection to bypass the feature information of the convolution layer and directly transfer the original feature information to the output layer. To alleviate the gradient descent problem and improve the information transfer of the network, we assume that the input is x, and the output result after the Rectified Linear Unit (RELU) nonlinear operations is F, then the calculation formula of the final output result o of the residual network is shown as:

$$o = x + F(x, W), \tag{3.2}$$

where $F(x, W)$ and W represent the nonlinear operation and network weight of the residual part, respectively.

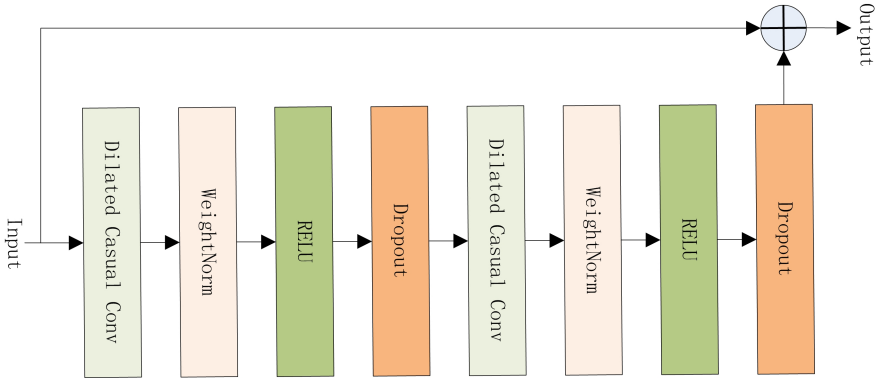

Figure 5. The structure of TCN network residual module.

The residual connection module in the TCN network is shown in Figure 5. The TCN network can well aggregate feature information and realize the interaction of feature information through methods such as multilayer convolution layers, dilate causal convolutions, and residual connections to achieve the goal of improving the overall network performance and feature expression ability. Accordingly, we have effectively integrated the SASEGAN model and the TCN network, as well as processed the final output result (latent variable Z) of the encoder in the generator through the two-layer TCN network to achieve the aggregation of feature information and improve the speech enhancement effect.

4. Experiments

4.1. Experimental parameters

This article uses the Valentini English dataset [31] and the THCHS30 Chinese dataset [32] with both audio sampling rates of 16 kHz. The Valentini dataset contains audio data from 30 pronunciation members in the Voice Bank corpus, and the training set was recorded by 28 pronunciation members. This pronunciation data was mixed with 10 different types of noise at signal-to-noise ratios of 15, 10, 5, and 0 db, respectively. The test set was recorded by 2 pronunciation members. After recording, it was mixed with 5 types of noise in the Demand audio library, with a signal-to-noise ratio of 17.5, 12.5, 7.5 and 2.5 db as the mixing conditions. First, we adjust the sampling rate of 15 audio signals in NoiseX-92 and concatenate them to form a long-term noisy audio data. Second, we traverse the training and testing sets in the THCHS30 dataset, then randomly select a long period of noisy audio data and mix it with mixing conditions of one of the four signal-to-noise ratios of 0, 5, 10 and 15 db. In this experiment, Table 1 shows the output data dimensions of each layer of the generator.

Table 1. Output dimensions of each convolutional layer in the generator.

Layer	1	2	3	4	5	6	7	8	9	10	11
Encoder	(8192×16)	(4096×32)	(2048×32)	(1024×64)	(512 × 64)	(256×128)	(128×128)	(64 × 256)	(32 × 256)	(16 × 512)	(8 × 1024)
Decoder	(16 × 512)	(32 × 256)	(64 × 256)	(128×128)	(256×128)	(512 × 64)	(1024×64)	(2048×32)	(4096×32)	(8192×16)	(16, 384 × 1)

4.2. Data processing

These experiments are conducted on a 2060 graphics card with 6 GB memory and the Windows system, and the software is used in Python version 3.7 and TensorFlow version 1.13. At training time,

the raw audio segments in the batch are sampled from the training data with 50% overlap, followed by a high-frequency pre-emphasis filter with a synergy efficiency of 0.95. Because the computer hardware configuration is limited, the TCN network used in this article has only two layers, and the number of channels is 32 and 16, respectively. The models are trained for 10 rounds with a batch size of 10, and the learning rates of the generator and discriminator models are both 0.0002.

To evaluate the effectiveness of the model experiments, this article will elaborate analysis based on various indicators. PESQ acts as an objective measure of speech quality, typically ranging from -0.5 to 4.5. A superior PESQ score indicates enhanced speech quality, and it's a pivotal metric for assessing the performance of speech encoding, decoding, and communication systems. As a comprehensive signal-to-noise ratio indicator, Channel State Information Gain (CSIG) evaluates the ratio of speech signals to noise, with a higher CSIG score reflecting an improved signal quality. Mean opinion score prediction of the intrusiveness of background noise (CBAK) serves as a comprehensive indicator for background noise suppression, and measures the extent of noise reduction in speech signals. A heightened CBAK score signifies more effective background noise suppression. Mean opinion score prediction of the overall effect (COVL) assesses the coverage of speech quality assessment algorithms across various quality levels and offers a more thorough evaluation of system performance. Lastly, Segmental Signal-to-Noise Ratio (SSNR), as a segmented signal-to-noise ratio indicator, is employed to assess the ratio between speech signals.

4.3. Experimental analysis

In order to verify the effectiveness of this method, this paper first conducts experiments on the Valentini dataset. It can be seen from Table 2 that SEGAN-TCN has improved in PESQ, STOI, SSNR, and other indicators compared with the SEGAN model. Specifically, PESQ, CBAK, COVL, and STOI reached 2.1476, 2.8472, 2.7079 and 92.61% and have been improved by 9.0, 16.7, 3.0 and 0.5% compared with noisy data, in addition, the SSNR increased by 5.3724 db. However, the CSIG has been slightly reduced due to improper selection of data processing methods and insufficient model training, which will be elaborated later.

Table 2. SEGAN and SEGAN-TCN experimental results of on the Valentini dataset.

	PESQ	CSIG	CBAK	COVL	SSNR	STOI
NOISY	1.97	**3.35**	2.44	2.63	1.68	92.11
SEGAN [23]	1.8176	3.0043	2.4423	2.3691	3.4108	91.24
SEGAN-TCN	**2.1476**	3.3388	**2.8472**	**2.7079**	**7.0524**	**92.61**

During the training process of the SEGAN model and the SEGAN-TCN model, the false sample loss value of the discriminator (d_fk_loss), the real sample loss value of the discriminator (d_rl_loss), the adversarial loss value of the generator (g_adv_loss), and the L1 loss value of the generator (g_l1_loss) curve chart are shown in Figure 6. This article records data every 100 steps and plots it. As can be seen from Figure 6, the SEGAN-TCN model loss value decline curves are smoother than the SEGAN model curves, and the training process is relatively stable. A decline in the d_fk_loss value denotes the discriminator's increased proficiency in distinguishing the generated samples as counterfeit, while a reduction the in the d_rl_loss value indicates the discriminator's heightened ability to accurately

classify genuine samples as authentic. The diminishing g_adv_loss value suggests the generator's success in outsmarting the discriminator and creating realistic samples. Meanwhile, the decrease in the g_l1_loss value signifies the similarity, at the pixel level, between the generator-produced sample and the authentic sample.

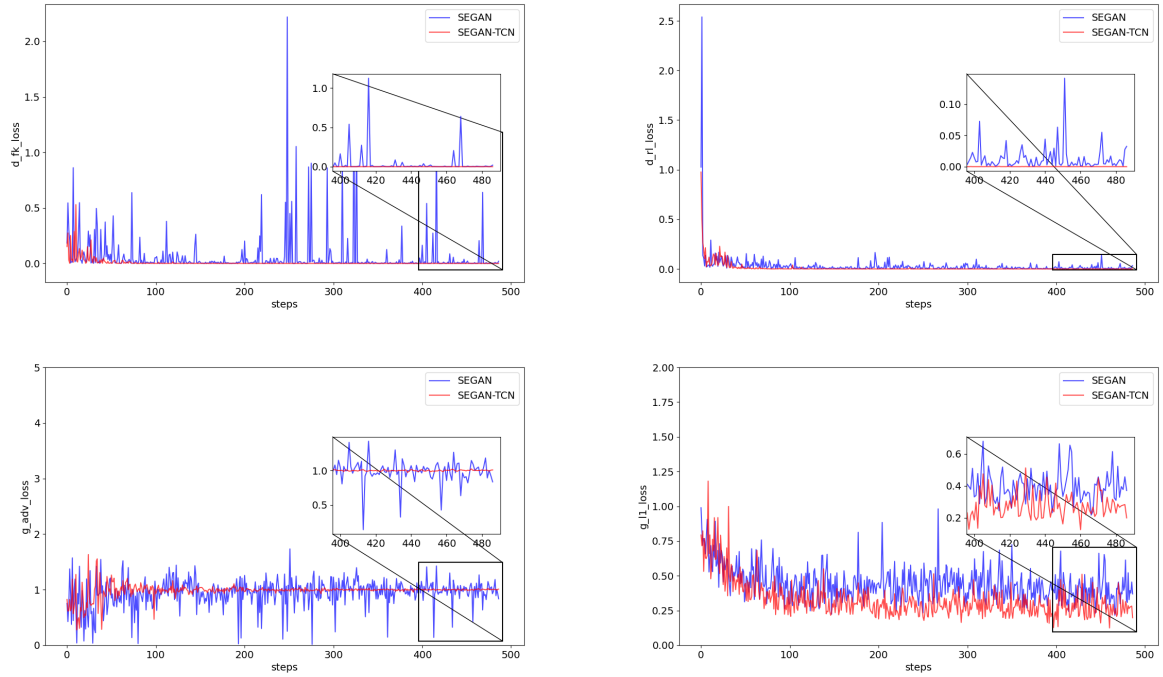

Figure 6. The loss curve of SEGAN and SEGAN-TCN on the Valentini dataset.

In order to further verify the generalization and effectiveness of the network, we will continue to conduct experiments based on the SASEGAN model. It can be seen from Table 3 that SASEGAN-TCN achieves 2.1636, 3.4132, 2.8272, 2.7631 and 92.78% on PESQ, CSIG, CBAK, COVL, and STOI on the Valentini dataset, and compared with the noise data, it's improved by 9.83, 1.9, 15.9, 5.1 and 0.7% besides the SSNR, which is improved by 4.4907 db. Data analysis reveals that the SASEGAN-TCN model has good performance in CSIG indicators, but it will reduce the quality of the speech signal when processing speech signals and the introduction of external noise will lead to a slight reduction in PESQ, CBAK, SSNR and other indicators. To effectively confront and resolve these issues, we will continue to conduct experiments and research analysis.

Table 3. SASEGAN and SASEGAN-TCN experimental results of on the Valentini dataset.

	PESQ	CSIG	CBAK	COVL	SSNR	STOI
NOISY	1.97	3.35	2.44	2.63	1.68	92.11
SASEGAN [30]	**2.2027**	3.3331	**2.9883**	2.7441	**8.3832**	92.56
SASEGAN-TCN	2.1636	**3.4132**	2.8272	**2.7631**	6.1707	**92.78**

As can be seen from Figure 7, we can clearly see that during the training phase, the SASEGAN-TCN model not only successfully fits to the optimal state, but also exhibits more stable loss curves

compared to the SASEGAN model. This strongly confirms the higher stability and easier convergence of SASEGAN-TCN during the training process. This result further emphasizes the superiority of the model in processing training data. The reduction in discriminator loss (d_fk_loss, d_rl_loss) indicates an improvement in the recognition of false and true samples. Lower g_adv_loss indicates successful generator deception, while lower g_l1_loss represents pixel level similarity between generated samples and real samples.

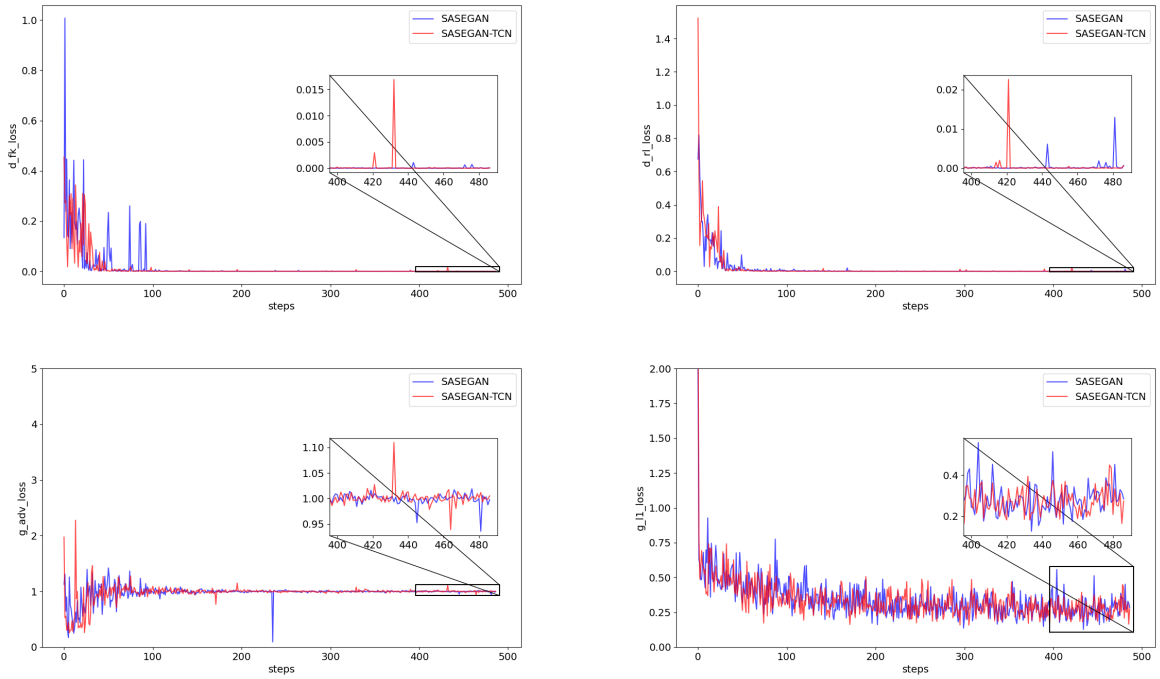

Figure 7. The loss curve of SASEGAN and SASEGAN-TCN on the Valentini dataset.

To tackle the issue that the SASEGAN model will reduce the quality of the speech signal and introduce external noise when processing Valentini data, this article will once again verify the effectiveness and applicability of the network on the THCHS30 Chinese dataset based on the SASEGAN model. The experimental results are shown in Table 4. PESQ, CSIG, CBAK, COVL, and STOI can reach 1.8077, 2.9350, 2.4360, 2.3009 and 83.54%, and the SSNR increases to 4.6332 db. After analyzing the experimental data, it can be seen that the SSNR in the SASEGAN model is higher, while the PESQ and STOI are lower, which proves that the SASEGAN model introduces additional noise during the training process and results in signal distortion. Nevertheless, the SASEGAN-TCN model proposed in this article not only ensures that SSNR does not attenuate too more, but also effectively improves PESQ and STOI levels.

Table 4. SASEGAN and SASEGAN-TCN experimental results of on the THCHS30 dataset.

	PESQ	CSIG	CBAK	COVL	SSNR	STOI
NOISY	1.3969	2.3402	1.9411	1.78	1.3101	80.33
SASEGAN [30]	1.7212	2.8051	2.3813	2.1815	**4.9159**	83.07
SASEGAN-TCN	**1.8077**	**2.9350**	**2.4360**	**2.3009**	4.6332	**83.54**

During the training phase, the training loss graphs of the SASEGAN and SASEGAN-TCN models on the THCHS30 dataset are shown in Figure 8. The SASEGAN-TCN model is still very stable and can achieve better fitting results than other models during the training process, which indicates that the model in this paper improved the discriminator's ability to distinguish between false and true samples, and also enhanced the generator's ability to generate false samples that are extremely similar to true samples. Through relevant experiments, it has been shown that there are also some problems that we should notice. Specifically, the integration of the TCN module increases the number of model parameters, which in turn requires higher experimental hardware costs. In addition, it has been experimentally proven that the model presented in this paper performs well in processing long speech data, while there may be poor performance in processing short speech data.

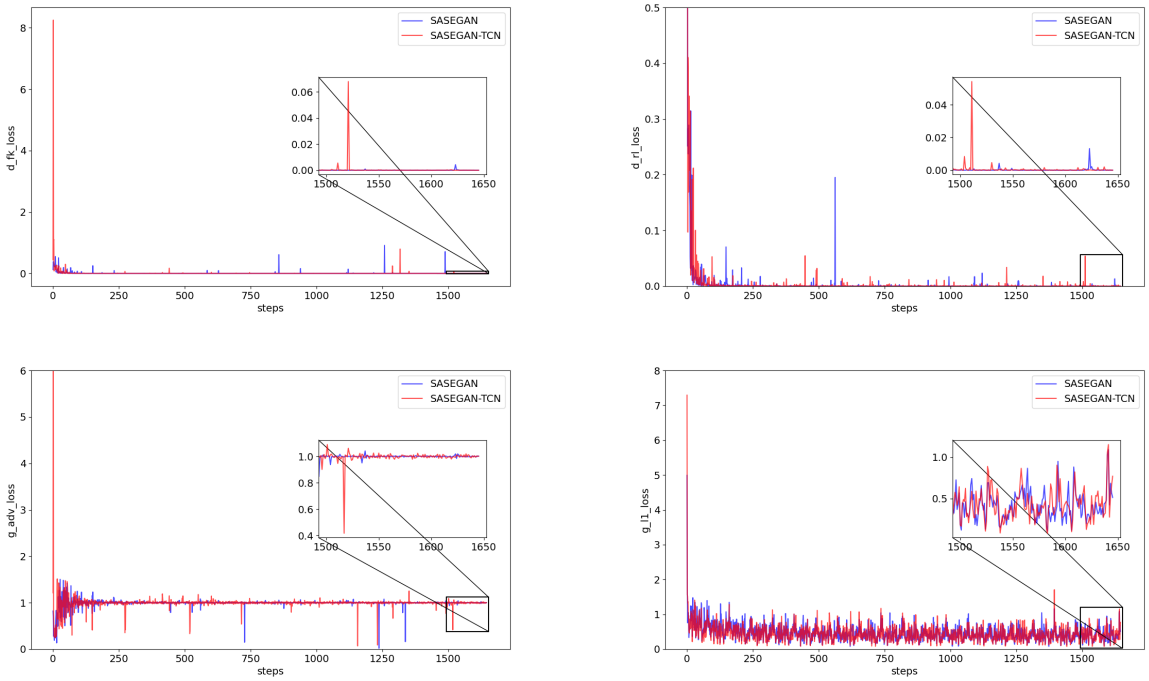

Figure 8. The loss curve of SASEGAN and SASEGAN-TCN on the THCHS30 dataset.

To sum up, this article verifies the recognition effect of enhanced audio data in the field of speech recognition technology. First, this article will use the last five saved model parameters during the SASEGAN-TCN model training process for testing and will obtain enhanced audio data corresponding to the five model parameters. Second, the test output speech data is used for a multi-core two dimensional causal convolution fusion network with attention mechanism for end-to-end speech recognition (ASKCC-DCNN-CTC) model [33] testing. The recognition results are shown in Table 5. The model proposed in this article obviously improves the quality and intelligibility of speech signals and significantly reduces the recognition error rate in speech processing technology.

Table 5. Identification results.

Type	Test wer
Noisy audio data	60.8189
First	50.9427
Second	51.5100
Third	51.3780
Fourth	52.5014
Fifth	**50.2238**
Average	51.3112

5. Conclusions

To enhance the quality and intelligibility of speech signals effectively, this paper analyzed the characteristics of the TCN network and used modules such as multilayer convolution layers, dilated causal convolution, and residual connections in the TCN network to effectively avoid problems like gradient vanishing. Moreover, the feature information between the encoder and decoder is also aggregated, thereby improving the performance and feature expression ability of network speech enhancement. Experimental results show that the proposed model has very obvious improvement on the Valentini and THCHS30 datasets, and exhibits a certain stability during the training process. In addition, we used the enhanced speech data in speech recognition technology, and the word recognition error rate is reduced by 17.4% compared with the original noisy audio data. The above content indicates that the SASEGAN-TCN model used the characteristics of the TCN network to solve the problem of non-aggregation, improved the model's speech enhancement performance and feature expression capabilities, and effectively elevated the quality and intelligibility of noisy speech data. Additionally, the speech recognition scheme proposed in this article can still maintain high recognition accuracy in noisy environments.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC, No. 61702320).

Conflict of interest

There are no conflicts of interest to report in this study.

References

1. A. R. Yuliani, M. F. Amri, E. Suryawati, A. Ramdan, H. F. Pardede, Speech enhancement using deep learning methods: A review, *J. Elektron. Telekomunikasi*, **21** (2021), 19–26. http://dx.doi.org/10.14203/jet.v21.19-26
2. D. Skariah, J. Thomas, Review of speech enhancement methods using generative adversarial networks, in *2023 International Conference on Control, Communication and Computing (ICCC)*, Thiruvananthapuram, India, (2023), 1–4. https://doi.org/10.1109/ICCC57789.2023.10164848
3. S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoust. Speech Signal Process.*, **27** (1979), 113–120. https://doi.org/ 10.1109/TASSP.1979.1163209
4. Y. Ephraim, H. L. Van Trees, A signal subspace approach for speech enhancement, *IEEE Trans. Speech Audio Process.*, **3** (1995), 251–266. https://doi.org/10.1109/89.397090
5. D. S. Richards, VLSI median filters, *IEEE Trans. Acoust. Speech Signal Process.*, **38** (1990), 145–153. https://doi.org/10.1109/29.45627
6. D. Burshtein, S. Gannot, Speech enhancement using a mixture-maximum model, *IEEE Trans. Speech Audio Process.*, **10** (2002), 341–351. https://doi.org/10.1109/TSA.2002.803420
7. B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, et al., Adaptive noise cancelling: Principles and applicationsl, *Proc. IEEE*, **63** (1975), 1692–1716. https://doi.org/ 10.1109/PROC.1975.10036
8. Y. Xu, J. Du, L. Dai, C. Lee, A regression approach to speech enhancement based on deep neural networks, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **23** (2014), 7–19. https://doi.org/10.1109/TASLP.2014.2364452
9. D. Michelsanti, Z. Tan, S. Zhang, Y. Xu, M. Yu, et al., An overview of deep-learning-based audio-visual speech enhancement and separation, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 1368–1396. https://doi.org/10.1109/TASLP.2021.3066303
10. M. Gutiérrez-Muñoz, M. Coto-Jiménez, An experimental study on speech enhancement based on a combination of wavelets and deep learning, *Computation*, **10** (2022), 102. https://doi.org/10.3390/computation10060102
11. T. Yadava, B. G. Nagaraja, H. S. Jayanna, A spatial procedure to spectral subtraction for speech enhancement, *Multimedia Tools Appl.*, **81** (2022), 23633–23647. https://doi.org/10.1007/s11042-022-12152-3
12. L. Chai, J. Du, Q. Liu, C. Lee, A cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **29** (2020), 106–117. https://doi.org/10.1109/TASLP.2020.3036783
13. E. M. Grais, M. U. Sen, H. Erdogan, Deep neural networks for single channel source separation, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, (2014), 3734–3738. https://doi.org/10.1109/ICASSP.2014.6854299

14. M. Strake, B. Defraene, K. Fluyt, W. Tirry, T. Fingscheidt, Fully convolutional recurrent networks for speech enhancement, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, (2020), 6674–6678. https://doi.org/10.1109/ICASSP40776.2020.9054230

15. J. Cole, F. Mohammadzadeh, C. Bollinger, T. Latif, A. Bozkurt, E Lobaton, A study on motion mode identification for cyborg roaches, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, (2017), 2652–2656. https://doi.org/10.1109/ICASSP.2017.7952637

16. T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. Mohamed, G. Dahl, et al., Deep convolutional neural networks for large-scale speech tasks, *Neural Networks*, **64** (2015), 39–48. https://doi.org/10.1016/j.neunet.2014.08.005

17. H. S. Choi, J. Kim, J. Huh, A. Kim, J. Ha, K Lee, Phase-aware speech enhancement with deep complex U-net, preprint, arXiv:1903.03107. https://doi.org/10.48550/arXiv.1903.03107

18. W. Chan, N. Jaitly, Q. Le, O. Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, (2016), 4960–4964. https://doi.org/10.1109/ICASSP.2016.7472621

19. T. A. Hsieh, H. M. Wang, X. Lu, Y. Tsao, WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement, *IEEE Signal Processing Lett.*, **27** (2020), 2149–2153. https://doi.org/ 10.1109/LSP.2020.3040693

20. Y. Lu, J. Zhou, M. Xu, A biologically inspired low energy clustering method for large scale wireless sensor networks, in *2019 IEEE International Conference of Intelligent Applied Systems on Engineering (ICIASE)*, Fuzhou, China, (2019), 20–23. https://doi.org/10.1109/ICIASE45644.2019.9074047

21. M. A. Kramer, Autoassociative neural networks, *Comput. Chem. Eng.*, **16** (1992), 313–328. https://doi.org/10.1016/0098-1354(92)80051-A

22. I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, et al., Generative Adversarial Networks, preprint, arXiv:1406.2661. https://doi.org/10.48550/arXiv.1406.2661

23. S. Pascual, A. Bonafonte, J. Serra, SEGAN: Speech enhancement generative adversarial network, preprint, arXiv:1703.09452. https://doi.org/10.48550/arXiv.1703.09452

24. Z. Zhang, C. Deng, Y. Shen, D. S. Williamson, Y. Sha, Y. Zhang, et al., On loss functions and recurrency training for GAN-based speech enhancement systems, preprint, arXiv:2007.14974. https://doi.org/10.48550/arXiv.2007.14974

25. M. H. Soni, N. Shah, H. A. Patil, Time-frequency masking-based speech enhancement using generative adversarial network, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, (2018), 5039–5043. https://doi.org/10.1109/ICASSP.2018.8462068

26. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, preprint, arXiv:1511.06434. https://doi.org/10.48550/arXiv.1511.06434

27. X. Hao, C. Shan, Y. Xu, S. Sun, L. Xie, An attention-based neural network approach for single channel speech enhancement, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, (2019), 6895–6899. https://doi.org/10.1109/ICASSP.2019.8683169

28. A. Pandey, D. Wang, Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, (2020), 6629–6633. https://doi.org/10.1109/ICASSP40776.2020.9054536

29. S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, preprint, arXiv:1803.01271. https://doi.org/10.48550/arXiv.1803.01271

30. H. Phan, H. Le Nguyen, O. Y. Chén, P. Koch, N. Q. Duong, et al, Self-attention generative adversarial network for speech enhancement, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, (2021), 7103–7107. https://doi.org/10.1109/ICASSP39728.2021.9414265

31. C. V. Botinhao, X. Wang, S. Takaki, J. Yamagishi, Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech, in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop*, Sunnyvale, USA, (2016), 146–152. https://doi.org/10.21437/SSW.2016-24

32. D. Wang, X. Zhang, THCHS-30: A free Chinese speech corpus, preprint, arXiv:1512.01882. https://doi.org/10.48550/arXiv.1512.01882

33. R. Lv, N. Chen, S. Cheng, G. Fan, L. Rao, X. Song, et al, ASKCC-DCNN-CTC: A Multi-Core Two Dimensional Causal Convolution Fusion Network with Attention Mechanism for End-to-End Speech Recognition, in *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Rio de Janeiro, Brazil, (2023), 1490–1495. https://doi.org/10.1109/CSCWD57460.2023.10151993

AIMS Press