Integration of Depthwise Separable CNN and Seq2Seq for Enhanced Chinese TTS Systems

Huanxin Dou, Zhenhua Zhao*
Bohai University, Jinzhou 121000, China
Email: nkzn89@163.com, zzhbhu2023@163.com
*Corresponding author

Keywords: convolutional neural network, speech synthesis, sequence-to-sequence model, deep separable convolution, highway structure

Received: June 16, 2025

Speech synthesis technology has become increasingly common in real-time broadcasting systems, smartphone voice assistants, and other applications thanks to the advancement of artificial intelligence technology. However, Chinese speech synthesis still faces issues such as the influence of Chinese tones, the impact of polyphonic characters on synthesis results, and the naturalness of tones. Therefore, this study designed a novel Chinese text to speech system that integrates sequence to sequence models and convolutional neural networks. Design multiple modules in the system to address challenges such as Chinese polyphonic characters and tones. Meanwhile, the convolutional neural network framework for sequence-to-sequence models adopts a hybrid architecture of depthwise separable convolution and highway network. Three aspects were optimized, including depthwise separable convolution to reduce parameter count, highway network to maintain gradient flow, and causal convolution to constrain temporal dependencies. The results show that compared to mainstream text to speech models, the proposed model has a Mel frequency cepstral distortion of 4.5288dB, an average opinion score of 4.15, a parameter count of 68.4487 × 106, and a training time of 16.57 hours. This system can ensure the naturalness of speech synthesis and improve synthesis efficiency. This study provides an efficient solution for the practical application of Chinese text to speech, suitable for scenarios with limited computing resources. Its lightweight design concept has guiding significance for the development of low-power speech synthesis technology.

Povzetek: Študija predstavi lahek TTS za kitajščino, ki s konvolucijsko arhitekturo ter moduli za tone in polifonijo izboljša naravnost govora in energetsko učinkovitost.

1 Introduction

Speech synthesis technology has been used more and more frequently in daily life recently thanks to the advancement of artificial intelligence technologies, including intelligent audio and navigation [1]. Speech synthesis is a technique for implementing Text-to-Speech (TTS), which often employs a sequence to sequence (Seq2Seq) framework. Seq2Seq represents the transformation from one sequence to another, which can avoid the occurrence of mismatched input and output sequence lengths [2]. Sequence to sequence Chinese TTS systems leveraging deep learning have made remarkable progress in synthesizing naturalness as a result of the rapid advancements in neural networks and deep learning. In current TTS tasks, Convolutional Neural Networks (CNN) are widely used to extract time-domain and frequency-domain features from input speech signals, and model these features through neural networks to generate synthesized speech. The advantage of CNN lies in its ability to automatically learn the features of input speech signals through multi-layer convolution operations, without the need for manually designing complex feature extractors. By gradually abstracting through convolutional layers, CNN can extract more representative and hierarchical information, significantly improving the quality and efficiency of speech synthesis in TTS systems, and demonstrating good adaptability in modeling complex Chinese syllables and intonations. However, in Chinese TTS, due to the influence of polyphonic characters and the complexity of intonation, models often need to handle a large amount of computation and storage [3]. To address the issue of the influence of polyphonic characters and the complexity of intonation in Chinese, this study optimized the Seq2Seq model using an introduced attention mechanism and developed an innovative Chinese TTS system. Simultaneously using depthwise separable convolution (DSC) and highway structure optimized CNN to solve the problem of large quantity and high complexity in Chinese TTS, thereby improving the quality, naturalness, and efficiency of speech synthesis. The core contributions of this study are reflected in the following aspects: (1) A multi module Chinese TTS system was designed to solve the problems of Chinese polyphonic characters and tones. (2) A Seq2Seq model incorporating scaled dot product attention has been designed to improve the quality of speech generation. (3) An integrated framework combining DSC and highway networks has been proposed, achieving a balance between parameter quantity and computational efficiency.

The research content mainly includes four sections. In section 2, the research status of TTS technology and Seq2Seq model at home and abroad was reviewed. In section 3, the optimization of Chinese TTS system based on Seq2Seq model and CNN is introduced. In section 3.1, the Chinese TTS System was first designed. Section 3.2 explains the optimization of Seq2Seq model based on attention mechanism. A CNN framework combining DSC and highway Network was built in section 3.3. Section 4 is the result analysis of the Chinese TTS system based on Seq2Seq model and CNN. Section 4.1 is the performance analysis of the optimized CNN framework, and Section 4.2 is the result analysis of the Chinese TTS system. Finally, the conclusion of the Chinese TTS system based on improved CNN and Seq2Seq model is drawn.

2 Related works

With the improvement of computer performance, Chinese TTS technology has been improving. Additionally, as deep learning and neural networks are constantly evolving, the combination of these technologies with Chinese TTS technology is progressively emerging as a new area of study in the field of Chinese TTS. Du's team focused on the high fidelity and speaker similarity of TTS speaker adaptation. By introducing PCA dimensionality reduction, regularization, and timbre normalization vector quantization acoustic features, and combining k-means quantization technology, they optimized the timbre independent style embedding trained together with the acoustic model. The results showed that this method performed better than existing methods in multi speaker text to speech synthesis [4]. Jiang et al. proposed a semantic dependency based self attention mechanism to improve the performance of TTS, and introduced one-dimensional CNN to better integrate local contextual information. The results showed that this method performed well in TTS [5]. Tan and other scholars have developed a variational autoencoder system for quality judgment in text to speech technology, which is used to generate different waveforms, avoid complex speech, and enhance text prior ability. The results show that this technology achieves generation quality without statistical differences [6]. Li et al. built a text-to-speech conversion model incorporating a transformer to limit the maximum length of audio. The results show that the model is feasible [7]. Deng and Lam scholars built a graph to sequence conversion model of graph neural network for graph to sequence learning and implemented communication between two remote nodes, the results show that the bilingual evaluation study score of the method improved by at least 1 point [8]. Han et al. proposed a one shot multi speaker TTS system that integrates a speaker encoder, FastSpeech2 acoustic model, and HiFi GAN encoder to synthesize personalized speech without encountering the target speaker's speech. The results indicate that the proposed model outperforms the baseline model in terms of naturalness and speaker similarity [9].

As a universal structure, Seq2Seq can solve complex speech problems and is widely used in the fields of Chinese TTS and speech recognition. Nazir et al. designed a Chinese TTS method that combines gated loop units and Seq2Seq to generate real-time speech for the robustness of system speech recognition. The results showed that the Mean Opinion Score (MOS) of this method was 4.02 [10]. Fujita and other professionals proposed a speaker embedding method based on speech rhythm to achieve multi speaker prosody modeling. The method combines rhythm embedding with acoustic features and generates target speaker style speech through attention mechanism. The results showed that the error rate of this method was reduced to 15.2% [11]. To solve the problem of multi response sorting in service conversations and improve the accuracy of Seq2Seq models, Widiyanto et al. proposed a CNN enhanced sequence to sequence model, which achieved a sorting accuracy of 86.7% by preprocessing chat history data such as standardized cleaning, conversation pair alignment, and word vector encoding [12]. Amin et al. used ResNet-18 and bi-directional long and short-term memory (BiLSTM) to extract spatiotemporal depth features and process long-range dependencies in order to achieve acoustic feature temporal modeling. The results showed a recognition accuracy of 90% on gait datasets [13]. Yang et al. addressed the limitations of traditional TTS technology in noisy environments by perceiving acoustic signals and mechanical motion through a mixed mode, enabling the artificial throat to accurately recognize speech elements such as phonemes, tones, and words at low frequencies. The results showed that this method achieved recognition accuracy of over 90% [14]. Bao et al. adapted the pre trained bidirectional Transformer encoder to the Seq2Seq task using a unified modeling method and a carefully designed self attention mask. The results showed that the method exhibited strong performance, effectiveness, and scalability in different tasks and languages [15].

In summary, although Chinese TTS technology and Seq2Seq model and TTS system have become the current mainstream, these single model joint training paradigms still face two major challenges in Chinese scenes, that is, polyphone disambiguation needs to rely on external language knowledge, and tone modeling is sensitive to the long-term dependence of attention mechanism. Therefore, the core objective of this study is to construct an efficient, lightweight, and highly natural Chinese TTS. To systematically achieve this goal, a Seq2Seq model based on scaled dot product attention mechanism was designed to effectively handle polyphonic characters and tone problems. Simultaneously adopting a hybrid architecture of DSC and highway network to reduce the number of model parameters and alleviate gradient problems in deep networks, thereby accelerating model convergence while maintaining sound quality.

3 Optimization of Chinese TTS system based on Seq2Seq model and CNN

The study first designed a new Chinese TTS system, which includes multiple modules designed to address the issues of Chinese polyphonic characters and tones.

Afterwards, a Seq2Seq model based on attention mechanism was built, and the CNN framework of the model was optimized from three aspects. The method flowchart of this study is shown in Figure 1.

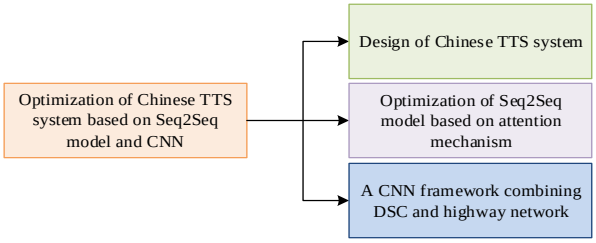

Figure 1 Method flowchart of this study

3.1 Design of Chinese TTS system

Chinese speech synthesis faces special language complexity challenges in terms of vocabulary complexity and tone characteristics. For example, Chinese has over 3500 commonly used Chinese characters, including more than 400 polyphonic characters. The pronunciation of many words may depend on the context. Different characters contain four basic tones, which increases the difficulty of Chinese TTS [16]. To overcome these issues, this study developed a Chinese TTS system. To overcome these issues, this study developed a Chinese TTS system based on Seq2Seq model. The structure of the system is shown in Figure 2.

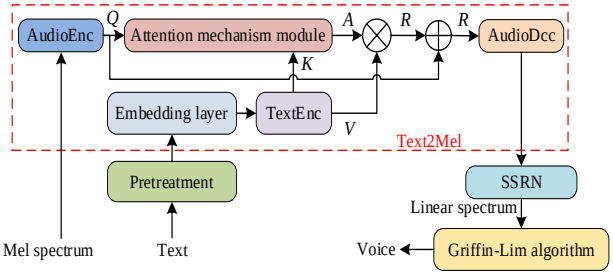

Figure 2: Chinese TTS system framework incorporating attention mechanism

In Figure 2, the construction of the Chinese TTS system can be divided into three key modules: Text2Mel module, SSRN module, and preprocessing module. The Text2 Mel module preprocesses the input text first, converting Chinese characters into pinyin as characters. The Text2Mel module consists of five parts: embedding layer, text encoding (TextEnc), audio encoding (AudioEnc), attention mechanism module and audio decoding (AudioDec). Through these modules, the system can map Chinese characters to Pinyin characters and further convert them into Mel spectrograms, ultimately achieving high-quality speech synthesis. The text encoder converts input characters into context aware representations through embedding layers and TextEnc architecture. In the embedding layer, each Chinese character is converted into a d -dimensional vector to form a matrix $K,V \in \square^{N\times d}$ representing N -length text. The TextEnc architecture is shown in equation (1), where the encoder generates two key matrices.

$$K,V = TextEnc(L) \tag{1}$$

In Equation (1), K is the key, encoding language features for attention matching. V is a value that contains pronunciation related information. e and d represent the vector dimensions. The AudioEnc module encodes the Mel spectrum $S(=S_{1:F,1:T}) \in R^{T\times F}$ of a sample of length T to form the matrix $Q \in \square^{T\times d}$, expressed as shown in Equation (2).

$$Q = AudioEnc(S_{1:F,1:T}) \tag{2}$$

In Equation (2), Q represents the query matrix of AudioEnc. T is the mel frame rate. F denotes the dimension of the Meier spectrum. The preprocessing module in the Chinese TTS system is responsible for preprocessing the input Chinese text, converting Chinese characters into pinyin. By querying the dictionary, select the corresponding pinyin for each Chinese character and provide it as input to the subsequent Chinese TTS model. Figure 3 depicts the preprocessing module's flowchart.

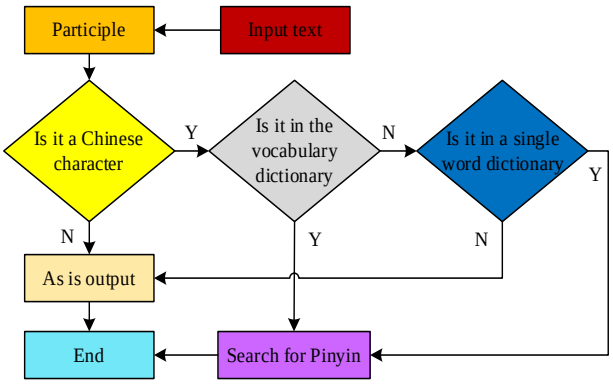

Figure 3: Preprocessing module flowchart

The input text in Figure 3 is first broken up into words. The appropriate pinyin for Chinese characters is then obtained using a word dictionary and a single-character dictionary query. The output is then used as-is for non-Chinese characters and characters that were not queried in the dictionary. The word dictionary is derived from the Chinese dictionary pinyin data of Handian.com, totaling more than 390,000 words, and the dictionary of individual Chinese characters totaling more than 40,000. In the embedding layer, hanyu pinyin with tones and common punctuation marks are used as character annotations to form the embedded encoding matrix. The system used both subjective and objective evaluation methods during the assessment. The subjective evaluation is performed using the internationally standardized MOS, and a 5-point scale is used to evaluate the audio quality. The objective evaluation used in Chinese TTS is measured by the Mel-cepstral distortion (MCD), which is measured in dB, as shown in Equation (3).

$$MCD\left(v^{t\arg},v^{ref}\right)=\frac{\alpha}{T}\sum\frac{T-1}{t=0\sqrt{\sum_{d=s}^{D}\left(v_d^{t\arg}(t)-v_d^{ref}(t)\right)^2}},\ \alpha=10\sqrt{2}/(\ln 10)=6.14185 \tag{3}$$

In Equation (3), $v^{t\arg}$ denotes the Mel Frequency Cepstrum Coefficient (MFCC) of the Chinese TTS, v^{ref} denotes the MFCC of the original speech, T is the number of speech frames, and D denotes the MFCC feature dimension.

3.2 Optimization of Seq2Seq model based on attention mechanism

The traditional Seq2Seq model has certain shortcomings for Chinese tones, polyphonic characters, word segmentation, etc. For example, the four tones and polyphonic characters in Chinese can affect the effectiveness of generating TTS, unlike English sentences that have separate representations. For the naturalness of tones, Mel spectral features are used for analysis. Among them, Mel spectral features are the characteristics obtained by converting audio signals on the Mel frequency scale, which can better simulate the characteristics of the human auditory system. The Mel frequency scale is closer to the way the human ear perceives frequency. By converting the linear frequency f to the Mel frequency f_{mel} through equation (4), the frequency resolution of the low-frequency part is higher, while the relative resolution of the high-frequency part is lower, which is in line with the characteristics of the human ear. The expression for equation (4) is as follows.

$$f_{mel}=2595\times\log_{10}\left(1+\frac{f}{700}\right) \tag{4}$$

In the Mel scale, when the Mel frequencies of two speech segments differ by a factor of 3, the human ear can perceive their pitches as differing by approximately the same factor (3×). The process of Mel spectrum extraction is shown in Figure 4.

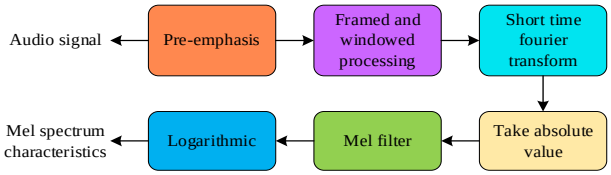

Figure 4: Mel spectrum extraction flowchart

In Figure 4, the audio signal is first pre emphasized, followed by frame segmentation and windowing. The short-time Fourier transform is used to capture the characteristics and take the absolute value. The logarithm is used to calculate the Mel specteum characteristics, which are then utilized to imitate the auditory properties of the human ear. Pre-emphasis specifically refers to compensating for the natural spectral tilt in speech signals by enhancing high frequencies. Due to the -6dB/octave roll off characteristic of glottal excitation, high frequencies typically have lower energy [17]. This process enhances the high-frequency components that are crucial for speech intelligibility. To improve the naturalness of speech synthesis, an attention mechanism based on scaled dot product attention is adopted to calculate the alignment weights between text features and audio features. This mechanism can calculate similarity based on the features of the input text and the current audio features, adjust the attention level to different text parts when generating speech, and enhance the fluency and naturalness of the speech. Specifically, in the attention mechanism module, the similarity between the text features of the n th character and the speech features at the t moment is calculated using the scaled dot product calculation method to obtain the attention matrix $A\in\square^{N\times T}$. This matrix A is used to adjust the model's focus on different text parts when generating speech, making the generated speech more natural. The expression of A is shown in Equation (5).

$$A=soft\max\left(\frac{K^TQ}{\sqrt{d}}\right) \tag{5}$$

In Equation (5), d represents the vector dimension of the encoding module. The result is scaled by dividing by $\sqrt{d}$ to prevent the dot product value from being too large when there is a large d, and to avoid falling into the region of very small gradient after the soft max function calculation. Next, the text-encoded V and the attention matrix A are multiplied to obtain $R^{T\times d}$ and the previous Q are spliced to obtain the input of the decoding module $R'\in\square^{T\times 2d}$. The process of extracting context is shown in equation (6).

$$R=AV=soft\max\left(\frac{K^TQ}{\sqrt{d}}\right),R'=[R,Q] \tag{6}$$

In equation (6), the context vector R is the result of the weighted sum of the value vector V with attention matrix A, which aggregates the most relevant information from all text characters to the current audio frame. Subsequently, R is concatenated with the current query vector Q to form the input R' of the decoder module, thereby injecting attention information into the decoding process. In AudioDec module, R' is input to the decoding network to get the predicted Mel spectrum $\hat{Y}\in\square^{F\times T}$, which is expressed as shown in Equation (7).

$$\hat{Y}=AudioDec\left(R'\right) \tag{7}$$

The AudioDec module finally converts the generated Mel spectrogram into a linear spectrum in the SSRN module, thereby generating the final speech signal. Specifically, in the SSRN module, the Meier spectrum $\hat{Y}\in\square^{F\times T}$ is converted into a linear spectrum $Z\in R^{F0\times 4T}$, F_0 denotes the dimension of the linear spectrum, the sampling from F to F_0 is achieved by increasing the number of one-dimensional convolution channels, and the up sampling from T to $4T$ in the time domain is achieved by one-dimensional deconvolution. This can achieve the conversion from frequency domain to time domain, generating playable audio signals. The expression for this process is shown in Equation (8).

$$Z = SSRN(Y) \tag{8}$$

The function of the SSRN module is to convert the Mel frequency spectrum into a linear spectrum, and achieve the conversion from frequency domain to time domain through time-domain deconvolution and one-dimensional convolution, ultimately generating audio signals that can be played. The final generated audio signal is transformed into waveform form using the Griffin-Lim algorithm [18]. The Griffin-Lim algorithm can avoid phase information loss in systems using Mel spectrogram as input. It updates the phase through multiple iterations until the difference between the recovered audio signal spectrogram and the original spectrogram is minimized. In the Chinese TTS system, to facilitate machine computation of natural language, it is necessary to convert ordinary text sentences into vectors before inputting them into the network for training. Among them, word vector is a form of word representation, and one hot text feature representation method is often used to convert words from symbolic form to vector form. The pseudocode of the scaled dot product attention mechanism is shown in Table 1.

Table 1: Pseudo code of the scaled dot product attention mechanism

Scaled Dot-Product Attention
Input: $K, V \in \mathbb{R}^{N\times d}$, $Q \in \mathbb{R}^{T\times d}$
Output: Z
1: $K, V = TextEnc(L)$
2: $Q = AudioEnc(S_{1:F,1:T})$
3: $A = soft\max\left(\frac{K^T Q}{\sqrt{d}}\right)$
4: $R = AV = soft\max\left(\frac{K^T Q}{\sqrt{d}}\right)$
5: $R' = [R, Q]$
6: $\hat{Y} = AudioDec(R')$
7: $Z = SSRN(Y)$

3.3 3.3 A CNN framework combining DSC and highway network

In the Seq2Seq model, both the encoder and decoder adopt a CNN network structure. Traditional CNN faces problems such as parameter explosion, high time complexity of standard convolution operations, and unstable gradients in Seq2Seq models. Specifically, it refers to the sharp increase in parameter count and the tendency for gradient vanishing or explosion when the number of network layers increases. Therefore, this study proposes a DSC-highway hybrid architecture, which solves the above problems by reducing the number of parameters through DSC, maintaining gradient flow through highway network, and constraining temporal dependence through causal convolution.

To solve the problems of excessively deep network layers, large model parameters, high computational complexity, and slow training, the study introduced the DSC structure to replace traditional convolution. DSC consists of two parts: channel wise convolution and Point-by-Point Convolution (PbP-C). Channel wise convolution means that each convolution kernel convolves only one channel, reducing computational complexity. PbP-C performs weighted summation in the depth direction, further reducing computational complexity. The schematic diagram of PbP-C processing is shown in Figure 5.

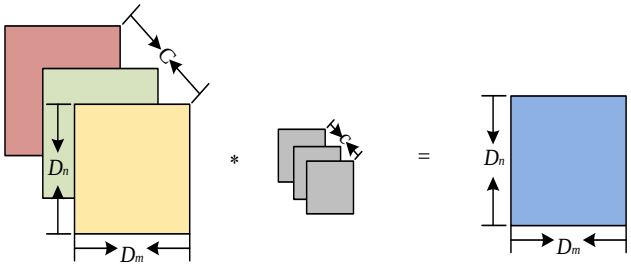

Figure 5: Schematic diagram of PbP-C

In Figure 5, the feature image of input size $D_a \times D_b \times C$ is first convolved channel-by-channel with a CK of size $K \times K \times C$ and the output feature image of size $D_m \times D_n \times C$, so the number of parameters of the channel-by-channel convolution layer is $(K \times K \times 1) \times C$. The PbP-C is then performed, where the size of the CK is $(1 \times 1 \times C)$ and the total number of such kernels is N, so the number of parameters in the PbP-C layer is $(1 \times 1 \times C) \times N$. The final parametric number of the depth-separable convolution layer is obtained by adding the depth convolution and PbP-C parametric numbers, and the expression is shown in Equation (9).

$$(K \times K) \times C + C \times N \tag{9}$$

in Equation (9), i.e., the ratio of depth-separable convolution to the number of conventional convolutional parameters, as shown in Equation (10).

$$\frac{(K \times K) \times C + C \times N}{(K \times K \times C) \times N} = \frac{1}{N} + \frac{1}{K^2} \tag{10}$$

Highway network divides the input data into two parts through a learnable threshold mechanism. Part of it is processed through convolutional layers, while the other part is passed directly. By using a gating mechanism to determine the weight of each part of information, efficient transmission of information can be achieved. The DSC-highway structure is shown in Figure 6.

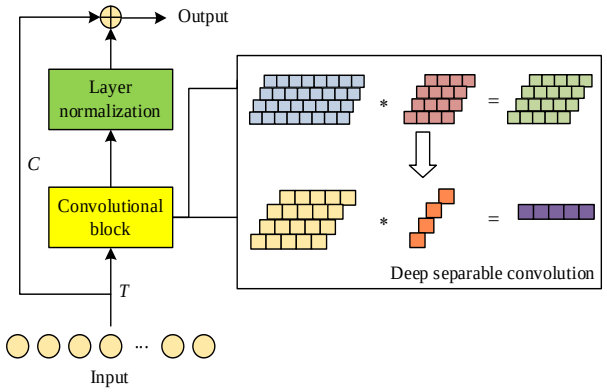

Figure 6: DSC-highway network structure diagram

In Figure 6, a deeply separable convolutional operation is performed on a portion of the input data using a learnable threshold mechanism. One part of the information is transformed through the convolutional layer, and the other part is passed directly without any transformation by passing the Carry gate C_{carry} and the transformation gate T to determine how much information is in each part. To speed up the convergence of the model, the two parts are finally summed and output. After the convolution operation, to accelerate the convergence speed of the model, a layer normalization mechanism was introduced in the study. Normalization of layers helps maintain the stability of data distribution during the training process, further optimizing the performance of the network. Equation (11), which depicts the highway structure's forward propagation.

$$y = x \bullet C_{\text{carry}}\left(x, W_{C_{\text{carry}}}\right) + H\left(x, W_H\right) \bullet T\left(x, W_H\right) \tag{11}$$

In Equation (11), y indicates the output, x indicates the input, H indicates the nonlinear function, and W_H indicates the weights of the network. Where $C_{\text{carry}} = 1 - T$, as shown in Equation (12).

$$y = H\left(x, W_H\right) \bullet T\left(x, W_H\right) + x \bullet \left(1 - T\left(x, W_{C_{\text{carry}}}\right)\right) \tag{12}$$

In Equation (12), the dimensions of x , y , , and $H\left(x, W_H\right)$ $T\left(x, W_T\right)$ are the same, i.e., as shown in Equation (13).

$$y = \begin{cases} x, T\left(x, W_T\right) = 0, \\ H\left(x, W_H\right), T\left(x, W_T\right) = 1. \end{cases} \tag{13}$$

The Jacobi transformation of this layer, as shown in Equation (14).

$$\frac{dy}{dx} = \begin{cases} 0 = 1, T\left(x, W_T\right) \\ 1 = H'\left(x, W_H\right), T\left(x, W_T\right) \end{cases} \tag{14}$$

A part of the data is processed, a part is passed directly, and the final output, as shown in Equation (15).

$$y_i = x_i * \left(-T_i\left(x\right) + 1\right) + H_i\left(x\right) * T_i\left(x\right) \tag{15}$$

In Equation (15), i indicates the final number of output layers. To prevent data leakage in the future, a convolution kernel (CK) $F = \left(f_1, f_2, \ldots, f_K\right)$ is used with an input sequence of $X = \left(x_1, x_2, \ldots, x_T\right)$, and the convolution definition expression at x_t is shown in equation (16).

$$\sum_{K=1}^{K} f_k x_{t+K-K} = \left(F * X\right)_{x_t} \tag{16}$$

A temporal processing method has been proposed for causal convolution constraints. In Figure 7, a schematic representation of the one-dimensional convolution in Equation (17) is presented.

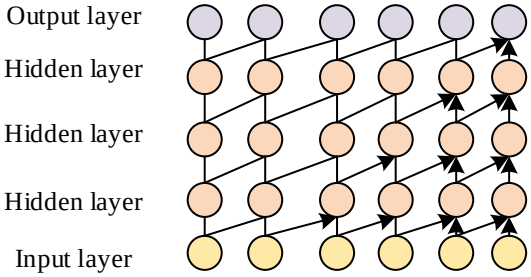

Figure 7 Schematic diagram of causal CNN

The CK of size 2 is consecutively shifted in Figure 5's input layer for convolution, and the output obtained is utilized as the input of the subsequent layer. The value of each node is only related to the nodes that have come before it, not the nodes that will come after. The closer the node is to the top layer, the larger its receptive field is. The formula for the receptive field RF_i of the first i layer is shown in Equation (17).

$$RF_i = RF_{i-1} + \left(k - 1\right)s \tag{17}$$

In Equation (17), i represents the number of layers, RF_{i-1} represents the receptive field of the previous layer, k represents the CK size, and s is the convolution step size. Therefore, the DSC-highway hybrid architecture significantly reduces the number of parameters through depthwise separable convolutions of equations (9) and (10). The stable propagation of gradient flow was achieved through the gating mechanism of equations (11) - (15), and finally the causal convolution of equations (16) and (17) ensured the causal constraints of the model in time. These equations collectively define an efficient and stable forward propagation process. The pseudocode of DSC-highway network is shown in Table 2.

Table 2: Pseudocode of DSC-highway network

DSC-highway network
Input: $X = \left(x_1, x_2, \ldots, x_t\right)$
Output: y_i
1: $y = x \bullet C_{\text{carry}}\left(x, W_{C_{\text{carry}}}\right) + H\left(x, W_H\right) \bullet T\left(x, W_H\right)$
2: $C_{\text{carry}} = 1 - T$
3: $y = H\left(x, W_H\right) \bullet T\left(x, W_H\right) + x \bullet \left(1 - T\left(x, W_{C_{\text{carry}}}\right)\right)$
4: $y = \begin{cases} x, T\left(x, W_T\right) = 0, \\ H\left(x, W_H\right), T\left(x, W_T\right) = 1. \end{cases}$

$$\frac{dy}{dx}=\begin{cases}1, T\left(x, W_T\right)=0 \\ H'\left(x, W_H\right), T\left(x, W_T\right)=1\end{cases} \tag{5}$$

$$\left(F * X\right)_{x_t}=\sum_{K=1}^{K} f_K x_{t-K+K} \tag{6}$$

$$RF_i=RF_{i-1}+\left(k-1\right)s$$

$$y_i=H_i\left(x\right) * T_i\left(x\right)+x_i *\left(1-T_i\left(x\right)\right) \tag{7}$$

Through these improvements, it is easier to accelerate the convergence speed of the model, reduce the computational burden caused by deep networks, and make the system more efficient and accurate in handling Chinese TTS tasks. The training is carried out using the Noam learning rate decay approach, and Adam is chosen as the optimizer. The purpose of Noam learning rate decay is to use a larger learning rate in the early stages of training to accelerate learning, and gradually reduce the learning rate as training progresses to avoid numerical instability in the later stages of training.

4 Result analysis of Chinese TTS system based on Seq2Seq model and CNN

4.1 Performance analysis of optimized CNN framework

The starting learning rate is 0.001. To quantify the statistical significance of performance differences, this study conducted paired t-tests on all indicators and calculated a 95% confidence interval (95% CI) for MOS. The significance level of all statistical tests is set to α=0.05. To investigate the impact of CK size on model performance, four models with CK sizes of 2, 3, 5, and 7 were set up. Using these four models to synthesize speech, the evaluation indicators are MOS, MCD, and the receptive field of the enhancement module. Among them, a total of 40 listeners participated in MOS scoring, and all listeners were trained and familiar with MOS scoring standards. The MOS rating is conducted in a controlled environment, using standardized audio equipment and ensuring that the ambient noise is below 30dB. From the test set, 30 texts were chosen at random. The effect of CK size on model performance is shown in Figure 8.

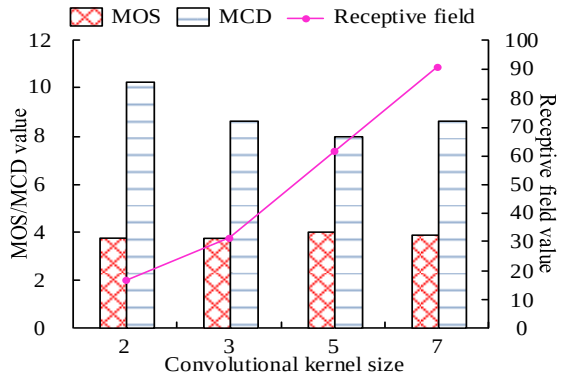

Figure 8: The Effect of CK Size on model performance

In Figure 8, the receptive field of convolution increases as CK increases. At a CK size of 5, the model MOS is 3.928±0.15 (95% CI) and the Chinese TTS quality is best. This is because a smaller CK will have a smaller receptive field and be unable to obtain longer information, while a larger CK will result in overfitting and large computation. Compared with the models of CK size 3 and CK size 57, the MCD reduction of CK size 5 was statistically significant (p<0.05). Therefore, CK size 5 is the most appropriate.

Thirty randomly selected texts from the test set were synthesized using DSC models with different channel numbers. The MOS, MCD and model parametric quantities of the Chinese TTS varied with the number of channels, as shown in Figure 9.

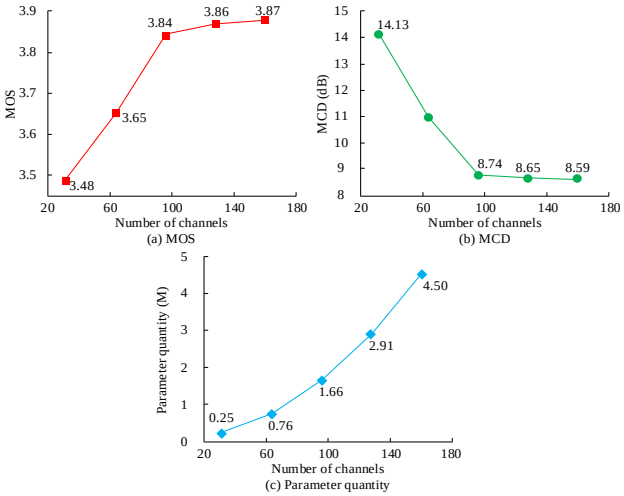

Figure 9: Comparison chart of the impact of channel number on model performance

In Figure 9(a), the Chinese TTS quality is at its lowest when the DSC model has 32 channels, and the MOS is at its lowest at this point, 3.48. When the number of channels is 160, the MOS reaches its highest value of 3.87 ± 0.11 (95% CI), indicating that the TTS quality in China is in the best state. Figure 9(b) shows that the MCD is at its highest when there are 32 channels, at 14.13 dB. When the number of channels is 160, the lowest MCD is 8.59 dB. Compared with 32 channels, both MOS enhancement and MCD reduction have high statistical significance (p<0.01). The MOS value steadily rises and the MCD value gradually falls as the number of channels rises, showing that the Chinese TTS quality improves and implying that adding more channels enables the DSC network to pick up more features. In Figure 9(c), the number of model parameters increases gradually from 0.25M to 4.50M as the number of channels increases from 32 to 160. Therefore, the number of channels is 96 to ensure the number of parameters and the quality of the Chinese TTS.

To explore the effect of having highway structure on network training, the training convergence curves at 10-, 50- and 100-layers depth were compared, as shown in Figure 10.

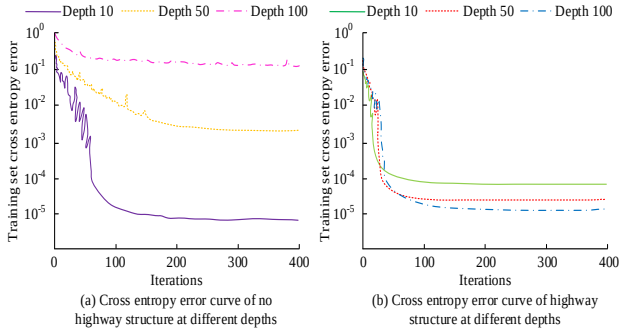

Figure 10: Comparison diagram of cross entropy error curve of training set

The cross-entropy error (CEE) curve of the training set for the network without highway structure in Figure 10(a) shows a decreasing trend as the number of iterations rises at 10-, 50-, and 100-layer depths. When the number of iterations is 200, the error curve gradually converges at the depth of 10 layers. In Figure 10(b), the curves for all three depths start to converge when the number of iterations is 50. The training set with the highway structure has a CEE of 1×10^{-4} when the depth is 10 layers. When the depth is 100 layers, the neural network with the highway structure converges more easily, and the CEE of the training set is 1×10^{-5}. This indicates that the highway structure can avoid the difficulty of training the network due to the obstruction of the gradient information backflow when the depth of the network deepens.

4.2 Analysis of the results of Chinese TTS system

The datasets selected for the experiment include the LJSpeech dataset and the LibriTTS dataset. The LJSpeech dataset is a public domain speech dataset containing 13100 short audio clips with a total duration of approximately 24 hours. Randomly select 5% of the data as the test set and the remaining 95% as the training set [19]. LibriTTS is an English multi speaker language library for TTS, originating from LibriSpeech. It has thousands of speakers (approximately 2456, gender balanced), a sampling rate of typically 24 kHz, and a total duration of approximately 585 hours [20]. In the designed system, to obtain higher quality synthesized audio, the filter group with the highest frequency $f_{\max} = 8000Hz$ and the number of filter groups $M = 24$ in the Mel spectrum extraction process is set. At the same time, to study the effectiveness of the Chinese TTS system, the study will compare the loss curves of the system before and after improvement, as shown in Figure 11.

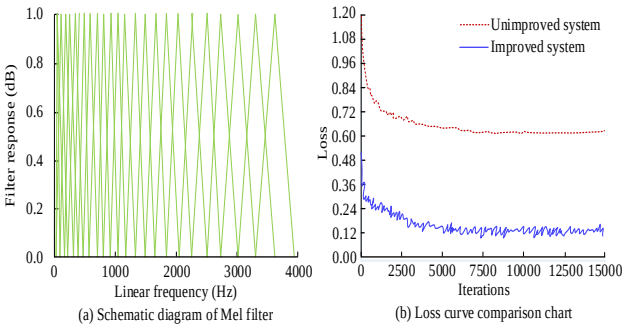

Figure 11: Improved Mel filter for system and its loss comparison

In Figure 11(a), the larger the linear frequency, the wider the weight distribution, the denser the weight distribution in the linear frequency within 1000 Hz, and the weight distribution from 1000 Hz to 4000 Hz gradually becomes wider filter response is kept at 0 to 1.0 dB. In Figure 11(b), the loss curve for the CNN model starts to converge at 10,000 iterations, and the loss value is 0.60 at this point. The loss curve for the DSC model starts to converge at 6000 iterations, and the loss value is 0.12 when the number of iterations reaches 10,000. All of the test set's sentences are used for evaluation in the objective test. The training time of the Chinese TTS system was 16.57h, MOS was 4.15, and MCD was 4.5288dB, indicating that the high-quality Chinese TTS was achieved in a short training time under the condition of limited corpus.

To verify the effectiveness of the optimized model, the accuracy of different optimized CNN models was compared. The comparison models are CNN-Bidirectional Long Short-Term Memory (BiLSTM) [21], Deep Neural Network and Two-Dimensional CNN (DNN-2DCNN) [22], and Deep Learning Speech Communication and Transmission (DeepSC-ST) [23]. The CNN BiLSTM system classifies speech signals using Mel spectrograms, short-term energy signal features, and a hybrid CNN-BiLSTM model. The DNN-2DCNN system selects DNN-2DCNN to predict speech data, and uses a neural encoder to achieve Chinese TTS in intracranial EEG signals. The DeepSC ST system identifies features through a joint semantic channel encoder and then generates speech signals in a neural network model. The comparison of accuracy obtained by different models on different datasets is shown in Figure 12.

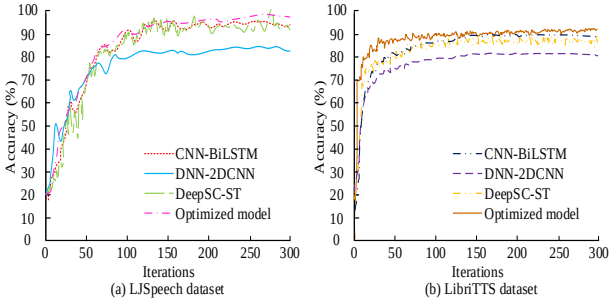

Figure 12: Comparison of accuracy obtained by different models on different datasets

In Figure 12(a), the optimized model has the highest accuracy on the LJSpeech dataset, at approximately 99.25%. In contrast, DNN-2DCNN has a lower value of approximately 82.96%. The accuracy of the optimized model is significantly higher than that of DNN-2DCNN, with a difference of 16.29%. This is because relying on the DNN-2DCNN model to predict speech data is unstable. The accuracy curve of DeepSC ST fluctuates more significantly, indicating weaker stability of the model. In Figure 12(b), CNN-BiLSTM performs better on the LibriTTS dataset, outperforming DeepSC ST by 8.50%. The highest accuracy of the improved system is 91.02%, because the optimized model improves accuracy by optimizing the network structure, solving the problems of training efficiency and performance stability.

The comparison of the training results of the system before and after improvement on different language datasets is shown in Figure 13.

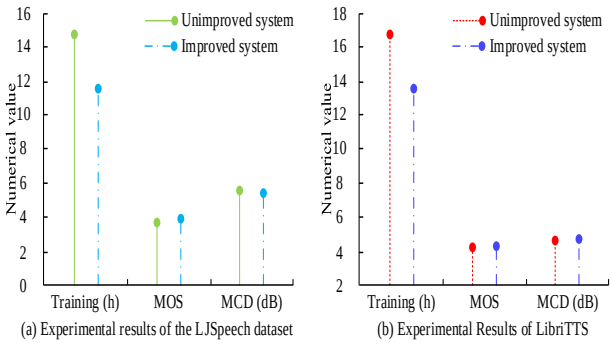

Figure 13: Comparison of system training results before and after improvement

In Figure 13(a), the improved model training time is reduced to 77.93% of the original value and the number of model parameters is decreased to 44.03% of the original value when the number of training steps is 90k at the same number of model training iteration steps. The improved system DSC slightly outperforms the CNN system in terms of synthetic speech quality. When there are 90k training steps, as shown in Figure 13(b), the improved model's training time is around 82.86% of the original, while the model's parameters are approximately 44.12% of the original. There isn't much of a difference between the two systems when comparing MOS and MCD values. In addition to cutting training time without compromising Chinese TTS quality, the improved system shrinks the size of the model, which saves storage space and computational resources while also achieving model optimization in both training time and space dimensions.

To verify the comprehensive performance of the proposed model, it was compared with the relative entropy alignment and multi-level attention fusion network (REA-MAF), FastSpeech2, Tacotron 2, and WaveGlow (TW). REA-MAF refers to the extraction of multi-level acoustic information under the action of relative entropy alignment and multi-level attention fusion network, and the complementary features between modalities to achieve information exchange [24]. FastSpeech2 refers to the ability to quickly generate high-quality synthesized speech on its own, enhancing audio reading and writing capabilities [25]. TW refers to fine-tuning and pre training the English Tacotron 2 and WaveGlow models to generate natural speech on the dataset [26]. These four models have been analyzed from different perspectives and subjected to t-tests to further validate the statistical significance of these performance differences. The test results are shown in Table 3.

Table 3: Performance results of different models

Model	REA-MAF	FastSpeech2	TW	The proposed model
MCD (dB)	4.7123	4.5891	4.8234	4.5288
MOS	4.02	4.10	3.95	4.15
Number of parameters ($\times 10^6$)	90.8431	81.8945	102.2741	68.4487
Training time (h)	20.12	18.34	21.45	16.57
FLOPs ($\times 10^{12}$)	2.20	2.00	2.50	1.75
Memory usage (GB)	8.5	7.8	9.3	6.2
Inference time (s)	0.18	0.17	0.20	0.15

In Table 3, after paired *t*-test verification, compared with the REA-MAF, FastSpeech2, and TW models, the differences in MCD, MOS, parameter quantity, and other indicators of this model reached statistical significance levels (all p-values<0.05). Specifically, the MOS score of this model is significantly higher than FastSpeech2 (p=0.032), and the MCD value is significantly lower than REA-MAF (p=0.017). Compared with the latest Transformer based models, this model still has a competitive advantage in the naturalness of synthesized speech. The MOS score is the highest at 4.15, while the MCD score is the lowest at only 4.5288 dB. This model achieves the best naturalness and accuracy in synthesized speech. The proposed model has a training time of 16.57 hours and a parameter count of 68.4487 × 10^6, which is lower than other models. Compared with the TW model, the training time is reduced by 4.88 hours and the parameter count is reduced by approximately 33.07%. The proposed model has significant optimization in computational efficiency. The FLOPs of the proposed model are 1.75 × 10^{12}, indicating better performance in terms of computational efficiency. In terms of inference time, the proposed model only takes 0.15 seconds, which is shorter than the inference time of other models. In comparison, the P value of the proposed model compared to the other three models was <0.05, indicating significant differences in all seven indicators. This is because the proposed model adopts DSC and efficient highway structure, reducing the computational complexity and

memory consumption of the network. This structure can significantly shorten the training time and reduce the number of model parameters while ensuring the quality of speech synthesis, thereby improving performance in multiple aspects.

5 Conclusion

To improve the effectiveness of Chinese TTS, this study introduced a scaled dot product attention mechanism in TTS. And a DSC highway hybrid architecture was built to solve the problems of parameter explosion and gradient instability in traditional CNN in Seq2Seq models. The results show that the training time of the Chinese TTS system is 16.57h, MOS is 4.15, and MCD is 4.5288dB. the loss curve of DSC model tends to be smooth at 10000 iterations, and the loss value is 0.12 at this time. When the DSC model has 32 channels, the MOS is the smallest at this point (3.48), and the MCD is the highest at this point (14.13), and this is when the Chinese TTS quality is at its lowest. The MOS value steadily rises and the MCD value gradually declines as the number of channels grows. The improved system slightly outperformed the original system in terms of Chinese TTS quality. When comparing the two systems' MOS and MCD values, there was minimal difference between them. The improved model's training time was roughly 82.86 percent of the original, and its model parameters were roughly 44.12 percent of the original. With more iterations, the CEE curve of the training set of the network without highway structure gets flatter at 10-, 50-, and 100-layer depths. When the number of iterations is 200, the error curve gradually converges at the depth of 10 layers. When the number of iterations is 50, the curves begin to converge for all three depths. When the depth is 10 layers, the CEE of the training set with the highway structure is 10^{-4}. When the depth is 100 layers, the neural network with the highway structure converges more easily, and the CEE of the training set is 10^{-5}. Therefore, the designed model can be applied to embedded devices such as smart speakers and car voice systems, and used as an educational aid in Chinese pronunciation evaluation systems. However, the research did not consider the real-time performance of the model, and in the future, dynamic pruning techniques can be combined to further compress the model to millisecond level response.

Acknowledge

This manuscript was supported by Project of Industry-University Cooperation and Collaborative Education of the Ministry of Education: Empowering International Chinese Language Teachers with Generative AI Technology (Project Number: 250305061134614).

References

[1] Kasparaitis P. Evaluation of Lithuanian Speech-to-Text Transcribers. Informatica, 2025, 36(2): 369-384. https://doi.org/10.15388/25-INFOR591

[2] Zheng J, Zhou J, Zheng W, Tao, L., & Kwan, H. K. Controllable Multi-Speaker Emotional Speech Synthesis with an Emotion Representation of High Generalization Capability. Affective Computing, IEEE Transactions on, 2025, 16(1):68-82. https://doi.org/10.1109/TAFFC.2024.3412152

[3] Kaur N, Singh P. Conventional and contemporary approaches used in text to speech synthesis: A review. Artificial Intelligence Review, 2023, 56(7): 5837-5880. https://doi.org/10.1007/s10462-022-10315-0

[4] Du C, Guo Y, Chen X, Chen, X., & Yu, K. Speaker adaptive text-to-speech with timbre-normalized vector-quantized feature. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 3446-3456. https://doi.org/10.1109/TASLP.2023.3308374

[5] Jiang C, Gao Y, Ng W W Y, Zhou, J., Zhong, J., & Zhen, H. SeDepTTS: Enhancing the naturalness via semantic dependency and local convolution for text-to-speech synthesis. Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 12959-12967. https://doi.org/10.1609/aaai.v37i11.26523

[6] Tan X, Chen J, Liu H, Cong, J., Zhang, C., Liu, Y., ... & Liu, T. Y. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4234-4245. https://doi.org/10.1109/TPAMI.2024.3356232

[7] Li N, Liu Y, Wu Y, Liu S, Liu M. RobuTrans: A robust transformer-based text-to-speech model. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8228-8235. https://doi.org/10.1609/aaai.v34i05.6337

[8] Deng C, Lam W. Graph transformer for graph-to-sequence learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7464-7471. https://doi.org/10.18653/v1/2020.acl-main.640

[9] Han S, Um J, Kim H. One-shot multi-speaker text-to-speech using RawNet3 speaker representation. Phonetics and Speech Sciences, 2024, 16(1): 67-76. https://doi.org/10.13064/KSSS.2024.16.1.067

[10] Nazir O, Malik A, Singh S, & Pathan, A. S. K. Multi speaker text-to-speech synthesis using generalized end-to-end loss function. Multimedia Tools and Applications, 2024, 83(24): 64205-64222. https://doi.org/10.1007/s11042-024-18121-2

[11] Fujita K, Ando A, Ijima Y. Speech rhythm-based speaker embeddings extraction from phonemes and phoneme duration for multi-speaker speech synthesis. IEICE TRANSACTIONS on Information and Systems, 2024, 107(1): 93-104. https://doi.org/10.1587/transinf.2023EDP7039

[12] Widiyanto W W, Rizki U, Susena E. Increased accuracy of sequence-to-sequence models with the CNN algorithm for multi response ranking on travel service conversations based on chat history. Jurnal Infotel, 2020, 12(2): 39-44. https://api.semanticscholar.org/CorpusID:225874443}

[13] Amin J, Anjum M A, Sharif M, Kadry S, Wang S H. Convolutional Bi-LSTM based human gait recognition using video sequences. Computers, Materials and Continua, 2021, 68(2): 2693-2709. https://doi.org/10.32604/cmc.2021.016871

[14] Yang Q, Jin W, Zhang Q, et al. Mixed-modality speech recognition and interaction using a wearable artificial throat. Nature Machine Intelligence, 2023, 5(2): 169-180. https://doi.org/10.1038/s42256-023-00616-6

[15] Bao H, Dong L, Wei P F. Fine-tuning pretrained transformer encoders for sequence-to-sequence learning. International journal of machine learning and cybernetics, 2024, 15(5):1711-1728. https://doi.org/10.1007/s13042-023-01992-6

[16] Li Y, Anumanchipalli G K, Mohamed A, et al. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. Nature Neuroscience, 2023, 26(12): 2213-2225. https://doi.org/10.1038/s41593-023-01468-4

[17] Enrico V, Pierre G, Tobias R. AVbook, a high-frame-rate corpus of narrative audiovisual speech for investigating multimodal speech perception. The Journal of the Acoustical Society of America, 2023, 153(5):3130-3137. https://doi.org/10.1121/10.0019460

[18] Chen X, Wang R, Khalilian-Gourtani A, et al. A neural speech decoding framework leveraging deep learning and speech synthesis. Nature Machine Intelligence, 2024, 6(4): 467-480. https://doi.org/10.1038/s42256-024-00824-8

[19] Metzger S L, Littlejohn K T, Silva A B, Moses, D. A., Seaton, M. P., Wang, R., et al. A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 2023, 620(7976): 1037-1046. https://doi.org/10.1038/s41586-023-06443-4

[20] Haider C L, Park H, Weisz H N. Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. Journal of cognitive neuroscience, 2024, 36(1):128-142. https://doi.org/10.1162/jocn_a_02059

[21] Ahmed G, Lawaye A A. CNN-based speech segments endpoints detection framework using short-time signal energy features. International Journal of Information Technology, 2023, 15(8): 4179-4191. https://doi.org/10.1007/s41870-023-01466-6

[22] Arthur F V, Csapó T G. Speech synthesis from intracranial stereotactic Electroencephalography using a neural vocoder. Infocommunications journal: a publication of the scientific association for infocommunications (HTE), 2024, 16(1): 47-55. https://doi.org/10.36244/ICJ.2024.1.6

[23] Weng Z, Qin Z, Tao X, Pan, C., Liu, G., & Li, G. Y. Deep learning enabled semantic communications with speech recognition and synthesis. IEEE Transactions on Wireless Communications, 2023, 22(9): 6227-6240. https://doi.org/10.1109/TWC.2023.3240969

[24] Lei J, Wang J, Wang Y. Multi-level attention fusion network assisted by relative entropy alignment for multimodal speech emotion recognition. Applied Intelligence, 2024, 54(17):8478-8490. https://doi.org/10.1007/s10489-024-05630-8

[25] Firdaus M R, Firdaus M R, Pasha P Y. Text-to-Speech Technology Development Using FastSpeech2 Algorithm for the Story of the Prophet. Khazanah Journal of Religion and Technology, 2024, 2(2): 55-62. https://journal.uinsgd.ac.id/kjrt

[26] Rai A, Shiwakoti S, Basukala S, & Dahal, S. S. Nepali Text-to-Speech Synthesis Using Tacotron2 and WaveGlow. KEC Journal of Science and Engineering, 2024, 8(1): 103-109. https://doi.org/10.3126/kjse.v8i1.69276