INTELLIGENT SYNTHESIS TECHNOLOGY OF CHINESE SPEECH FOR SPEECH NAVIGATION

Kade Tuerxun*

Abstract

To improve the speech synthesis (SS) technology in speech navigation APP, and to improve its SS quality and synthesis speed, the study proposed ALBert multi-syllable disambiguation method and used it in text-phoneme conversion processing. And the study also constructed a non-autoregressive Chinese SS technique based on Transformer. The research indicates that ALBert possesses the optimum disambiguation outcome, with an average accuracy of 94.2% for its polyphonic character disambiguation, 83.4% for maximum entropy model (MEM) algorithm, 83.7% for tree-guided transformation-based learning (TGTBL) algorithm, 84.3% for pypinyin tool library, and 87.1% for conditional random fields (CRF). Among the common polysyllabic words, "chao" has the highest recognition accuracy of 98.5%, and "wei" has the highest frequency of 11%. The highest performance of the FastSpeech2-GAN model is achieved at 100 k training steps, with a mean opinion score (MOS) of 3.94 and a Mel Cepstral distance (MCD) of 2.8911. The MOS scores and MCD values of the SS models are compared. The MOS score of FastSpeech2-GAN model is 3.94, and the MCD value is 2.8911, followed by FastSpeech2 model with MOS score of 3.88 and MCD value of 2.9168. 0.011, and FastSpeech2 has the same real-time rate. The studied improved Transformer-based non-autoregressive Chinese SS technology has made some progress in SS speed and SS quality.

Key Words

ALBert, Chinese speech synthesis, FastSpeech2, Transformer structure, GAN

1. Introduction

The communication of human civilisation relies on voice and text. But along with the development of artificial intelligence technology, voice communication is no longer limited to human-to-human, and human–computer voice communication has been realised in map navigation, intelligent customer service, and voice assistants [1]. In these practical applications of human–computer communication, speech synthesis (SS) technology is commonly used. Initially, the SS technology is mainly based on the recurrent neural network. But the synthesis efficiency of this model is low. At present, more excellent SS models have appeared with simpler rhyme, such as English, while Chinese has been lagging behind in the SS as it has complex rhyme structure [2]. How to improve the efficiency and accuracy of Chinese SS models is nowadays the main research direction of scholars. Therefore, the study applies the synthesis model FastSpeech2 from other languages to Chinese SS technology, and uses the PostNET structure in Tacotron2 synthesis model and generative adversarial networks (GAN) to synthesise the speech of FastSpeech2. The complete Transformer non-autoregressive Chinese SS model is optimised, and the model's effectiveness in the navigation APP is discussed.

SS technology is being used more and more frequently in intelligent fields, but the problems of traditional SS technology still exist. Wang proposed a variational self-coding model to improve the statistical parametric SS model to solve the incapability of simulating human intonation in SS systems leading to emotion deficit. And the results showed that the model successfully synthesised synthetic speech with intonation [3]. To solve the communication problem of people with dysarthria, Celin *et al.* proposed a synthesis model based on a linear microphone array with multi-resolution feature extraction and two-level data enhancement in the speech with dysarthria. The results showed that the error rate of this method was decreased by 32.79%, and its speech intelligibility was improved by 35.75% relative to the traditional method [4]. The fundamental frequencies are predicted frame by frame and cannot represent larger fundamental frequency contours. To solve these problems in conventional SS, a syllable-level fundamental frequency model was proposed by Janyoi and Seresangtakul. And the results illustrated that the method could completely represent the fundamental frequency parameter relationships in syllables [2]. Wakabayashi presented a phase estimation method

* The School of Chinese Language and Culture, Xinjiang University of Finance and Economics, Urumqi, China 830012; e-mail: tuerxunkade@163.com
Corresponding author: Kade Tuerxun

(DOI: 10.2316/J.2024.206-1050)

based on harmonic structure for enhancing the speech perception during speech enhancement. And the results illustrated that this method could accurately describe the important parameters of phase estimation [5]. To design the key components in text-to-phoneme transformation in Bengali, Ahmad *et al.* [6] proposed an encoder–decoder-based sequence to sequence (STS) model, and the results showed that the model had only 12 errors out of 135,000 training samples.

To classify household activities using sound signals, Lee and Pang proposed a non-negative matrix decomposition of the Meier spectrum feature extraction method. And the results showed that the performance of the method was very superior, and its F1 score performance was improved by 6%–12% compared to the traditional feature extraction method [7]. To enhance the inaccurate expression of rhyme in Tacotron SS technique, and improve the inaccurate rhythmic representation in Tacotron SS, Liu *et al.* proposed that Tacotron structure was extended by optimising the Mel frequency spectrum features and phrase breaks. And the outcomes demonstrated that the scheme improved the quality of synthesised speech in Mongolian and Chinese [8]. Pawlowski *et al.* presented a new method based on Transformer without deep learning network layers for addressing the long and tedious training process in Mel frequency filter bank method. This method showed obvious advantages of fast learning and solved the limitations of the traditional deep learning method of sequential computation [9]. Zhou *et al.* proposed an STS acoustic model for spliced SS to measure the dependencies between consecutive units. And the outcomes indicated that the method was superior to the HMM model with higher robustness and faster inference [10].

In summary, there are already excellent SS models in small languages and simple rhyme systems, but there is no more perfect SS model that can completely express Chinese rhyme in Chinese SS. And the Transformer has a wide range of applications in SS technology, and the Tacotron method has more applicable scenarios. Therefore, this study proposed a non-autoregressive Transformer model based on FastSpeech2, and used it to study the application of Chinese SS technology in speech navigation APP.

2. Research on the Application of Speech Intelligent Synthesis Technology in Voice Navigation

The main content of this chapter is the related research of speech intelligent synthesis technology, and the research content will be expanded from two parts. The first part is the front-end processing research of SS technology based on rule constraints, and the second part is the application of transformer in Chinese SS.

2.1 Front-End Processing of Speech Synthesis Based on Rule Constraints

There are many hard languages in the world, and Chinese is considered one of them, so it is more difficult to improve the efficiency and accuracy of Chinese SS. In building a Chinese SS model, the study proposes to add a text processor to the front-end text processing of Chinese to convert Chinese text into phonemes as a way to cut down the redundancy of the SS model. In this scheme, firstly, the Chinese text is processed with a word division and lexical annotation model, secondly, a text regulariser is used to convert special symbols to facilitate subsequent phoneme conversion, and finally, a polyphonic disambiguation model is added to the text conversion to reduce the conversion errors due to many-to-many Chinese characters and pinyin and for enhancing the quality of the synthesised speech [11]. The study selected a disambiguation and lexical annotation model based on ALBert, which is a stacked Transformer model, and the Transformer structure can input all the text into the model at once. The training efficiency of the traditional deep learning model is much lower than that of the Bert model and due to the multi-headed attention mechanism of the Transformer, the Bert model can be trained with multiple sub-models at the same time to better detect the correlation and dependency between each input data. Since the efficiency and precision of the Bert model fluctuate greatly when the amount of data is too large, the ALBert model proposes to decrease the dimensionality of the embedded data by using matrix operations in the factorisation of the input data. The fully connected layer of the ALBeert model is shared with the attention layer using parameter sharing techniques, a move that improves the dimensionality reduction of the parameters and increases the ALBert model [12]. Figure 1 demonstrated the structure of the ALBert model.

For enhancing the effect of the ALBert, it replaces the word mask operation with an n-gram mask operation. The length of the n-gram mask is chosen randomly, and its probability distribution is shown in (1).

$$p(n) = \frac{1/n}{\sum_{k=1}^{N} \frac{1}{k}} \tag{1}$$

In (1), n is the mask parameter and k is the mask length. In the actual language environment, there are a large quantity of non-standard words, and the text regulariser can transform these non-standard words in the text, into synonymous Chinese characters. The text regularisation needs to consider a more complex situation. But at this stage, there is a lack of Chinese text regularisation dataset, and personally constructed data has the problem of incomplete data. Therefore, the study uses rule-based regularisation processing, and its process is shown in Fig. 2 [13].

In the presence of numbers, the ALBert model will divide special symbols and numbers into one word and occupy one label. So this text regularisation processor mainly targets special symbols and Arabic numbers. After the regularisation, the front-end text processing also needs to convert the text to pinyin. If the deep learning method is used to predict the reading of polyphonic words, it requires a lot of manual and professional knowledge. So the Pypinyin tool library is mostly used in the processing of text to pinyin, and the rule-based constraint-based disambiguation model for polyphonic words used in the

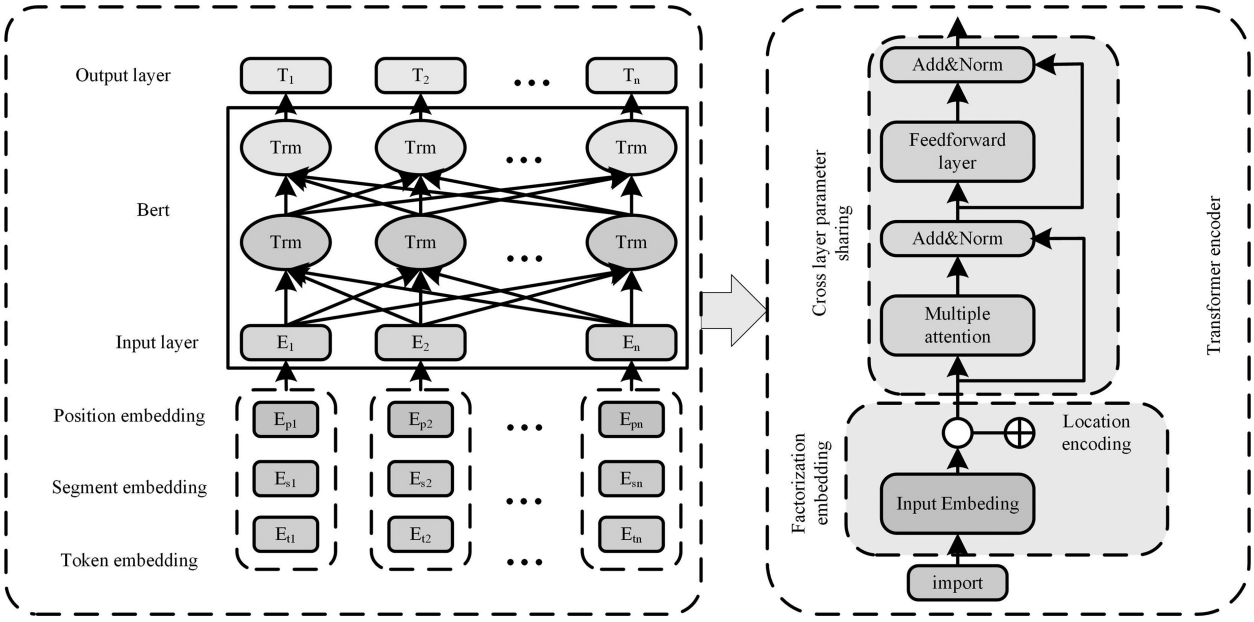

Figure 1. Structure diagram of ALBert model.

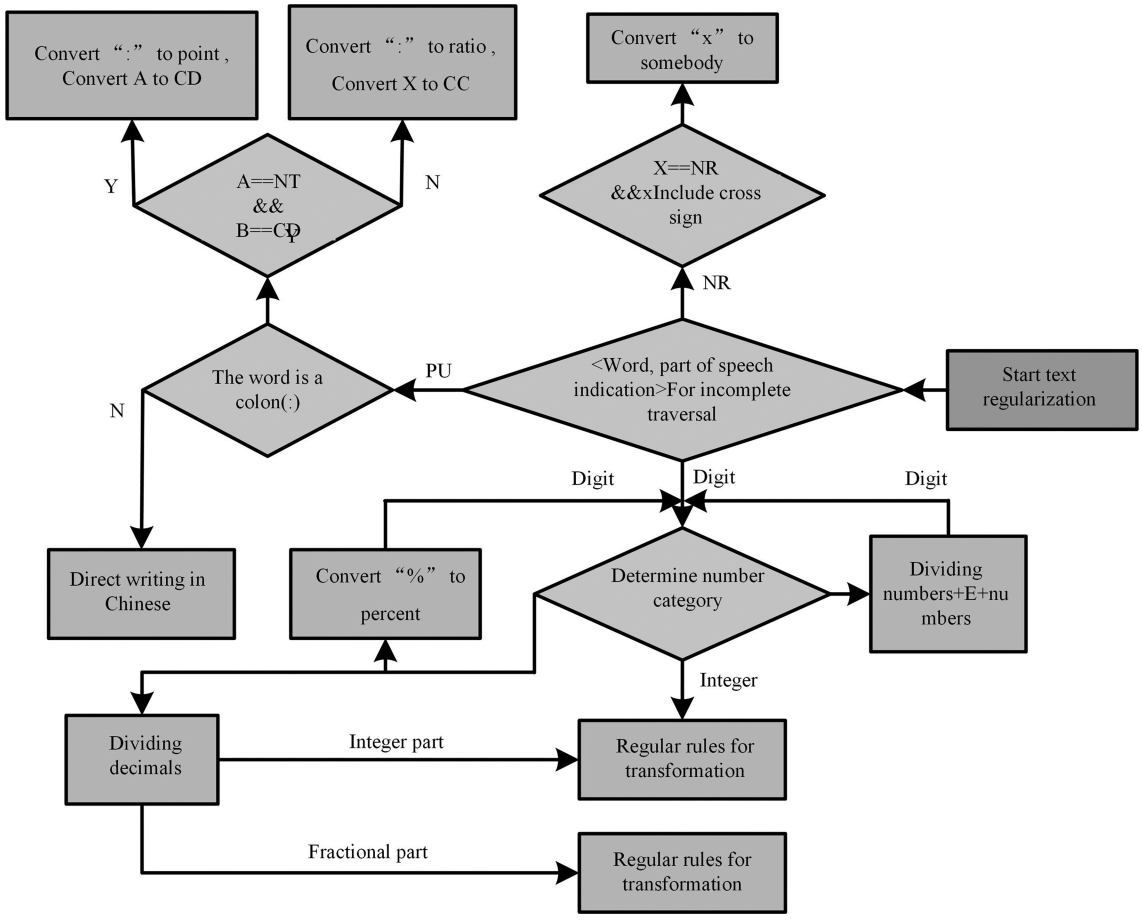

Figure 2. Processing flow of rule-based text regularisation.

study. In the process, the data to be converted are labelled. If the corpus embodies a word, they will be converted directly. If not, the model would traverse each Chinese character in the word one by one for conversion. In this process, if it is not a polyphonic word, it will be converted directly. If it is a polyphonic word, it will be converted after querying the specified constraints according to the rules, and its flow is shown in Fig. 3 [14].

Usually, a phonetic action corresponds to a phoneme. But Chinese Pinyin is divided into rhymes and consonants,

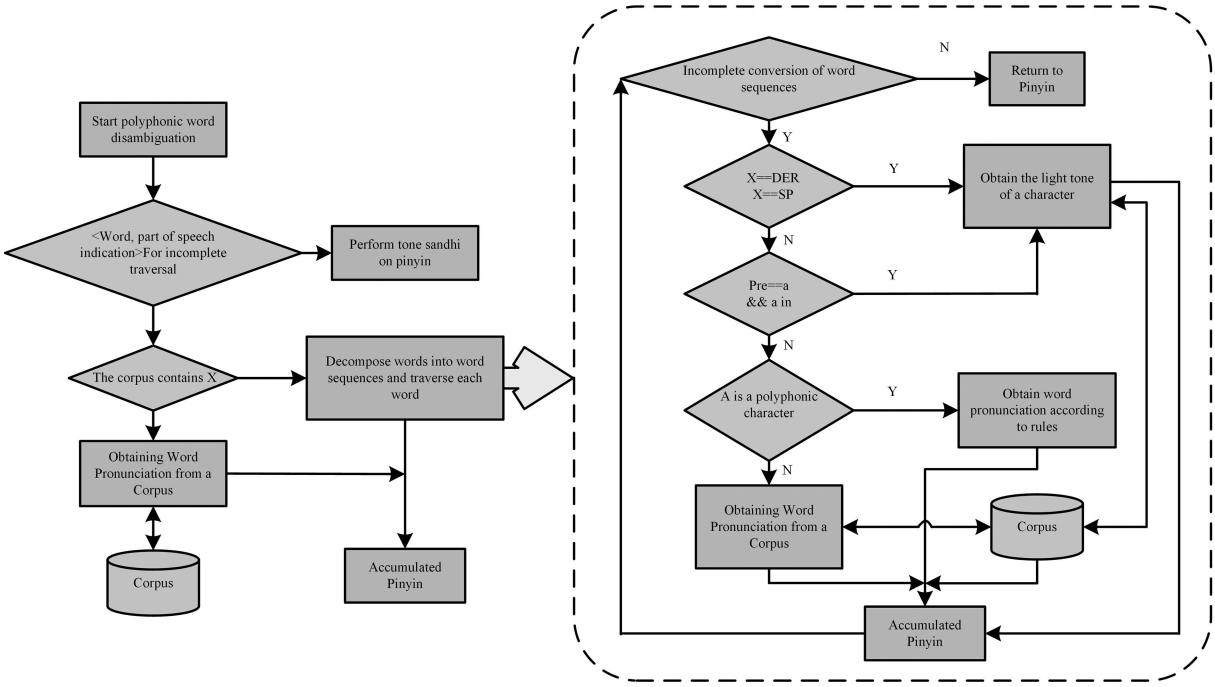

Figure 3. Flow chart of polyphonic word disambiguation model.

and there are four basic tones and one special tone. Therefore, to facilitate labelling during phoneme transformation, the study labels the tones with numbers. The front-end processing of the study is evaluated by comparing the accuracy rate (AR) and the character error rate (CER) of the model. AR is calculated as shown in (2) [15].

$$AR = \frac{N - S}{N} \times 100\% \tag{2}$$

In (2), S denotes the total number of typos and N is the total quantity of words in the text. CER is calculated as in (3).

$$CER = \frac{S}{N} \times 100\% \tag{3}$$

2.2 Chinese Speech Synthesis Model Based on Transformer

Most of the current Chinese SS models are autoregressive synthesis models, which need a lot of training before they can be used in SS. The phoneme modelling of Chinese is more complex, and the synthesis effect on Chinese datasets is inferior to that on English and other language datasets at this stage of SS technology. To solve this problem, the study proposes a non-autoregressive SS method [16]. The study first proposes the FastSpeech2-GAN Mel spectrum generator, and then uses the vocoder to transform Mel spectrum features into real audio. Since FastSpeech2-GAN uses the Transformer structure to build the Mel spectrum generator, the problem of long model training time in Chinese SS is fundamentally solved. The Transformer SS structure is a model of self-attentive mechanism, and the SS technology mainly uses RNN combined with the model of attention mechanism before the introduction of this structure. Although this model can stably solve the timing problem in the synthesis process, it also makes its own are SS speed is severely constrained, and cannot efficiently use the parallelism of GPU [17]. In the timing problem, Transformer SS structure uses multi-head attention mechanism instead of the RNN structure in the Tacotron2 model, extending the focus of the fully connected layer in different positions. So the model can take into account the efficient utilisation of GPU, and its structure is shown in Fig. 4.

Although Transformer SS is a new model proposed by the study, its core is still the Transformer structure. The attention weights are calculated as shown in (4) [18].

$$\begin{cases} \text{Multi Head}\,(Q,K,V) = \text{Concat}\,(\text{head}_1,\ldots,\text{head}_h)\,W^o \\ \text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{cases} \tag{4}$$

In (4), Q, K, V denote a set of matrices composed of query, key, valu in the model, $\sqrt{d_k}$ is the scaling factor, W_i^Q is the mapping matrix of query, W_i^K is the mapping matrix of key, W_i^V is the mapping matrix of valu, and W^O is the mapping matrix after all the attentions are connected. The model also requires the dot product operation, which is given in (5).

$$\text{Attention}\,(Q,K,V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

It has been mentioned in the above description that the study designs a structure in which a multi-headed attention mechanism is used instead of RNN structure. Therefore, the model designed by the study cannot obtain the location information autonomously. To solve this problem, the study adds positional encodings (PE) to the input information of the structure, and the calculation of the location information encoding is usually done using

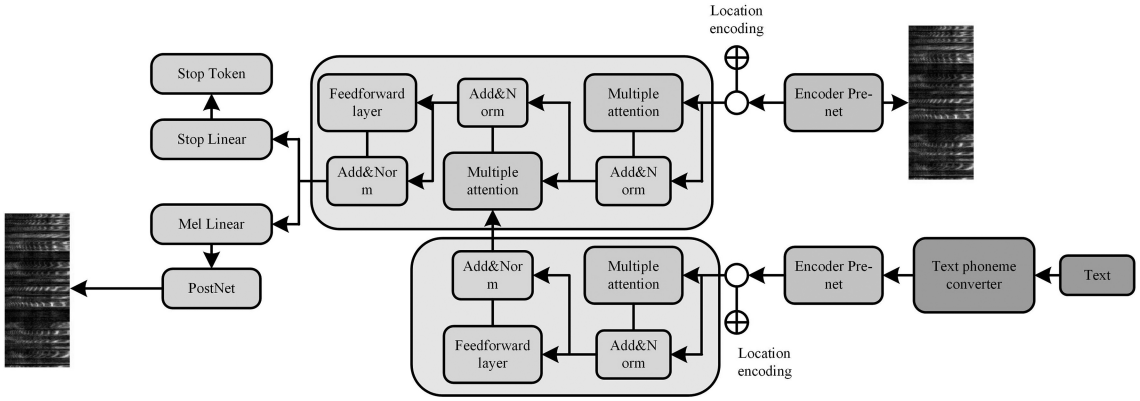

Figure 4. Overall framework of Transformer TTS.

several different frequency functions for calculation, and the calculation equation is given in (6).

$$\begin{cases} PE_{(\mathrm{pos},2i)} = \sin\left(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}\right) \\ PE_{(\mathrm{pos},2i+1)} = \cos\left(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}\right) \end{cases} \tag{6}$$

In (6), pos serves as the time step index and $2i, 2i+1$ serve as the channel indexes. Due to the different target domains, Transformer structure requires architectural migration. In the migration, Transformer TTS needs to add triangle location embedding to PE so that the model can be automatically adjusted to match the output ratio of the preprocessing network in the encoder, and its processing equation is shown in (7).

$$x_i = \mathrm{prenet}\left(\mathrm{phoneme}_i\right) + \alpha PE\left(i\right) \tag{7}$$

In (7), $\mathrm{phoneme}_i$ is the input phoneme of the model and α is the parameter. Although the Transformer TTS model takes into account the efficient use of GPU parallelism, each of its output results depends on the previous one. Therefore, to achieve the construction of a non-autoregressive SS model, the study proposes a Meier spectrum generator based on the Transformer structure, which uses the FastSpeech2 structure as a framework, as specified in structure is shown in Fig. 5.

FastSpeed2-GAN uses the addition of Duration Predictor module and Length Regulator module to solve the problem that the output result of Transformer TTS model depends on the previous result. To alleviate the pressure of one-to-many modelling during the conversion of text to audio due to the presence of polyphonic characters, the study adds two Predictor modules to FastSpeed2-GAN to achieve the prediction of acoustic energy information. In the construction of the Predictor module, the study adopts the structure of Duration Predictor, which is used in the training. This model optimises the parameters by calculating the mean square error loss of the output data of the predictor, compared with the actual data. The mean square error loss in this model is shown in (8).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \widehat{y}_i\right)^2 \tag{8}$$

In (8), MSE denotes the mean square error, m serves as the quantity of data, y is the true Predictor value, and $\widehat{y}$ is the prediction value of the predictor. In the model training process, it is also necessary for the model to count the average absolute error loss of the Mel spectral features before PostNet processing, Mel spectral features after PostNet processing and the true Mel spectral features, and the calculation equation is shown in (9).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \widehat{y}_i\right| \tag{9}$$

In (9), MAE is the mean absolute error. Therefore, the equation for the total generator loss for this model is given in (10).

$$\begin{aligned} L_G = {} & MSE\left(d_{\mathrm{GT}}, d\right) + MSE\left(e_{\mathrm{GT}}, e\right) + MSE\left(f_{\mathrm{GT}}, f\right) \\ & + MAE\left(\mathrm{mel}_{\mathrm{GT}}, \mathrm{mel}_{\mathrm{before}}\right) \\ & + MAE\left(\mathrm{mel}_{\mathrm{GT}}, \mathrm{mel}_{\mathrm{after}}\right) \end{aligned} \tag{10}$$

In (10), L_G denotes the total generator loss, d is the phoneme duration prediction, e is the sound energy prediction information, f is the treble prediction information, d_{GT} is the true phoneme duration, e_{GT} is the true sound energy information, and f_{GT} is the true treble information. The generator of FastSpeed2-GAN, after synthesising the audio signal, needs a discriminator to judge the truth or falsity of the segment, and the accuracy of the discriminator on the truth or falsity of the audio signal is 50%. The discriminator of FastSpeed2-GAN, which is composed of a simple multi-layer neural network, is used to enhance the robustness of the GAN training process. The study uses universal normalisation instead of layer normalisation in the discriminator of FastSpeed2-GAN, and the Lipschitz constant of the discriminator is constraint to limit the local float of the function. The Lipschitz constraint requires that the function needs to satisfy (11) in the definition domain.

$$\frac{\left\|f\left(x\right) - f\left(x'\right)\right\|_2}{\left\|x - x'\right\|_2} \leq M \tag{11}$$

In (11), $\left\|\cdot\right\|_2$ denotes L2 regularisation, x, x' are any values in the domain of function definition, and the smallest M satisfying the condition is the Lipschitz constant. The

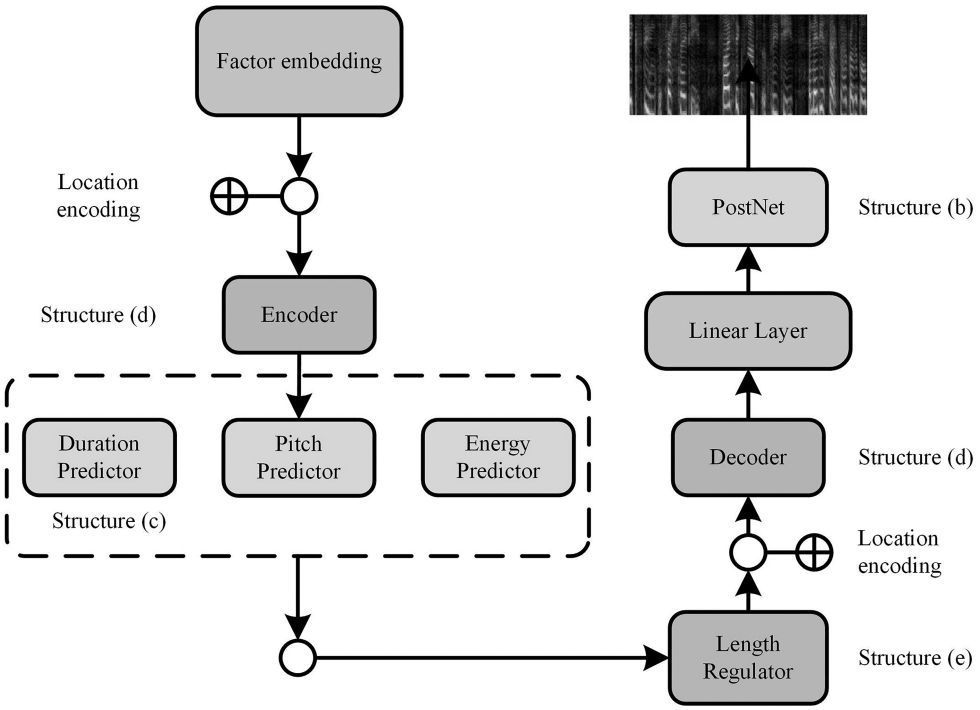

Figure 5. FastSpeed2-GAN generator structural model.

multi-band MelGAN vocoder used in the study of the conversion of the Mel spectrum into real audio signals contains a generator and a judge, and its structure is shown in Fig. 6.

To accurately determine the authenticity of audio signals, multi-band MelGAN uses short-time Fourier transform (STFT) loss as the loss function of the generator. In the STFT loss, multi-band MelGAN needs to calculate the convergence loss and amplitude loss of STFT. The convergence loss L_{SC} is calculated in (12).

$$L_{\mathrm{SC}}(x,s)=\frac{\left\|\left|\mathrm{STFT}(x)\right|-\left|\mathrm{STFT}(G(s))\right|\right\|_F}{\left\|\left|\mathrm{STFT}(x)\right|\right\|_F} \tag{12}$$

In (12), x is the real audio, s is the Meier spectrum feature generated by the generator G, $\|\cdot\|_F$ denotes the Frobenius paradigm, $\|\cdot\|_1$ denotes the L1 paradigm, and $|\mathrm{STFT}(\cdot)|$ denotes the STFT function used to calculate the amplitude. logSTFT amplitude loss function is calculated in (13).

$$L_{\mathrm{mag}}(x,s)=\frac{1}{N}\left\|\log\left|\mathrm{STFT}(x)\right|-\log\left|\mathrm{STFT}(G(s))\right|\right\|_1 \tag{13}$$

In (13), N indicates the number of elements in the repetition. When training the generators individually in multi-band MelGAN, it is necessary to calculate the generator loss. Equation (14) indicates the calculation equation.

$$L(G)=\frac{1}{2}\left(L^{\mathrm{full}}_{\mathrm{fmr\_stft}}(G)+L^{\mathrm{sub}}_{\mathrm{smr\_stft}}(G)\right) \tag{14}$$

In (14), $L^{\mathrm{full}}_{\mathrm{fmr\,stft}}$ is the full-band loss of multi-scale STFT, and $L^{\mathrm{sub}}_{\mathrm{smr\,stft}}$ is the sub-band loss. In the multi-band MelGAN model, there are three modules of the discriminator. The first module extracts the features directly from the original audio, and the remaining two modules extract the features in the two-fold and four-fold down sampling of the original audio, respectively. After the construction of the model, the study evaluated the effect of the model by mean opinion score (MOS), which is used to evaluate the quality and distortion of the synthesised speech, and the higher the score, the better the quality of the speech.

$$\mu_i=\frac{1}{N_i}\sum_{k=1}^{N_i}m_{i,k} \tag{15}$$

In (15), N_i is the model sample and $m_{i,k}$ indicates the score of the k generated sample. The final evaluation also requires the calculation of the 95% confidence interval for the mean score, which is given in (16).

$$CI_i=\left[\widehat{\mu}_i-1.96\frac{\widehat{\sigma}_i}{\sqrt{N_i}},\widehat{\mu}_i+1.96\frac{\widehat{\sigma}_i}{\sqrt{N_i}}\right] \tag{16}$$

In (16), $\widehat{\sigma}_i$ is the standard deviation of the collected MOS scores. Apart from the MOS score, it is also necessary for the model performance evaluation to consider the SS speed, and the real time factor (RTF) is a commonly used metric to measure the SS speed.

3. Analysis of Speech Synthesis Results

The main content of this chapter is to analyse the synthesis results of the SS model proposed, which will be analysed from two aspects. The first aspect is to analyse the transformation results of text-phoneme, and the second aspect is to analyse the results of SS based on Transformer.

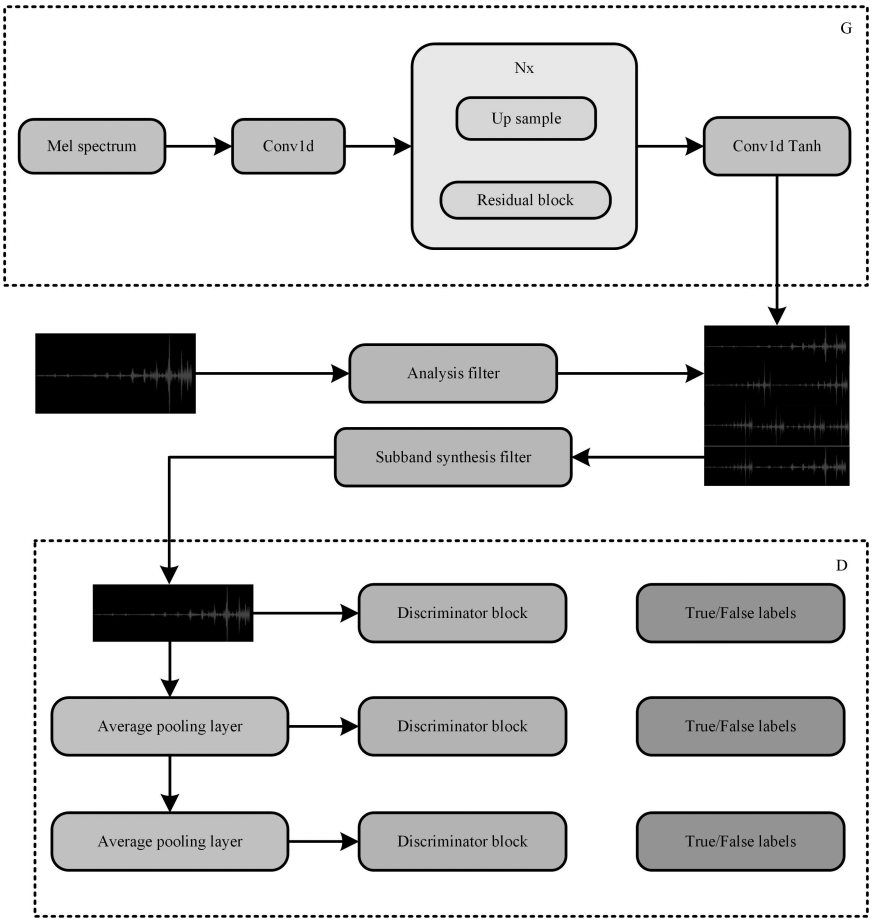

Figure 6. Structure diagram of multi-band MelGAN model.

3.1 Analysis of the Results of the Text-Phoneme Transformer Based on Rule Constraints

To verify the superiority of the proposed ALBert polysyllabic disambiguation algorithm, the study used Pycharm software to compare the accuracy and error rates of different methods. ALBert, maximum entropy model (MEM) polysyllabic disambiguation method, tree-guided transformation-based learning (TGTBL), pypinyin tool library, and conditional random fields (CRF)-based disambiguation and lexical annotation methods were compared in polyphonic disambiguation, as shown in Fig. 7.

From Fig. 7(a), it can be seen that the accuracy rate of ALBert algorithm is more than 90% in the comparison of the accuracy rates of five kinds of multi-syllable disambiguation methods, among which the recognition accuracy rate of the character "zhe" is the lowest, 92.3%. The average accuracy of the ALBert algorithm is 94.2%, while the accuracy of the rest of the algorithms is lower than 90%, with the highest being 89.8%, and the average accuracy of the MEM algorithm is 83.4%. The average accuracy of TGTBL algorithm is 83.7%, the average accuracy of pypinyin tool library is 84.3%, and the average accuracy of CRF is 87.1%. It can be seen that the polyphonic word disambiguation accuracy of ALBert algorithm is much higher than the remaining four algorithms, and ALBert algorithm has the best performance. The transformation results of this algorithm were observed in the study, and the results are shown in Fig. 8.

Figure 8(a) shows the accuracy rate of polysyllabic words, and Fig. 8(b) shows the frequency of common polysyllabic words. In Fig. 8(a), among all polyphonic characters, the character with the highest labeling accuracy is "chao", with an accuracy rate of 98.5%, followed by "cang", with an accuracy rate of 98.0%. In Fig. 8(b), it can be seen that among the common polysyllabic characters, "cang" has the lowest frequency of 3%, and when it appears, 69.4% of the time, its pinyin is "cang". Among the common polysyllabic characters, "wei" has the highest the frequency of "for" is the highest, at 11%, and when it occurs, the pinyin is "wei" in 63.1% of cases.

3.2 Analysis of Transformer-Based Speech Synthesis Results

For testing the potency of the non-autoregressive model presented in the study, the study compared the synthetic effects of Tacotron2, FastSpeech2, and FastSpeech2-GAN on the Baker dataset, and Table 1 illustrates the results.

In Table 1, it can be seen that, except for natural speech, the non-autoregressive FastSpeech2-GAN model proposed in the study has the highest MOS score of 3.94 and the autoregressive Tacotron2 model has the lowest MOS score of 3.88. The FastSpeech2-GAN model has the

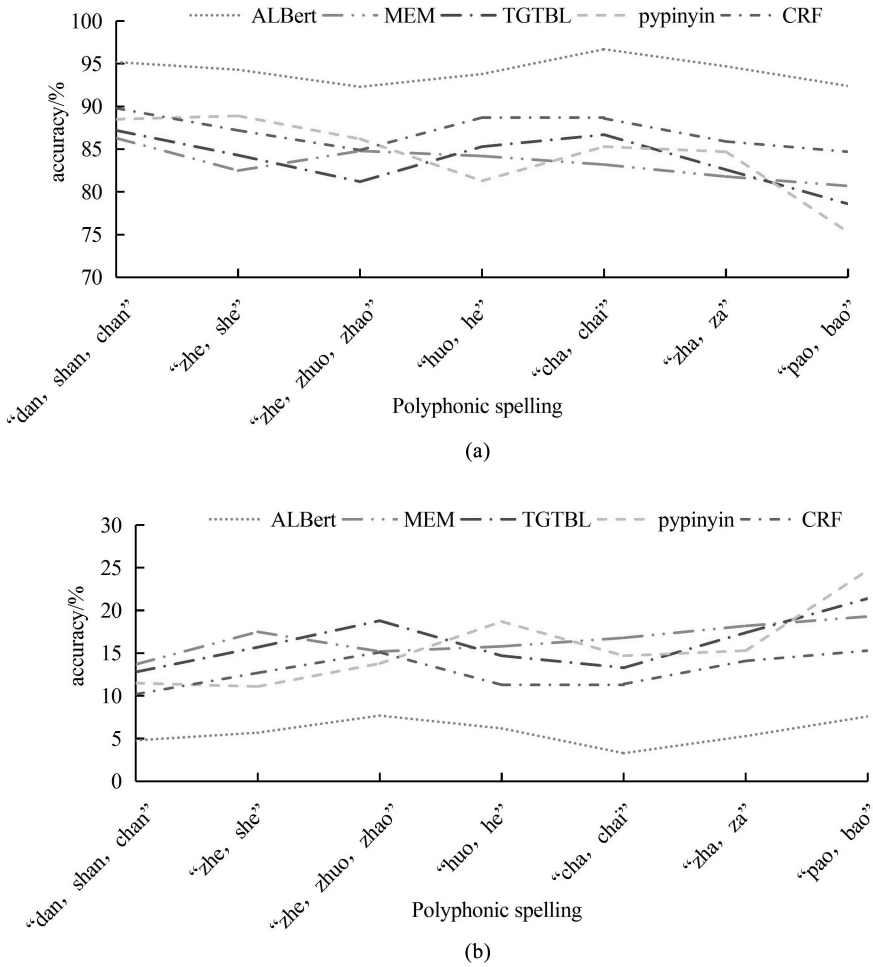

Figure 7. Performance comparison of polyphonic word disambiguation method: (a) comparison of accuracy of polyphonic word disambiguation methods and (b) comparison of error rates in polyphonic word disambiguation methods.

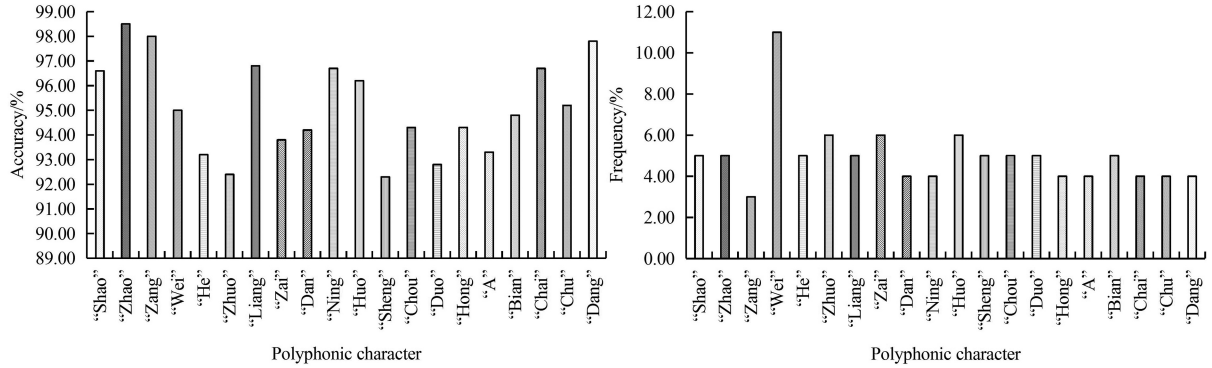

Figure 8. ALBert polyphonic word disambiguation effect.

lowest MCD value of 2.8911 and the Tacotron2 model has the highest MCD value of 2.9934. The FDSD and cFDSD values of the Tacotron2 model are 0.0512 and 0.0178, respectively, which are much higher than the remaining two models. For testing the effect of discriminator on the SS effect in FastSpeech2-GAN model, the study compared the MOS score, Mel Cepstral distance (MCD), mean FDSD of deep speech distance of two audio samples and depth of two audio samples for the synthesised speech at different number of steps with the addition of discriminator speech distance variance cFDSD. The results are shown in Fig. 9.

Figure 9(a) shows the MOS score and MCD values of the model, where the MOS score is taken as the mean value. Figure 8(b) shows the FDSD and cFDSD values of the model. In Fig. 8(a), it can be seen that the highest MOS score of the model is 3.94 and the lowest MCD value is 2.8911 when the discriminator is added at step 100 k. And the lowest MOS score of the model is 3.88 and the

Table 1
Synthesis Effect of Tacotron2, FastSpeed2, and FastSpeed2-GAN

	MOS	MCD	FDSD	cFDSD
Natural speech	4.43 ± 0.06	/	/	/
Tacotron2	3.80 ± 0.08	2.9934	0.0512	0.0178
FastSpeech2	3.88 ± 0.09	2.9168	0.0156	0.0019
FastSpeech2-GAN	3.94 ± 0.08	2.8911	0.0158	0.0015

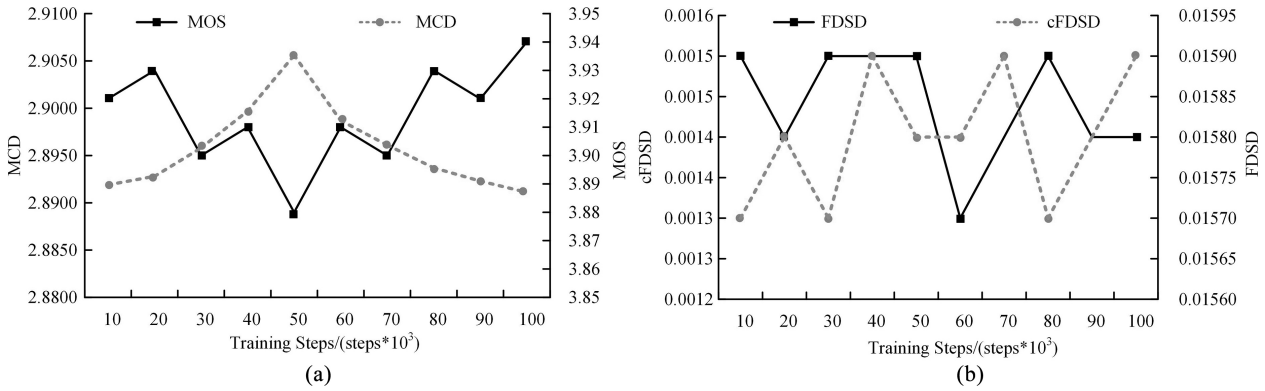

Figure 9. FastSpeed2-GAN synthesis effect when asynchronous number is added to discriminator: (a) MOS score and MCD values under different training steps and (b) FDSD and cFDSD values under different training steps.

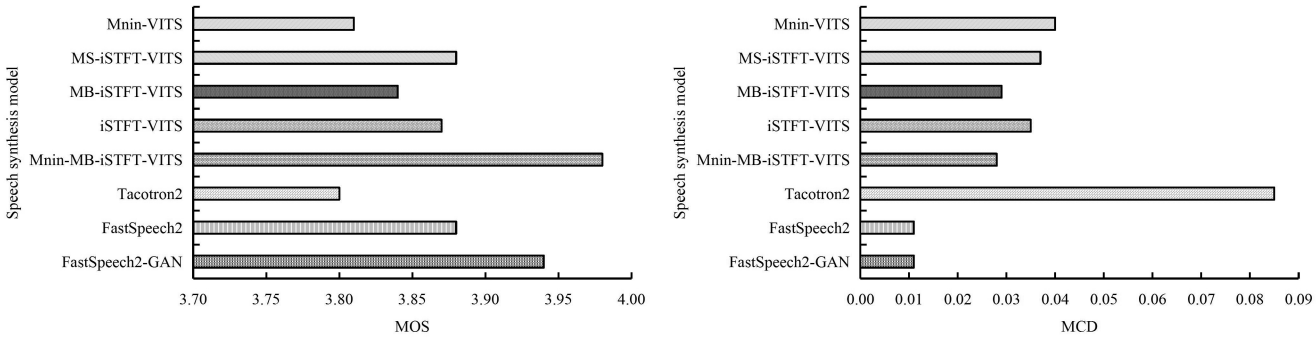

Figure 10. Mnin-MB-iSTFT-VITS model.

highest MCD value is 2.9056 when the discriminator is added at step 50 k. The higher the MOS score of the model and the lower the MCD value, the better the model synthesis. From Fig. 8(b), there is almost no effect on the FDSD value and cFDSD value of the model at which step the discriminator is added, and the difference is within 0.0002. Therefore, the model with the discriminator added at the 100 kth step is used in the study. The function of the SS model needs to consider not only the SS treatment but also the SS speed. Figure 10 shows the MOS score and the SS speed of the FastSpeech2-GAN SS model compared with the Vits model.

In Fig. 10, it can be seen that among all models, the Mnin-MB-iSTFT-VITS model has the highest MOS score of 3.98, followed by the FastSpeech2-GAN model with the highest MOS score of 3.94, which is 0.04 lower than the Mnin-MB-iSTFT-VITS model, and the rest of the models have MOS scores of 3.90 or below. The models with the lowest real-time rate are FastSpeech2-GAN and FastSpeech2, both with a real-time rate of 0.011. The next lowest real-time rate is Mnin-MB-iSTFT-VITS with a real-time rate of 0.028. The Tacotron2 model has a real-time rate of 0.084, which is much higher than the FastSpeech2 GAN. The FastSpeech2-GAN model constructed in the study is stronger than the rest of the models in terms of comprehensive performance of SS effect and SS speed. This study also compares the loss function curves of Transformer autoregressive SS model and Transformer-based non-autoregressive SS model, as shown in Fig. 11.

In Fig. 11(a), it can be seen that the loss functions of both models are slowing down during the training. The Transformer non-autoregressive model has a slower decline of the model loss function than the Transformer autoregressive model during the training period. This

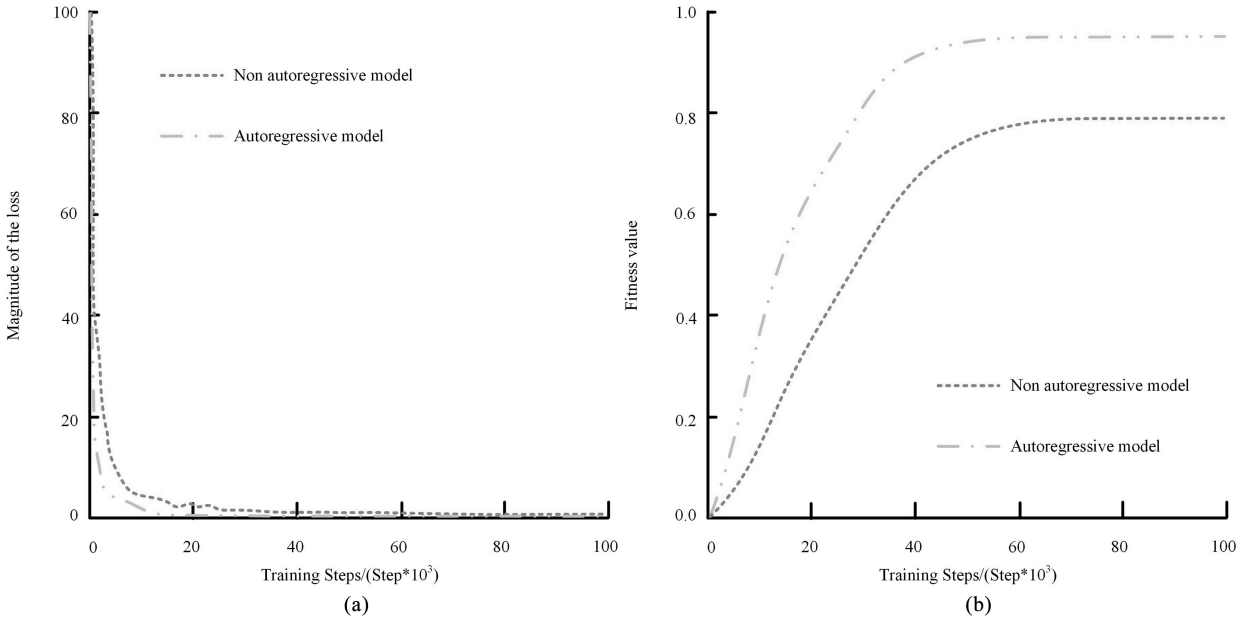

Figure 11. (a) Change curve of loss function and (b) change curve of fitness function.

model slows down the decline of the loss function because the SS method proposed in the study changes the modelling of the temporal sequence. As the training increases, the loss functions of two models are gradually aligned. In Fig. 11(b), it can be seen that the Transformer non-autoregressive model has a lower adaptation and better training results.

4. Conclusion

To design Chinese SS model with better synthesis effect and faster synthesis speed, the study proposes ALBert polyphonic word disambiguation algorithm, which is used for phoneme transformation of text. And the study proposes Transformer-based Chinese SS model, which uses FastSpeech2-GAN algorithm to complete the output of synthesised speech. The model indicates that ALBert has the best disambiguation outcome, with an average accuracy of 5.8% for its polyphonic character disambiguation, 16.6% for MEM algorithm, 16.3% for TGTBL algorithm, 15.7% for pypinyin tool library, and 12.9% for CRF. Among the common polysyllabic words, "chao" has the highest conversion accuracy of 98.5% and its frequency is 5%, while "wei" has the highest frequency of 11% and its accuracy is 95%. The FastSpeech2-GAN model has the worst synthesis effect when the training is 50 k, the best synthesis effect when the training is 100 k, the lowest MOS score and the highest MCD value are 3.88 and 2.9056 at 50 k steps, respectively, and the MOS score and MCD value are 3.94 and 2.8911 at 100 k steps, respectively. The proposed FastSpeech2-GAN algorithm has the highest MOS score of 3.94, while its MCD value is the lowest at 2.8911. Among the remaining synthesis algorithms, the highest MOS score is the FastSpeech2 model with the highest score of 3.88. The FastSpeech2-GAN model has the best synthesis with its MOS score of 3.94 and its time rate of 0.01. The research successfully designs and built Transformer-based Chinese SS model, and makes it possible to apply it in speech navigation APP, but its SS quality still needs to be improved.

References

[1] Z. Mu, X. Yang, and Y. Dong, Review of end-to-end speech synthesis technology based on deep learning, 2021, *arXiv:2104.09995*.

[2] P. Janyoi and P. Seresangtakul, F0 modeling for Isarn speech synthesis using deep neural networks and syllable-level feature representation, *International Arab Journal of Information Technology*, *17*(6), 2020, 906–915.

[3] Y. Wang, A speech synthesis model with mood based on variational autoencoder, *Computer Science and Application*, *10*(12), 2020, 2159–2167.

[4] T. Celin, T. Nagarajan, and P. Vijayalakshmi, Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition, *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 2020, 346–354.

[5] Y. Wakabayashi, Speech enhancement using harmonic-structure-based phase reconstruction, *Acoustical Science and Technology*, *40*(3), 2019, 162–169.

[6] A. Ahmad, M.R. Selim, M.Z. Iqbal, and S.R. Mohammad, An encoder-decoder based grapheme-to-phoneme converter for Bangla speech synthesis, *Acoustical Science and Technology*, *40*(6), 2019, 374–381.

[7] S. Lee and H.S. Pang. Feature extraction based on the non-negative matrix factorization of convolutional neural networks for monitoring domestic activity with acoustic signals, *IEEE Access*, *8*(7), 2020, 122384–122395.

[8] Y. Lin, F. Wang, X. Cui, L. Hong, and Y. Liu, A parameterized representation for self-motion manifold of crawler crane robots, *International Journal of Robotics and Automation*, *37*(2), 2022, 219–226.

[9] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, Modeling prosodic phrasing with multi-task learning in Tacotron-based TTS, *IEEE Signal Processing Letters*, *27*(8), 2020, 1470–1474.

[10] X. Zhou, Z. Ling, and L.R. Dai, UnitNet: A sequence-to-sequence acoustic model for concatenative speech synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*(6), 2021, 2643–2655.

[11] B. Pérez-Canedo and J.L. Verdegay, On the application of a lexicographic method to fuzzy linear programming problems, *Journal of Computational and Cognitive Engineering*, *2*(1), 2023, 47–56.

[12] S. Oslund, C. Washington, A. So, T. Chen, and H. Ji, Multiview robust adversarial stickers for arbitrary objects in the physical world, *Journal of Computational and Cognitive Engineering*, *1*(4), 2022, 152–158.

[13] A. Lanza, S. Morigi, I.W. Selesnick, and S. Fiorella, Sparsity-inducing nonconvex nonseparable regularization for convex image processing, *Siam Journal on Imaging Sciences*, *12*(2), 2019, 1099–1134.

[14] A. Walker, S. Abedi, and S. Kwon, Design of threshold-based energy storage control policy based on rule-constrained two-stage stochastic program, *International Journal of Electrical Power & Energy Systems*, *137*, 2022, 107798–107810.

[15] E. Brown, I.A. Jamsek, L. Liang, F. Rachael, and T.B. Holt, Predicting children's word recognition accuracy with two distance metrics, *The Journal of the Acoustical Society of America*, *145*(3), 2019, 1798.

[16] V.Y. Semenov, Methods for calculating and coding the parameters of autoregressive speech model when developing the vocoder based on fixed point signal process, *Journal of Automation and Information Sciences*, *51*(2), 2019, 30–40.

[17] Y. Liu and J. Zheng, Es-Tacotron2: Multi-task Tacotron 2 with pre-trained estimated network for reducing the over-smoothness problem, *Information (Switzerland)*, *10*(4), 2019, 131–152.

[18] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, ChangeMask: Deep multi-task encoder-Transformer-decoder architecture for semantic change detection, *ISPRS Journal of Photogrammetry and Remote Sensing*, *183*(5), 2022, 228–239.

Biographies

Kade Tuerxun received the B.S. and M.S. degrees in Chinese language and literature from Xinjiang University in 2001 and 2006, respectively. In 2011, he finished the Linguistics and Applied Linguistics program from Minzu University of China and received the Ph.D. in literature. Since then, he has been engaged in language teaching and research with the School of Chinese Language and Culture, Xinjiang University of Finance and Economics. Until now, he has published over 20 research papers, more than ten scientific research projects and one academic monograph. He is currently an Associate Professor with the School of Chinese Language and Culture, Xinjiang University of Finance and Economics. His main research areas include phonology, phonetics, and speech engineering.