Improved VITS-Based Multilingual AI Speech Synthesis Model with Domain Adaptors and Acoustic Feature Optimization

Xing Yang
Communication University of Shanxi, Shanxi Film Academy, Jinzhong, 030619, China
E-mail: soundgoal0506@outlook.com

Keywords: speech synthesis technology, acoustic characteristics, domain adaptors, audio conversion, multilingual speech synthesis

Received: November 19, 2024

Speech synthesis technology plays an important role in global economic and cultural exchanges, and multilingual speech synthesis and output are still unable to meet the current development needs of the global market. The study proposes the use of acoustic feature conversion methods and steps for decoupling multilingual information, combined with modules of domain adaptors to improve end-to-end text to speech variational inference and adversarial learning models, to adapt to the application of multilingual speech synthesis. Through the evaluation of speech synthesis technology indicators, it was found that the average selection score of the model after removing the regularization term for similarity in different languages was 4.93. The synthesis model without domain adaptors significantly reduced the naturalness of speech synthesis by 0.8 compared to multilingual speech synthesis models, indicating that domain adaptors have a good effect on the naturalness of speech synthesis. In cross-lingual indicator analysis, the model proposed by the research achieved the highest naturalness result, with an average selection score of 4.26 and 3.96 for naturalness and similarity in transit English. In the intermediate day voice synthesis with a data volume of 200, the highest accuracy was 94.58%, which was 16.53% higher than traditional speech synthesis frameworks. Comparing the cross-lingual synthesis performance of the synthesis model, it was found that the model had an accuracy rate of 94.58% and a time of 3.12 seconds for the synthesis of Chinese to Japanese conversion with a data volume of 200. The above results demonstrate the feasibility and superiority of the multilingual speech synthesis model based on domain adaptors, which adds multilingual imagery to speech synthesis applications in the field of artificial intelligence and promotes the industrial development and intelligent services of speech synthesis technology.

Povzetek: Raziskava predstavi izboljšan govorni sintetiziator z modelom VITS in domenskimi adapterji, kar poveča naravnost, točnost in časovno učinkovitost sinteze govora v različnih jezikih.

1 Introduction

With the development of technology, artificial intelligence (AI) is constantly integrating into people's lives 0. Speech recognition and synthesis technology have promoted robots to “speak” and become important means of human-computer interaction. Speech recognition can convert speech into text, while synthesis technology can transform textual information into speech patterns, and the emotions, timbre, rhythm, and other aspects of synthesized speech are increasingly catering to human image and scene needs. In addition, the diversity of global language types has also led to a wider range of objects for speech synthesis, and the transformation of multilingual synthesized speeches can meet different language environments and multimedia platforms 0. In personalized development and social demand, speech synthesis and speech recognition technologies typically perceive language environment features to adapt to the user’s environment. In the oral application of patients with hearing impairment, autism, and language disorders, speech recognition technology can extract features from the original voice and enhance the test speech to improve the language ability of patients with disabilities [3]. Speech synthesis technology, based on speech recognition, adds text processing and audio conversion to language, enhancing the quality of speech synthesis through analysis of audio signals. In recent years, scholars at home and abroad have conducted a lot of research on speech recognition and synthesis technology. In speech recognition technology, Jiang et al. proposed to address the technical issues of text to speech synthesis by using semantic dependencies to extract speech information from the original text. They combined convolutional neural networks and self attention mechanisms to improve pronunciation accuracy, and separated tones and phonemes in model training to demonstrate the reliability of their model in speech synthesis [4]. Liu et al. proposed a multi-modal speech synthesis method in the field of dubbing, which utilizes multi-modal and multi-scale text to speech technology to provide style embedding for video input. Multi-scale styles are reused to convert the speech style of reference videos, thereby generating high-quality video aligned speech [5]. In the problem of distinguishing speech emotions, Long et al. proposed a multi-distribution method based on Mel frequency cepstral graph and parameter transfer, and proved its effectiveness in classifying language emotions through model training [6].

In the subsequent field of speech synthesis technology, breakthroughs were made in speech classification and recognition applications, and the development of speech conversion and output was also expanded. Ghosh et al. proposed the use of visual transformers and attention mechanisms to capture the spectral and temporal features of speech in the audio-visual field, thereby improving the quality of speech synthesis [7]. Okamoto et al. proposed using a neural encoder and a multi-speaker corpus to resample acoustic features for speech rate conversion technology, demonstrating that their method has high-quality speech rate conversion [8]. Masood et al. proposed the use of generative adversarial networks and deep forgery generation tools to detect audio and video data in response to speech conversion and forgery issues on social media, demonstrating the rapid development of speech conversion and the application of synthesis technology [9]. In the application of language and culture, Zhang et al. proposed a cross-lingual pre-training method for adversarial training of text and image generation, thereby increasing the robustness of the model to the target language [10]. Kaur et al. proposed a method for text to speech synthesis based on deep learning to construct an acoustic model, which can improve the fidelity of synthesized speech and achieve high performance evaluation [11]. However, in terms of specific multi-lingual synthesis techniques and audio conversion, current research still lacks in-depth analysis, and the phonemes, pronunciations, and duration of language data have not been adjusted. Due to the significant differences between different languages, it is difficult to establish a cross-lingual sound learning model. At the same time, sound collection and datasets for multiple languages are scarce, which leads to model training only with a small number of language data. The generalization ability and speech output quality of the model cannot meet people's actual needs. Currently, cross-lingual speech synthesis cannot fully achieve the mutual conversion of multiple languages. Therefore, based on the variational inference with adversarial learning for end-to-end text to speech (VITS) model, this study innovatively uses neighborhood adaptors to extend multi-lingual language rules and improves the structure of the VITS model to construct a multi-lingual AI speech synthesis model based on domain adaptors. To meet the application needs of speech synthesis in various fields and globalization, research is being conducted on the mutual conversion and recognition of output sounds from multiple languages. At the same time, advanced technology is being used to establish an AI image with multi-lingual speech synthesis capabilities, aiming to improve the fluency and naturalness of speech synthesis in different languages, and provide personalized speech experience and technical reference for the field of AI.

This study is mainly divided into five sections. The first section elaborates on the existing research to introduce the content of this study. The second section is to analyze speech synthesis technology and acoustic feature conversion methods, and establish a VITS model, providing a foundation for the subsequent construction of synthetic models. The third section is to optimize and improve the VITS model using domain adaptors and multi-lingual information modules to establish a multi-lingual speech synthesis model. The fourth section is to combine the improved multi-lingual speech synthesis model with the AI field to form a multi-lingual AI speech synthesis model. The final section uses publicly available speech datasets to analyze the evaluation indicators of the proposed speech synthesis model, to demonstrate the reliability and superiority of the research method.

To demonstrate the novelty and innovation of the research design ed method, a summary of the above work has been made, as shown in Table 1.

Table 1: Summary of related works

Reference	Method	Result	Shortcomings
Jiang et al [4]	Using semantic dependencies to extract speech information, convolutional neural networks and self attention mechanisms provide speech training.	Model training can separate tones and phonemes, improving the processing of pauses, stress, and intonation in speech.	The method of extracting speech information from text is prone to limitations and dependencies on contextual information.
Liu et al [5]	The multi-modal and multi-scale text to speech technology provides style embedding and conversion for video input.	Transform speech styles in the field of dubbing and generate high-quality video aligned speech.	The speech extracted from the public grid corpus may have limitations in a single language, making it impossible to achieve dubbing videos between different languages.
Long et al [6]	Multi-distribution method based on Mel frequency cepstral graph and parameter transfer.	The model training proves its effectiveness in classifying language emotions.	The collection environment and equipment for voice data are not universal, and the network's processing of voice is relatively

			complex.
Ghosh et al [7]	In the audio-visual field, visual transformers and attention mechanisms are used to capture the spectral and temporal features of speech.	Improved the quality of speech synthesis.	Only comparing the clarity of speech in the speaking environment lacks quality detection of the synthesized input speech.
Okamoto et al [8]	Use neural encoder and multi-speaker corpus to resample acoustic features.	Its method has high-quality speech rate conversion.	The experimental results of the assistive robot cannot provide reference for the application of real-world scenarios or human voice input.
Masood et al [9]	Generate adversarial networks and deep forgery generation tools to detect audio and video data.	Improved the rapid development of speech conversion and the application of synthesis technology.	There is a lack of more advanced technology in the process and detection of deepfake videos and speech.
Zhang et al [10]	Use cross-language pre training methods for adversarial training of text and image generation.	Increased the robustness of the model to the target language.	Lack of low resource language training for parallel texts in practice can limit the generation of some cross linguistic text to image results.
Kaur et al [11]	Build acoustic models based on deep learning.	Improving the fidelity of synthesized sound, the performance of text to speech synthesis is highly evaluated.	Speech processing heavily relies on the use of synthesizer acoustic models, and the processing of acoustic models is still relatively complex.
Li et al [18]	Use error correction code to improve the Bidirectional Encoder Representation from Transformers model and handle contextual information.	The ability to extract contextual information features and natural language classification has been improved.	Natural language processing is still constrained by the error correcting codes of the model, resulting in a lower ability to recognize its structure.
This study	A domain adaptive multilingual AI sound synthesis model.	In multi-lingual cross-speech synthesis technology, the training efficiency and speech quality of the synthesis model have been improved.	The language input lacks intervention for issues such as accent, and the text model needs to be refined.

2 Methods and materials

Through the development of speech synthesis technology and acoustic feature conversion, the VITS model and the addition of neighborhood adaptors are studied for module improvement, and a series of loss functions are designed to ensure the quality of speech synthesis.

2.1 Speech synthesis technology and acoustic feature conversion methods

In the development process of computers and acoustics, the basic content of speech synthesis technology includes audio processing technology and front-end processing methods for speech synthesis [12-13]. Among them, audio pre-emphasis processing is a common signal processing technique, which enhances the energy of high-frequency signals by adding an emphasis filter to the audio signal, thereby improving the frequency response of the audio signal, as shown in equation (1).

$$J(s) = d(s) - \lambda \times d(s-1) \tag{1}$$

In equation (1), $J(s)$ represents the audio signal after emphasis. $d(s)$ is the initial, unprocessed audio signal. Among them, s is the audio signal, λ is the audio pre-emphasis coefficient, and $[0.9,1.0]$. Fourier transform and wavelet transform are common frequency domain analysis methods in signal processing technology, mainly converting time-domain signals into frequency-domain signals and analyzing the effectiveness of digital signal processing.

In addition, the front-end processing technology of speech synthesis, as an important part of speech synthesis methods, focuses on the processing and conversion of text information, converting human text into appropriate rhythms and vocalizations, thereby providing signals and parameters for the establishment of acoustic models [14]. In the field of speech synthesis, it is necessary to convert the processed acoustic features into playable speech waveforms, as shown in Figure 1.

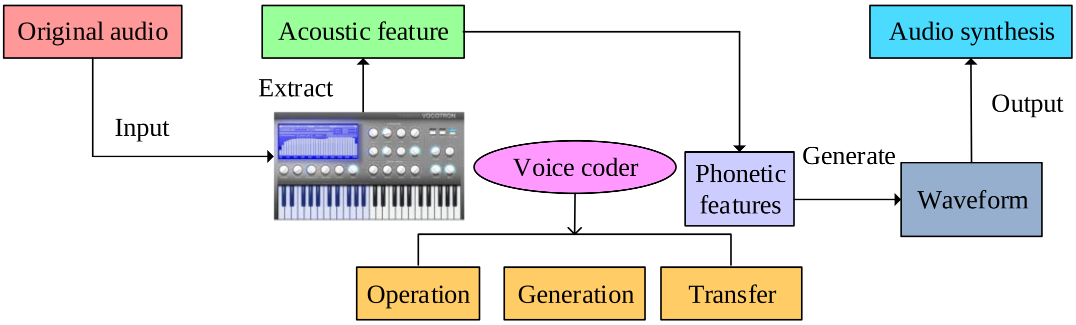


Figure 1: Schematic diagram of the operation of the voice encoder

From Figure 1, the encoder extracts acoustic features from the original audio, and the speech encoder generates audio waveforms through feature operations, generation, and transmission, ultimately achieving the playback of synthesized speech. As the digitalization level improves, codecs have better processing technology in terms of speech quality and compression efficiency. A speech synthesis model based on deep learning technology is developed for end-to-end speech synthesis, which includes modules of encoder and decoder to generate speech waveforms in text sequence input and acoustic feature output. The VITS model is an advanced speech synthesis model in recent years, which shows good performance in generating speech quality and fidelity, as shown in Figure 2.

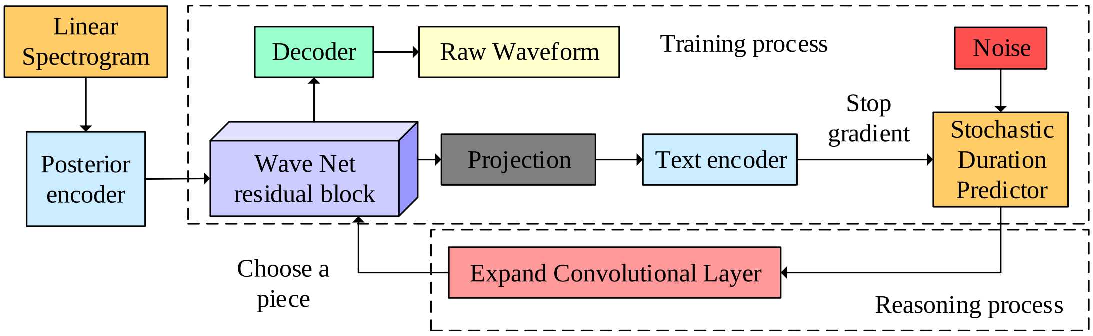


Figure 2: Schematic diagram of the training and inference process of the ITS model

From Figure 2, the structure of the VITS model mainly includes modules such as text encoder, decoder, posterior encoder, and random duration predictor. The original waveform can be converted into text information through the processing of the encoder and residual block. In the inference process, the text encoder is a converter encoder, and the input phonemes are passed through the linear projection layer on top of the text encoder to obtain corresponding representations from the input phonemes. At the same time, the position of this layer is used to construct the mean and variance of the prior distribution, and then the normalization process of monotonic alignment search can be completed. Among them, post encoding selects a residual block from the residual blocks to train the convolutional layer, and embeds the voice of the sound input in the residual block, which can help improve the performance of subsequent speech conversion. In addition, speech noise can be parameterized in the speech layer through random duration predictors and discriminators to improve the performance and quality of the conversion.

2.2 Improvement of domain adaptors and multi-lingual prediction methods

Due to the fact that multilingual speech synthesis technology and its model training have become a research hotspot in cross-lingual communication, a cross-lingual speech synthesis model is established by integrating multi-lingual information and grammar rules into the VITS model. When incorporating multi-lingual information, the VITS model needs to replace the text encoder in the model structure, and add language embeddings and time predictors in the standardized residual block to comprehensively consider the duration differences of language factors. In the VITS model, adversarial learning is used to optimize the variational auto encoder (VAE) to improve the quality of synthesized speech and real-world context [15]. The loss function of VAE is shown in equation (2).

$$L_{VAE}=\left(\mu,\lambda;d\right)=L_{\mathrm{Re}}\left(\mu,\lambda;d\right)+L_{En}\left(\mu,\psi;d\right) \quad (2)$$

In equation (2), ψ represents the optimization parameter for gradient reversal. To maximize the input data, the principle of VAE is studied and explained. Assuming the input data d and its hidden representation z, the relationship between the two is first calculated using conditional probability distribution $y(d|z)$ and prior probability distribution $y(d)$, which also represents the prior distribution of unknown data in the hidden space. Edge likelihood is performed on the posterior distribution of hidden variables, as shown in equation (3).

$$y(d)=\int y(d|z)\times y(z)\times z' \tag{3}$$

In equation (3), $y(d)$ represents the distribution of real data, z is the dimension of the hidden space, and z' is the partial derivative of the dimension. Due to the uncertainty of dimensions, the prior distribution is not a single result, so an approximation of the posterior distribution is introduced to solve the integral. Under variational theory, the prior distribution also needs to be close to its approximate value, as shown in equation (4).

$$y_{\mu}(d)=\int y_{\lambda}(z|d)\times\frac{y_{\mu}(d|z)y(z)}{y_{\lambda}(z|d)}\times z' \tag{4}$$

In equation (4), $y_{\mu}(d|z)$ represents an approximate posterior distribution, $y_{\lambda}(z|d)$ is an approximate prior distribution, and $y(z)$ is a prior distribution. After incorporating multilingual information, the study added a classifier for voice input in the synthesis model, namely the domain adaptors, as shown in Figure 3.

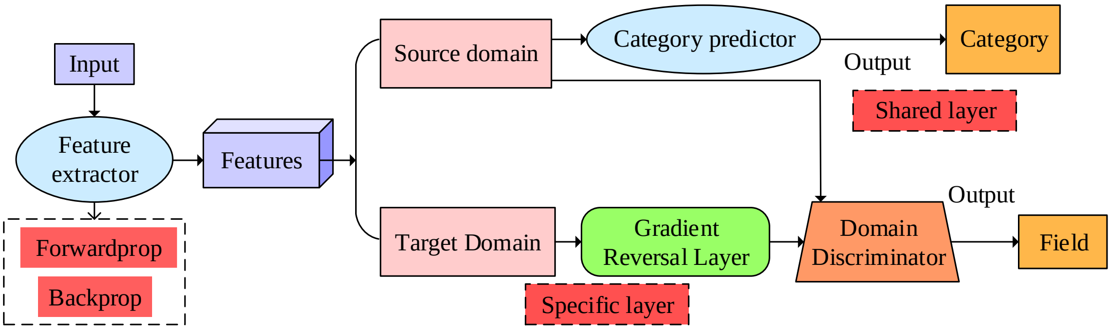


Figure 3: Structure and gradient reverse layer of domain adversarial neural networks

From Figure 3, the domain adversarial neural network will use a feature extractor to separate the features into the source domain and the target in two layers when inputting content. These two layers are the shared layer and the specific layer, with the source domain and target domain as the two assumed datasets in the model, and trained separately at each layer to update weights. In addition, each dataset has a specific classifier and domain classifier used to determine whether the input samples come from the source domain or the target domain. The source domain is located in the category predictor to classify the domain, and the weights are updated through backpropagation. The target domain uses a gradient reverse layer (GRL) in a specific layer to reverse the sign, and then passes it to the shared layer, so that the model ignores the differences between domains [16-17]. In the process of network training, the weights of the shared layer are updated through the source domain classifier and its loss. In the GRL, the classifier's loss is added to the shared layer's loss, and then gradient backpropagation is used to update the weights. Due to GRL being a layer in the network structure, it evolves into negative values through gradient reversal in reverse broadcasting, transferring data from one domain to another, thereby prompting the network model to ignore differences and preserve common data features. In addition, in a specific layer, the target domain needs to be combined with a neighborhood discriminator to determine the type of input domain. At the same time, the introduction of domain classifiers not only achieves classification of each sample, but also helps GRL train its discriminator in backpropagation, making the network model robust to input samples from different datasets, as shown in equation (5).

$$g_{\gamma}(a)=a \quad \frac{dg_{\gamma}}{da}=-\gamma I \tag{5}$$

In equation (5), g represents the gradient backpropagation period, a is the input content in the network structure, and γ is the parameter value of dynamic transformation. dg_{γ} represents the process of reversing the gradient direction, while I represents the identity transformation in the forward propagation process. Therefore, GRL can help train domain classifiers in the backpropagation process, allowing the model to ignore differences between domains. At the same time, gradient inversion to negative values can not only achieve domain adaptation, but also preserve the common features of the data during transmission. Due to the fact that the learning rate parameter also changes with the iteration process in network architecture, to ensure the form of identity transformation, the dynamic expression of the parameter is shown in equation (6).

$$\gamma_L = 2 \times \frac{1}{1+\exp(-\gamma \times L)} - 1 \tag{6}$$

In equation (6), L is the linear variation during the learning process of the network model, and $\exp$ is an exponential function with the natural constant e as the base. This can force domain classifiers to ignore irrelevant information while forcing feature extraction, thereby increasing the model's generalization ability and improving data processing capabilities in other domains. However, to distinguish the accent and text information of the voice input, a regularization loss term is added in the decoupling process between the input and speech information, as shown in equation (7).

$$R = \left\| E_{x \in X}\left[conv(S_x) \right] \right\|_2 \tag{7}$$

In equation (7), R is the regularization loss term, E represents the embedding of the speech input, $conv$ is the convolutional layer, and the kernel is. S_x is the embedding vector of the speech input in data point x. Specifically, the representation of the sound input is extracted, the average value of the embedding vector of the sound input is solved, and then the hidden average value is pushed to the zero vector through a convolutional layer, thereby helping the synthesis model to separate the sound input from its content when generating audio, and the generated sound is not affected by language or accent.

2.3 Construction of multilingual AI speech synthesis model

In the development of the AI era, the field of speech synthesis is closely linked to the processing of AI technology, while speech synthesis technology has been widely applied in the AI field. Therefore, the improved multi-lingual speech synthesis model will be combined with the AI field to form a multilingual AI speech synthesis model, as shown in Figure 4.

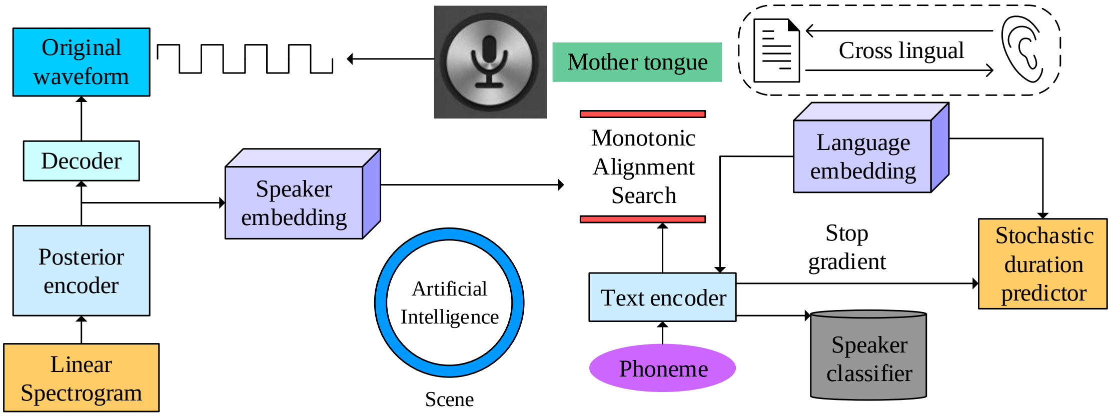


Figure 4: Schematic diagram of the process of multi-lingual AI speech synthesis model

From Figure 4, through the propagation description of text and sound, multi-lingual information input can be achieved, and the task of text editor can be realized under language embedding. The classifier that reuses sound input performs audio processing on the input sound information. The original waveform generated by the input speech is processed by the decoder's processor, which can improve the efficiency of sound generation and the naturalness of synthesis. When processing different audio text data, the synthesis model uses a random duration predictor and classifier for the overall embedding of input text or speech to avoid a large difference in audio length between converted data. This can reduce the instability of the sound synthesis model and solve the task of decoupling phonemes and information of the sound input in the model, achieving multilingual sound synthesis. In addition, in AI scenario applications, the multi-lingual speech synthesis model is integrated with the speaker embedding module to facilitate the implementation of multi-lingual AI speech synthesis models for other languages through sound input and encoder processing.

The improved VITS model utilizes a joint loss function to improve the conditional encoder during training, including recombination loss function, relative entropy loss function, and adversarial function. The first two are used to optimize the conditional VAE, while adversarial loss is used to optimize the discriminator network. The recombination loss function can balance the difference between the recombined speech and the initial speech in the speech synthesis model, as shown in equation (8).

$$L_{\text{Re}} = \left\| d_{MEL} - \hat{d}_{MEL} \right\|_1 \tag{8}$$

In equation (8), L_{Re} represents the difference calculation of the recombination loss function on the audio. MEL represents Mel spectrum. d_{MEL} and $\hat{d}_{MEL}$ are the Mel spectra of the initial audio and generated audio, respectively. The relative entropy loss function is shown in equation (9).

$$L_{En} = \log y_{\lambda}\left(H \middle| d_{lin}\right) - \log y_{\mu}\left(t_C, A_C\right) \tag{9}$$

In equation (9), L_{En} represents the relative entropy loss function, d is the sample of the input model, that is, the speech or text description. H is the output latent variable, and $y_{\lambda}\left(H \middle| d_{lin}\right)$ and $y_{\mu}\left(t_C, A_C\right)$ are the posterior and prior distributions, respectively. Among them, λ and μ are the parameters of the encoder and decoder, respectively. At this point, the use of encoders brings stronger nonlinear capabilities to the model, reducing reliance on coupling layers and streamlining model size. In addition, $\left(t_C, A_C\right)$ is the encoder output result of upsampling in the model structure, which specifically includes the phoneme t_C extracted from the text and its permutation combination A_C with variables. The latent variable is shown in equation (10).

$$H \sim y_{\lambda}\left(H \middle| d_{lin}\right) = N\left(H; M_{\lambda}\left(d_{lin}\right), \sigma_{\lambda}\left(d_{lin}\right)\right) \tag{10}$$

In equation (10), N is the maximum number of samples for the VAE model. N is the mathematical expectation of a normal distribution, and σ is the standard deviation. To provide higher resolution information to the posterior encoder y_{λ}, the above equation uses Mel as the input content. At the same time, to improve the expressive power of prior distributions, a standardized flow is introduced in the text encoder to perform reversible transformations between simple distributions and complex distributions corresponding to latent variables, thereby generating samples that approach reality. However, the reversible transformation of the encoder output is shown in equation (11).

$$y_{\mu}\left(H \middle| t_C\right) = N\left(f_{\mu}(H); M_{\mu}(t_C), \sigma_{\mu}(t_C)\right) \times \left|\det \frac{\partial f_{\mu}(H)}{\partial_H}\right| \tag{11}$$

In equation (11), f is an invertible function, f_{μ} is the mathematical concept of normalized flow, where det represents the determinant and is used to calculate the linear transformation properties of a square matrix, and ∂ is the partial derivative of the latent variable H. Due to the misalignment of the input and extraction of the prior encoder with the true labels, the model estimates the alignment information during training iterations to provide more high-resolution information for subsequent encoders. Therefore, the study adopts a monotonic alignment search normalization process, which is also a search alignment method that maximizes the possibility of parameterized data in the normalized flow, thereby achieving calibration of output text and target sound, as shown in equation (12).

$$A_C = Arg\max \text{Logp}\left(d \middle| t_C, \hat{A}_C\right) \tag{12}$$

In equation (12), $Arg\max$ represents the function expression for finding parameters, and p is the prior distribution of the standardized flow. The random duration predictor in the synthetic model can achieve search calibration using VAE, thereby maximizing the possibility of normalized parameterized data, as shown in equation (13).

$$\underset{\hat{A}_C}{Arg\max}\log p_{\mu}\left(d_{MEL} \middle| H\right) - \log \frac{y_{\mu}\left(H \middle| d_{MEL}\right)}{y_{\mu}\left(H \middle| t_C, \hat{A}_C\right)} = \log N\left(f_{\mu}(H); M_{\mu}\left(t_C, \hat{A}_C\right), \sigma_{\mu}\left(t_C, \hat{A}_C\right)\right) \tag{13}$$

In equation (13), $\hat{A}_C$ represents the maximum degree of sorting combination of hidden variable logarithms, which simplifies the sorting combination while ensuring its rationality and readability. Due to the fact that the duration of input phonemes is extracted from random noise samples by a random duration predictor, its lower limit of variation is shown in equation (14).

$$\log_{y_{\mu}}\left(D \middle| t_C\right) \geq E_{y_{\lambda}\left(u,v \middle| D,t_C\right)}\left[\log \frac{y_{\mu}\left(D-u, v \middle| t_C\right)}{y_{\lambda}\left(u, v \middle| D, t_C\right)}\right] \tag{14}$$

In equation (14), u and v are added random variables to balance the time length. D is the spatial channel, which is the discriminator added in the VITS model to distinguish between the decoder output content and the real waveform. Through variable sampling, high-dimensional transformed integers can be obtained. Therefore, the joint loss function of the multilingual AI speech synthesis model is shown in equation (15).

$$L = L_{\text{Re}} + L_{En} + L_{dur} + L_{Adv} + L_{Fm} \tag{15}$$

In equation (15), L_{dur} is the duration function, L_{Adv} is the adversarial loss function, and L_{Fm} is the linear regression function. Through the conditions of a multilingual AI speech synthesis model and a joint loss function, efficient decoupling of speech input and language information can be achieved in AI applications, thereby synthesizing multilingual speech outputs.

3 Results

Based on the improvement method of multi-lingual AI speech synthesis model, this study used evaluation indicators of speech synthesis technology to train the synthesis model, and tested and analyzed single-language synthesis and cross-lingual synthesis separately to verify the feasibility and superiority of the model.

3.1 Analysis of evaluation indicators for speech synthesis models in a single language

Due to the diversity of languages, multi-lingual speech synthesis data plays an important role in testing the training of synthesis models. Therefore, the research chose speech datasets with multiple scenarios, environments, and continuous updates, and these datasets have broad openness and professional labeling, suitable for analysis in specific fields. Specifically, it includes the Chinese synthesized dataset AIshell-2, while the English synthesized dataset includes the centre for speech technology voice cloning toolkit (CSTR VCTK) dataset, text to speech (LJSpeech) dataset, and library transcriptional speech (libriSpeech) dataset. The Japanese synthesis dataset is a text transcription and reading audio (JSUT) dataset. To ensure the training of the multilingual synthesis model, the selected audio sampling rate was 22.05KHz and the sampling depth was 16 bits.

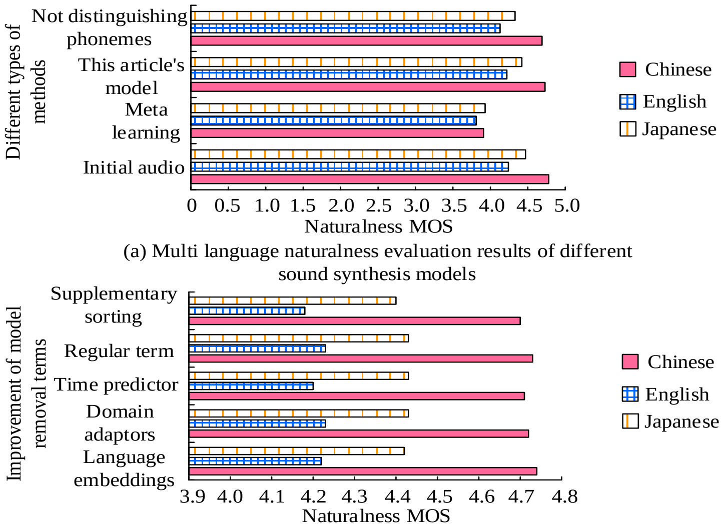
According to the evaluation metrics of speech synthesis technology and speech synthesis models, the natural effects of speech synthesis are usually evaluated using the mean option score (MOS) and similarity methods. The dataset and parameters involved in synthesizing the model are shown in Table 2.

Table 2: Results of dataset selection and model parameter setting

Data sets	The number of speech input users (unit)	Data duration (h)
AIshell-2 Data set	250	400h
CSTR VCTK Data set	500	250h
JSUT Data set	130	40h
Total	880	690h
Model parameters and training configuration	Input classifier	$\gamma = 10$
	Video Memory	Tesla V100 2G
	Batchsize	64
	Training frequency	800 epochs
	Training duration	110h

From Table 2, the multi-lingual speech synthesis dataset was selected, resulting in a total of 880 input data with a duration of 690h. Among them, 250 input data were selected for the Chinese speech synthesis data, and a 300ms silence segment was added to the beginning and end of each language audio data to ensure the unified output of the model during training. Considering the diversity of speech datasets and the processing of synthesis techniques, the model parameters set the weight coefficients of the classifier to 10, and the training times and duration to 800 and 110 hours, respectively.

The MOS evaluation index consisted of a 1–5-point scoring table and an intelligent platform. The study selected a meta learning synthesis model and sequentially removed modules from the proposed model to verify its synthesis effect. This type of model experiment that removes a single component from the model is called ablation experiment, which tests the changes in the initial audio and indistinguishable phonemes to demonstrate the evaluation of language MOS by different components in the model, and also proves the optimization level of cross language synthesis model. The naturalness MOS results are shown in Figure 5.



(a) Multi language naturalness evaluation results of different sound synthesis models

(b) The multilingual evaluation results of different terms in the sound synthesis model presented in this article

Figure 5: Evaluation results of different models on the naturalness MOS of multiple languages

From Figure 5 (a), the initial audio and different speech synthesis models had a naturalness MOS of around 4.4 in multiple languages. The model with the lowest evaluation result was the meta learning speech synthesis model, with naturalness MOS of 3.91, 3.81, and 3.93 for Chinese, English, and Japanese, respectively. The speech synthesis model proposed in the study rated 4.73, 4.22, and 4.42 for the three languages, with the overall evaluation performance being the best. In Figure 5 (b), after removing the domain adaptors, the results of the multilingual naturalness MOS were 4.72, 4.23, and 4.43, respectively. The results after replacing the time predictor were 4.71, 4.20, and 4.43, respectively. The similarity of multiple languages was evaluated and analyzed, as shown in Figure 6.

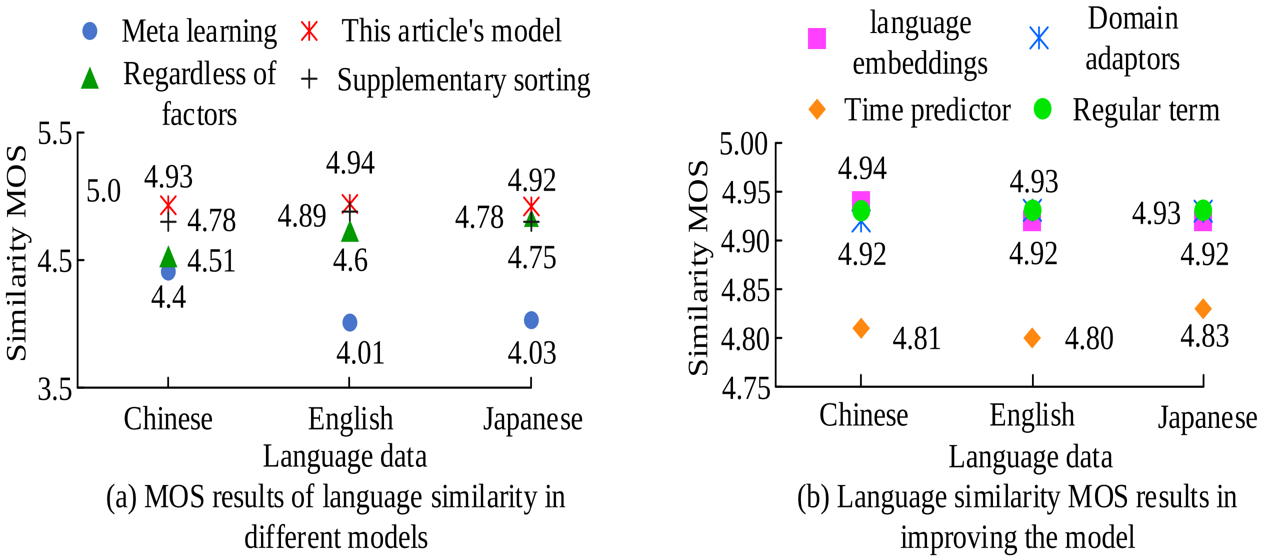
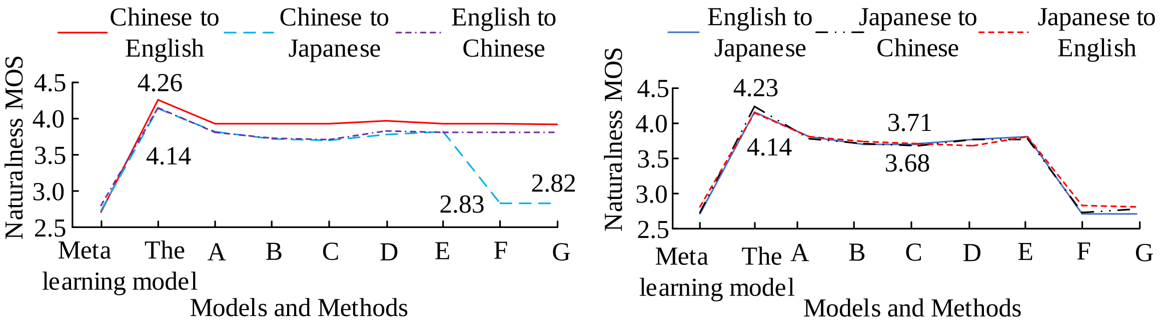


Figure 6: Comparison of MOS similarity results for multiple languages under different models and states

From Figure 6 (a), the similarity results of the three languages in the meta learning synthesis model were relatively low, with values of 4.4, 4.01, and 4.03, respectively. The speech synthesis model without the supplementary sorting component had similar similarity values in the three languages, with values of 4.78, 4.89, and 4.78, respectively. After removing the module that does not distinguish phonemes, the improved synthesis model had the lowest similarity of 4.51 in the Chinese language, while the proposed speech synthesis model had the highest similarity MOS values in the three languages, specifically 4.93, 4.94, and 4.92. In Figure 6 (b), the similarity MOS values of the research model significantly decreased after removing the time predictor, with values of 4.81, 4.80, and 4.83, respectively. After removing the regularization term, the language similarity MOS was 4.93 for all three languages. After removing the domain adaptors, the improved synthesis model had similar similarity scores for the three languages, which were 4.92, 4.93, and 4.93, respectively. Compared with other methods, the improved model with domain adaptors removed had better similarity MOS effect, which proves the reliability of domain adaptors in the model. Taking into account the MOS values of all individual languages, it can be concluded that the proposed speech synthesis model performed well in synthesizing individual languages.

3.2 Analysis of evaluation indicators for cross-lingual speech synthesis models

The multi-lingual AI speech synthesis model proposed by the research, which involves domain adaptors, random duration predictors, and regularization terms in the improvement method, cannot reflect the conversion between language audio and text in the synthesis effect of a single language. Therefore, in the same comparative model and method, the naturalness and similarity of the cross-lingual synthesis effects of Chinese, English, and Japanese were analyzed. The proposed model was translated into A~G to simplify the results by removing language embeddings, domain adaptors, replacement time predictors, regularization terms, supplementary sorting, zero vectors, and uniform phonemes. In the analysis of the naturalness MOS index, the results are shown in Figure 7.



(a) The naturalness results of Chinese English conversion using different model methods

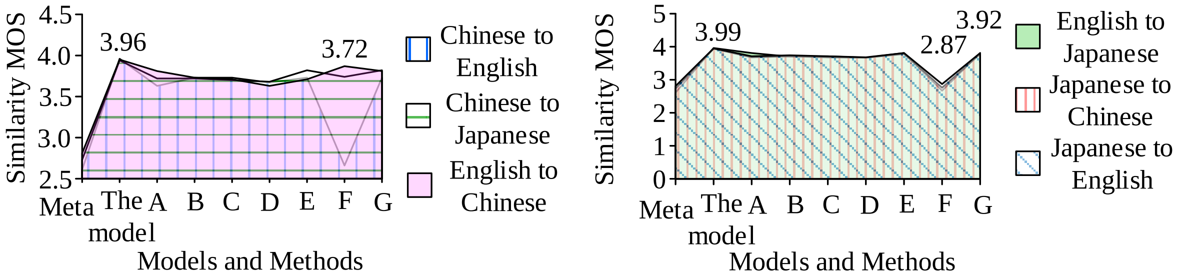
(b) The naturalness results of English Japanese conversion using different model methods

Figure 7: The naturalness MOS results of cross-lingual synthesized speeches using different models

From Figure 7 (a), the meta learning synthesized speech model had the lowest naturalness in Chinese English, Chinese Japanese, and English Japanese conversion, all below 3.0. The naturalness results of the model proposed by the research reached the highest level, with a naturalness MOS value of 4.26 for transit English and 4.14 for speech synthesis naturalness MOS on transit day. When zero vectors and unified phonemes were removed, the naturalness results of the model for Chinese Japanese conversion were the lowest, at 2.83 and 2.82, respectively.

In Figure 7 (b), all models had consistent changes in the naturalness of speech synthesis during the Chienese to Japanese and English to Japanese conversions. When replacing the time predictor values, the result values were 3.71, 3.68, and 3.71, indicating that the multi-lingual AI speech synthesis model had the best naturalness effect in language conversion.

Further analysis was conducted on the cross-lingual speech synthesis model based on similarity, and the results are shown in Figure 8.



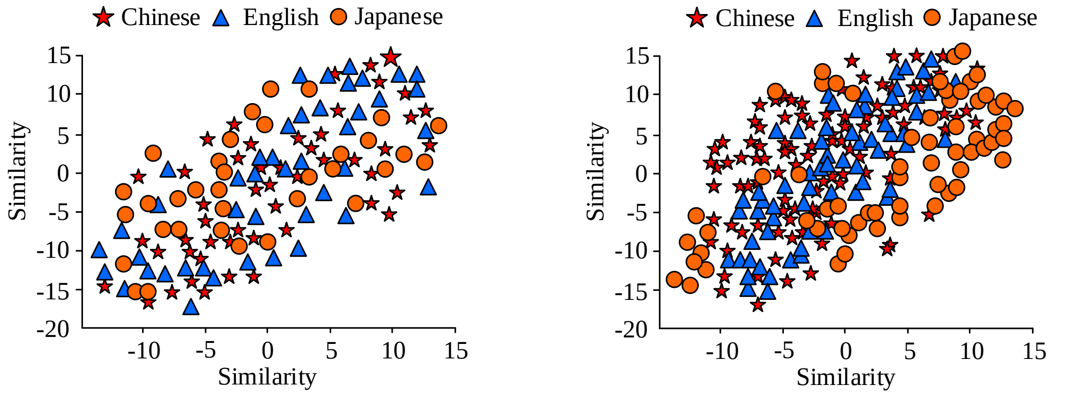
(a) The similarity results of Chinese English conversion using different model methods

(b) The similarity results of English Japanese conversion using different model methods

Figure 8: Similarity MOS results of cross-lingual speech synthesis model

From Figure 8 (a), the similarity of all models for Chinese English conversion was below 4.0. Among them, the similarity MOS values of our model for Chinese English conversion, Chinese Japanese conversion, and English Chinese conversion were 3.96, 3.94, and 3.95, respectively, which were significantly higher than other improved models. In Figure 8 (b), among all the improved synthesis models for cross-lingual conversion, the one with the lowest similarity MOS was the one with zero vector removed. Its MOS values for English Japanese conversion, Japanese Chinese conversion, and Japanese English conversion were 2.67, 2.76, and 2.87, respectively. Therefore, it indicated that the removal of the zero-vector module had a significant impact on the cross-lingual sound synthesis model in cross-lingual training, making the synthesis model less fluent in expressing the rhyme of language input, and the overall naturalness and similarity were lower than those of single language synthesis. This also proved that the proposed model had good fluency and authenticity in cross-lingual sound synthesis.

Afterwards, to minimize the coupling between the speech input and language information, an unsupervised nonlinear technical analysis was conducted on the proposed model and its improvement scheme, using t-distributed stochastic neighbor embedding (t-SNE), as shown in Figure 9.

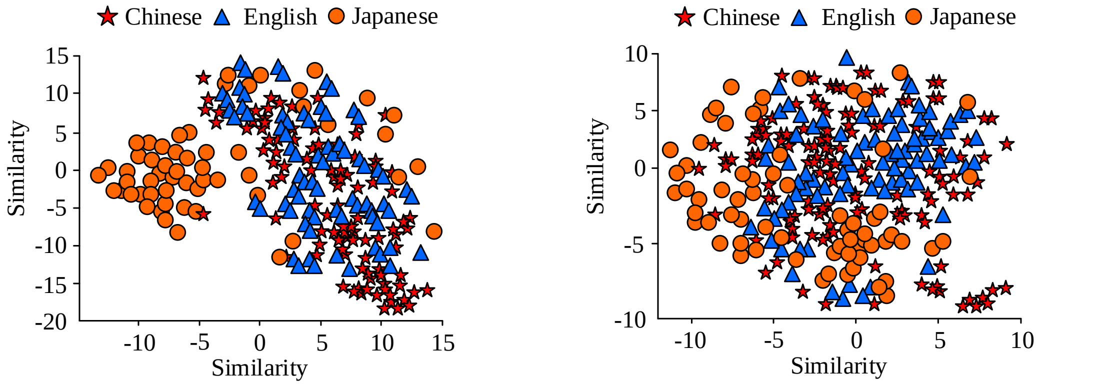


(a) The t-SNE diagram of the model in this article

(b) Remove t-SNE graph of domain adaptors

Figure 9: t-SNE diagram of the model in this article and the method of removing domain adaptors

From Figure 9 (a), the decoupling degree of the proposed model for speech embedding information in different languages was relatively uniform. The t-SNE graph of the three languages showed a good proportion of similarity distribution, which was basically within the range of -15 to 15. In Figure 9 (b), after removing the domain adaptors, the degree of information decoupling in Chinese language was mainly concentrated between -20 and 15, and the distribution in the left domain was relatively dense. However, the decoupling of language information in Japanese was not thorough enough, and the similarity distribution was quite extreme, mainly between 10 and 15. The degree of information decoupling in English language was relatively uniform compared to other languages, with a similarity between -10 and 10. However, the t-SNE graph obtained by removing regularization terms and language embeddings from the research model is shown in Figure 10.



(a) T-SNE graph with regularization term removed

(b) Remove t-SNE graph with language embedding

Figure 10: T-SNE graph with regularization term and language embedding removed

From Figure 10 (a), the model without regularization terms had poor decoupling of language information. Among them, the similarity information of Japanese was more concentrated, which was different from other languages, with a value range of -10 to 5, and the horizontal distribution range was relatively dense, with a value range of -10 to -5. The similarity between Chinese and English languages mainly ranged from -5 to 15, and Chinese language was more densely distributed than English language. In Figure 10 (b), the model without language embeddings had a relatively uniform distribution of t-SNE for languages, and the overall distribution of the three languages was irregular with a wide range. The decoupling degree of Japanese language was mainly on the left and bottom, with values ranging from -10 to 10 and horizontally distributed between -10 and 5. The distribution of English languages was relatively concentrated, ranging from -5 to 5, while the distribution of Chinese languages was relatively even and had a wide range, ranging from -10 to 10. The visualization results of t-SNE above demonstrate that the proposed multilingual AI speech synthesis model has the best decoupling degree for language information.

Finally, different models were used to analyze the accuracy, real-time performance, and scalability of three mcross-lingual synthesized speeches. The Tacotron 2 speech synthesis technology of the spectrogram prediction network and the bidirectional encoder representation from transformers (BERT) technology of the transformer were selected. Tacotron 2 speech synthesis technology is an end-to-end speech synthesis framework that introduces attention mechanisms to address the limitations of traditional models in processing long sequences, improving the quality and naturalness of speech synthesis. In addition, its framework can directly convert text into speech during the training process. However, BERT technology is a pre trained language model based on a bidirectional encoder representation architecture. In the technology of text to speech conversion, it can process text to adapt to the scene and user needs. In text to speech output, more natural and accurate speech can be generated. Compared with the model proposed in this article, the results are shown in Table 3.

Table 3: Performance comparison of different models for cross-lingual speech synthesis

Model	Cross-lingual speech synthesis	Data volume and performance results					
		50		100		200	
		Accuracy (%)	Real time(s)	Accuracy (%)	Real time (s)	Accuracy (%)	Real time (s)
Tacotron 2 speech synthesis framework	Chinese to English	89.03%	5.89s	86.94%	8.12s	79.38%	13.23s
	Chinese to Japanese	87.11%	5.48s	89.71%	8.07s	78.05%	13.49s
	English to Chinese	87.62%	4.19s	84.02%	9.54s	76.98%	15.84s
BERT speech synthesis	Chinese to English	88.67%	5.01s	87.49%	6.49s	83.54%	11.57s
	Chinese to Japanese	87.84%	5.04s	87.55%	5.13s	83.69%	11.62s
	English to Chinese	84.31%	5.62s	86.31%	5.98s	82.11%	10.48s
This article's model	Chinese to English	98.45%	0.68s	95.48%	1.15s	93.67%	3.05s
	Chinese to Japanese	98.36%	0.59s	96.16%	1.54s	94.58%	3.12s
	English to Chinese	97.26%	1.01s	94.59%	2.07s	93.24%	4.08s

From Table 3, as the amount of input speech data increased, the accuracy and real-time performance of the model gradually decreased, while scalability was also related to the amount of data and real-time conversion. Among them, the accuracy of Tacotron 2 speech synthesis technology was the lowest, and when the data volume was 50, the accuracy of Chinese to English conversion was the highest at 89.03%. The speech synthesis framework of BERT had the highest accuracy of 87.49% among 100 data sets. The model proposed by the research had an accuracy rate of up to 94.58% in 200 data sets, and the fastest real-time conversion time was 3.05 seconds, thus proving the superiority of the improved model in multilingual synthesis.

The signal-to-noise ratio (SNR) and Mel Cepstral Distortion (MCD) of 100 speech and audio data were further objectively evaluated to validate the results of MOS and the synthesis quality of the model. The calculation of SNR involved using analytical instruments to measure the power of audio signals and background noise. The higher the value, the stronger the relative strength of the audio signal to the background noise, resulting in more reliable and accurate detection results. MCD is the difference between the Mel frequency cepstral coefficients of synthesized speech and initial speech. After weighted averaging, the smaller the difference, the higher the similarity between synthesized speech and initial natural speech, and the better the speech quality. The results are shown in Table 4.

Table 4: Objective evaluation results of three models on 100 speech audio datasets

Model	Cross-lingual speech synthesis	SNR（dB）	MCD
Tacotron 2 speech synthesis framework	Chinese to English	4.9583 dB	14.56
	Chinese to Japanese	5.1389 dB	13.29
	English to Chinese	4.0534 dB	14.09
BERT text to speech	Chinese to English	4.1059 dB	12.38
	Chinese to Japanese	4.6615 dB	13.47
	English to Chinese	4.8915 dB	12.95
This article's model	Chinese to English	5.2649 dB	9.82
	Chinese to Japanese	5.4184 dB	8.96
	English to Chinese	5.1123 dB	9.51

According to Table 4, the average SNR of the Tacotron 2 speech synthesis framework for three language conversions was 4.7169, The average SNR of BERT technology was 4.5530. The model proposed in the study had a high SNR for speech and audio, which was 0.7122 higher than the average SNR of BERT technology, indicating that its model has accurate detection results for synthesized speech. In MCD calculation, the proposed model was much lower than other models, with MCD values of 9.82, 8.96, and 9.51 in Chinese English, Chinese Japanese, and English Japanese speech and audio, respectively. The above results demonstrate that the proposed sound synthesis model has a higher similarity to natural speech and better synthesis quality.

4 Discussion

Speech synthesis technology has become an important part of human-computer interaction in the field of AI. Research proposed a multi-lingual AI speech synthesis model based on domain adaptors, and module improvements were made to the VITS model to adapt to the technical operations of multilingual conversion. In the context of multi-lingual background and application

requirements, domain adaptors were incorporated into the VITS model and other module structures were improved to solve the coupling relationship between speech input phonemes and language information, thereby improving the quality and output of multilingual speech synthesis. This is because domain adaptors use classifiers and predictors to weight speech datasets in the structure of adversarial neural networks, which can improve the model's generalization ability and add a regularization loss term to the subsequent decoupling process. The regularization loss term utilized the structure and convolutional layers of neural networks to separate the audio and content of the synthesized model, thereby achieving decoupling of language information in the audio input domain. After analyzing the evaluation indicators of speech synthesis technology, it was found that the improved speech synthesis model had the highest naturalness MOS of 4.73 and similarity of 4.93 in the Chinese language. However, in cross-lingual conversion, the naturalness and similarity of the synthesized model decreased due to significant differences in pronunciation methods such as factors, rhythm, and rhyme between different languages. At the same time, the speech conversion process also needed to consider the duration of input and output, which made the speech synthesis not smooth and realistic enough. The model proposed in the study had a synthetic similarity MOS of 3.96 in English, which was lower than the similarity of 4.93 and 4.94 in Chinese and English language synthesis. Compared with Li et al., the extraction of contextual information features and natural language classification problems were improved. This is because BERT's ability to recognize the structure of natural language processing was still limited by the model's error correction codes [18]. The multi-lingual speech synthesis model proposed in the study could achieve the task of text editor through language embedding when integrating AI scenes, processing input audio to improve the naturalness of generation. The random duration predictor could reduce the instability in the speech synthesis process and combine it with monotonic alignment search to achieve the alignment of the target sound. Therefore, in AI applications, cross-lingual speech synthesis models can not only separate the input voice and content while outputting, but also ensure the fluency and authenticity of AI synthesized sounds across multiple scenarios and languages.

5 Conclusion

Cross-lingual speech synthesis technology is currently the main research direction in the field of speech synthesis. The VITS model and its algorithm were studied, and module technology and domain adaptors were improved to construct a domain adaptive multilingual AI speech synthesis model. According to the analysis of the synthesis indicators of Chinese, English, and Japanese languages, the naturalness MOS of these three languages in the meta learning sound synthesis model were 3.91, 3.81, and 3.93, respectively. By analyzing the synthesis indicators of Chinese, English, and Japanese languages, a meta learning speech synthesis model was obtained, with a naturalness MOS of 3.91, 3.81, and 3.93 for Chinese, English, and Japanese, respectively. The evaluation of the proposed speech synthesis model for the three languages was 4.73, 4.22, and 4.42, respectively. In the cross-lingual speech synthesis experiment, the improved synthesis model that removes zero vectors and unified phonemes achieved naturalness MOS values of 2.83 and 2.82 for Chinese to Japanese conversion, respectively. The model proposed in this paper had the highest similarity MOS values for Chinese English, Chinese Japanese, and English Chinese, which were 3.96, 3.94, and 3.95, respectively. After removing the domain adaptors, the degree of information decoupling in Chinese language was mainly concentrated between 0 and 10, and the t-SNE graph distribution was uneven, indicating the feasibility of a multilingual AI speech synthesis model based on domain adaptors. In the cross-lingual conversion test, the accuracy of our model for language conversion was above 93%, and the conversion efficiency was the fastest. However, the BERT model was affected by a large amount of data resources during training and computation, which reduced its synthesis processing performance. Therefore, it proves the superiority of the research model. However, research on the quality of synthesized audio across languages still lacks specific factors to consider, and there is a lack of intervention in issues such as accents for language input types. There is also a lack of corresponding constraints for model training. Therefore, in future research, the text model in the early stage should be refined to train high-quality phoneme and audio data, and a reference module should be added to constrain the accent of the voice input to improve the quality of synthesized audio. In addition, in the parameter training process of multi-lingual cross-speech synthesis technology, the idea of distillation model can be borrowed to effectively reduce the training workload of the synthesis model and ensure the feasibility and accuracy of the model.

Fundings

The research is supported by Shanxi Provincial Education Science "14th Five Year Plan" 2023 Education Digitalization and Teaching Reform Research Special Project, Research on the Teaching Path of "Dual-Qualification Teacher Classroom" Assisted by Virtual Digital Human, Project Number: SZH-230013.

References

[1] Sachdeva S, Ruan H, Hamarneh G, Behne D M, Jongman A, Sereno J A, Wang Y. Plain-to-clear speech video conversion for enhanced intelligibility. International journal of speech technology, 2023, 26(1): 163-184. https://doi.org/10.1007/s10772-023-10018-z

[2] Maruoka M, Tsujimura S, Asakura T. Effects of Artificial Synthetic Speech Control of SNR and Speech Rate on the Intelligibility of Train Station Announcements. Acoustics Australia, 2024, 52(1): 77-86. https://doi.org/10.1007/s40857-023-00306-8

[3] Revathi A, Sasikaladevi N, Arunprasanth D, Amirtharajan R. A Strategic Approach for Robust Dysarthric Speech Recognition. Wireless Personal Communications, 2024, 134(4):2315-2346. https://doi.org/10.1007/s11277-024-11029-y

[4] Jiang C, Gao Y, Ng W W Y, Zhou J, Zhong J, Zhen H, Hu X. Semantic dependency and local convolution for enhancing naturalness and tone in text-to-speech synthesis. Neurocomputing, 2024, 608(12): 1-11. https://doi.org/10.1016/j.neucom.2024.128430

[5] Liu Y, Wei L F, Qian X C. M3TTS: Multi-modal text-to-speech of multi-scale style control for dubbing. Pattern recognition letters, 2024, 179(3):158-164. https://doi.org/10.1016/j.patrec.2024.02.005

[6] Long L, Liang T. Multi-Distributed Speech Emotion Recognition Based on Mel Frequency Cepstogram and Parameter Transfer. Chinese Journal of Electronics, 2022, 31(1): 155-167.

[7] Ghosh S, Sarkar S, Ghosh S, Zalkow F, Jana N D. Audio-visual speech synthesis using vision transformer–enhanced autoencoders with ensemble of loss functions. Applied Intelligence, 2024, 54(6):4507-4524. https://doi.org/10.1007/s10489-024-05380-7

[8] Okamoto T, Matsubara K, Toda T, Shiga Y, Kawai H. Neural speech-rate conversion with multispeaker WaveNet vocoder. Speech Communication, 2022, 138(3): 1-12. https://doi.org/10.1016/j.specom.2022.01.003

[9] Masood M, Nawaz M, Malik K M, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, 2022, 53(4): 3974-4026. https://doi.org/10.1007/s10489-022-03766-z

[10] Zhang H, Yang S, Zhu H. CJE-TIG: Zero-shot cross-lingual text-to-image generation by Corpora-based Joint Encoding. Knowledge-based systems, 2022, 239(3): 108006-108017. https://doi.org/10.1016/j.knosys.2021.108006

[11] Kaur N, Singh P. Conventional and contemporary approaches used in text to speech synthesis: a review. Artificial Intelligence Review, 2022, 13(11): 1-44. https://doi.org/10.1007/s10462-022-10315-0

[12] Bahraini T, Sadigh A N. Proposing a robust RLS based subband adaptive filtering for audio noise cancellation. Applied acoustics, 2024, 216(1):109755-109765. https://doi.org/10.1016/j.apacoust.2023.109755

[13] Wang J, Yue K, Duan L. Models and techniques for domain relation extraction: a survey. Journal of Data Science and Intelligent Systems, 2023, 1(2): 65-82. https://doi.org/10.47852/bonviewjdsis3202973

[14] Higuchi Y, Moritz N, Roux J L, Hori T. Momentum Pseudo-Labeling: Semi-Supervised ASR With Continuously Improving Pseudo-Labels. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1424-1438. https://doi.org/10.1109/jstsp.2022.3195367

[15] Zhou Y, Ding Z, Liu X, Shen C, Tong L, Guan X. Infer-AVAE: An attribute inference model based on adversarial variational autoencoder. Neurocomputing, 2022, 483(4): 105-115. https://doi.org/10.1016/j.neucom.2022.02.006

[16] Meng J, Zhu F. Seek for commonalities: Shared features extraction for multi-task reinforcement learning via adversarial training. Expert Syst. Appl. 2023, 224(8): 119975-119986. https://doi.org/10.1016/j.eswa.2023.119975

[17] Haifeng Z, Fengqian Z, Shengtian S, Li Y, Li X, Hu K, Chen Y. An unsupervised intelligent fault diagnosis research for rotating machinery based on NND-SAM method. Measurement Science & Technology, 2023, 34(3): 35906-35922. https://doi.org/10.1088/1361-6501/aca98f

[18] Li S, Hu X, Huang Z Z J. ECC-BERT: Classification of error correcting codes using the improved bidirectional encoder representation from transformers. IET communications, 2022, 16(4): 359-368. https://doi.org/10.1049/cmu2.12357