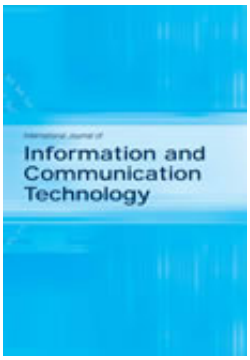
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Personalised English listening teaching design based on natural language processing and speech synthesis

Yanling Han

DOI: 10.1504/IJICT.2025.10070833

Article History:

Received:	10 March 2025
Last revised:	21 March 2025
Accepted:	22 March 2025
Published online:	06 May 2025

Personalised English listening teaching design based on natural language processing and speech synthesis

Yanling Han

Teaching Department of Public Courses,
Hunan Communication Polytechnic,
Changsha 410132, China
Email: hanyanling123@126.com

Abstract: Traditional English listening instruction tends to use a one-size-fits-all model, making it difficult to meet individualised learning needs. For this reason, this paper firstly analyses the English listening teaching text based on natural language processing (NLP), and designs encoders and decoders based on multi-head long- and short-term self-attention to convert the text feature sequences into Mel Spectrograms. The Mel spectrogram is then converted into a speech waveform using an improved generating adversarial network (GAN) generative model, and the grouped convolution-based discriminant model is responsible in distinguishing between real and synthesised speech, prompting the generative model to synthesise more realistic speech waveforms. Finally, a personalised application model of the proposed text to speech (TTS) method in English listening teaching is constructed. The experimental outcome indicate that the proposed method not only improves students' performance, but also the synthesised speech has a high degree of naturalness.

Keywords: English listening instruction; natural language processing; NLP; text to speech; TTS; attention mechanism; generative adversarial network.

Reference to this paper should be made as follows: Han, Y. (2025) 'Personalised English listening teaching design based on natural language processing and speech synthesis', *Int. J. Information and Communication Technology*, Vol. 26, No. 11, pp.21–37.

Biographical notes: Yanling Han received her Master's degree from Hunan University in 2011. She is currently a Lecturer in Hunan Communication Polytechnic. Her research interests include digitally enabled teaching, intelligent classroom, natural language processing and applied linguistics.

1 Introduction

Nowadays, with the increasing internationalisation of communication, English language proficiency has become more and more important. English listening, as the 'input' of language, is crucial for the formation of learners' phonological sense (Gilakjani and Ahmadi, 2011). For a long time, due to the influence of exam-oriented education in the college entrance examination, English teachers tend to focus on reading and writing, which are the main subjects for knowledge mastery, and neglect the teaching of listening.

This bias leads to relatively strong reading and writing skills, while listening and speaking skills lag behind (Yang, 2019). The rapid rise of information technique has created more possibilities and choices for English listening instruction. However, due to the complexity of technology operation, difficulty in accessing resources, inapplicability and other problems, the leading role of students and teachers is ignored, increasing the burden of teachers and learners (Aji, 2017), so that teachers and students do not know how to adapt to the new technological environment of English listening teaching, and students' interest cannot be well stimulated, and the results of the English listening teaching obtained are not very good (Fang, 2021). Therefore how to use artificial intelligence technology to realise efficient English listening teaching design and improve the quality of English teaching is a topic of practical value.

In English listening classroom teaching, teachers can select personalised teaching resources and complete the input of the textual information to be converted, and the text to speech (TTS) technology can convert it into the corresponding sound information. Nakai (2019) states that the traditional mode of teaching English is showing more and more problems, while the use of TTS technology can improve the listening skills of learners. Fitria (2023) proposed the use of TTS technology to create TOEFL listening questions, and his results showed that the technology has achieved realism comparable to real speech. Early TTS methods mainly include TTS based on speech-in-speech (Khan and Chitode, 2016), resonance peak TTS (Adiga and Prasanna, 2019), and tandem splicing TTS (Mattheyses and Verhelst, 2015), but the speech synthesised by these methods is unnatural and affective. With the growth of statistical machine learning, the Statistical Parametric Speech Synthesis (SPSS) approach (Reddy and Rao, 2020) has been proposed, which focuses on establishing parameters such as fundamental frequency, spectrum, and continuous time in acoustic models. Acoustic models generally use linguistic features to correspond to acoustic feature mapping relationships, such as those based on hidden Markov model speech synthesis (Kayte et al., 2015).

Neural networks have a powerful self-learning function that can automatically extract features from a large amount of data and optimise their internal parameters through training, which can be adapted to different tasks and scenarios. Valizada et al. (2021) researchers explored the exponential improvement of TTS driven by deep learning, showing through subjective ratings and behavioural measurements that modern speech synthesis systems are close to natural speech. Wajdi et al. (2021) improved the English speech synthesis system text2speech in various ways to achieve high quality English TTS. Yan (2024) fused Transformer and text2speech to design an intelligent TTS. Huang et al. (2020) proposed to take the English teaching text as input, then generate the corresponding Meier acoustic spectrograms by Seq2Seq prediction network which introduces the self-attention mechanism (SAM), and finally use the LPCNet model to reduce to the speech waveforms, but the naturalness of the synthesised speech is not high. Valin and Mustafa (2024) proposed an end-to-end TTS approach based on natural language processing (NLP) techniques, which uses a pre-trained encoder to extract contextual information from the input textual content, drastically reducing the training parameters and training time.

The key technology in TTS is the vocoder, i.e., the reduction of speech parameters to speech waveforms. Yasuda et al. (2021) proposed the use of an autoregressive convolutional neural network (CNN) to implement the design of the vocoder in TTS, synthesising speech with high quality but slow speed. Kong et al. (2020) based on an generating adversarial network (GAN), HiFi-GAN, which can simultaneously achieve

high efficiency and high fidelity sound synthesis, modelling the sound cycle and effectively improving the quality of sound samples. Yu et al. (2023) conducted a comprehensibility experiment based on HiFi-GAN and applied it to the innovation of English listening teaching mode to improve the quality of English teaching.

The research of TTS has achieved a large number of research results and has been applied in English listening teaching to a certain extent. This paper designs a personalised English listening teaching method based on NLP and TTS. The main research work of the method is as follows.

1 Based on the NLP analysis of English teaching text, each word in the text sequence is encoded as a sub-phonemes sequence, and the phoneme sequence and the expanded sub-phonemes sequence are aligned and summed to get the mixed phoneme sequence, which is converted into a vector representation.

2 The encoder and decoder based on multi-head long short-term self-attention are designed. The encoder is responsible for converting the linguistic feature embedded sequences into linguistic feature hidden sequences, and the decoder is responsible for converting the linguistic feature hidden sequences into Mel spectrograms for further synthesis of speech waveforms by TTS.

3 A TTS method based on GAN and speech time-frequency features is designed. The generative model adopts residual units that are more suitable for modelling speech signals, and is responsible for synthesising speech waveforms using Mel spectrograms; the discriminant model based on group convolution is responsible for distinguishing between real and synthesised speech in different speech domains, forcing the generative model to synthesise waveforms that are closer to real speech.

4 Combined with the proposed TTS method, a personalised application model in English listening teaching is constructed. The experimental results show that the equal error rate (EER) of the proposed method is reduced by at least 0.26%, which not only synthesises high-quality speech, but also has an obvious promotion effect on the improvement of students' performance.

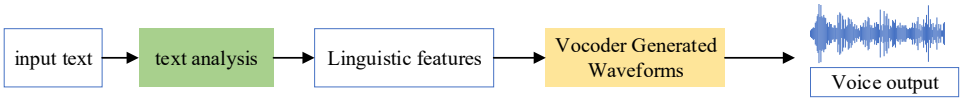
2 Relevant technologies

2.1 Speech synthesis system architecture

TTS technology is dedicated to converting textual information into speech signals, with the aim of producing smooth and natural speech output through speech synthesis systems. The TTS process involves a text front-end, an acoustic model, and a vocoder module (Adam, 2020), and its system architecture is shown in Figure 1.

The text front-end module converts the input Chinese text into a sequence of phonemes, covering such sub-modules as text normalisation, word segmentation, and polyphony prediction, whose processing accuracy directly affects the quality and naturalness of the final speech output. In the acoustic modelling stage, the system further converts the sequence of phonemes into a sequence of audio features, mainly using features such as the Meier spectrogram and fundamental frequency. The vocoder module algorithmically reduces the output of the acoustic model to a raw audio waveform, completing the conversion from text to auditory output.

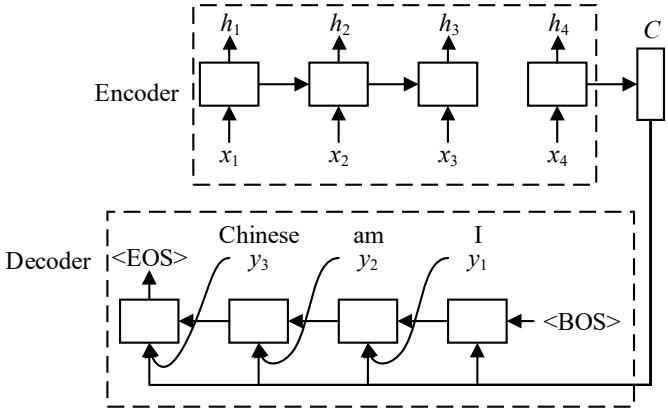
Figure 1 TTS system architecture (see online version for colours)



2.2 Sequence-to-sequence model

RNN can realise the processing of sequence data and map the source sequence to the target sequence of the same length. However, in the TTS task, the lengths of the source and target sequences are not the same, and RNN alone cannot realise the relevant functions. To solve the issue, Seq2Seq (Zhang and Xiao, 2018) is proposed, which consists of an encoder, a semantic vector (context) and a decoder.

Figure 2 Seq2seq model



The encoder encodes all inputs into a uniform semantic vector, which is decoded by the decoder. The front-end models of deep learning-based TTS systems all use seq2seq models to realise the conversion of text to acoustic features. The seq2seq encoder and decoder can use RNN, LSTM and GRU architectures. Figure 2 shows the schematic diagram of seq2seq model.

2.3 Generating adversarial networks

Compared with the traditional TTS method, GAN has stronger generation capability and higher flexibility. Compared with CNN and other classical deep learning models, GAN can generate high-quality images, audio, text and other data, especially in the field of image generation, and can generate a variety of samples, to avoid the problem of pattern collapse, and generate a data distribution closer to the real data. It can not only generate diverse speech samples, but also control the quality and diversity of the generated speech by adjusting the network structure of the generator G and discriminator D as well as the training parameters. This makes GAN have a greater potential for application in TTS.

G is responsible for capturing the training data distribution and generating data consistent with the distribution of the training samples. The objective of G is to maximise

the probability of error of D in order for D to determine the data it generates as real data, with the following objective function.

$$\begin{aligned}&\max_{\theta_g}\left(E_{z\sim p_z(z)}\left[\log\left(D\left(G\left(z;\theta_g\right);\theta_d\right)\right)\right]\right)\\&=\min_{\theta_g}\left(E_{z\sim p_z(z)}\left[\log\left(1-D\left(G\left(z;\theta_g\right);\theta_d\right)\right)\right]\right)\end{aligned}\tag{1}$$

where θ_g is the parameter of G, θ_d is the parameter of D, $p_z(z)$ is the prior distribution of the input noise variable, $D(x, \theta_s)$ is the probability that the sample x comes from the true distribution, and $G(z, \theta_g)$ is the projection function.

D is responsible for recognising whether the current data belongs to the data generated by the generative model or the real training data, the backbone of D is usually a CNN, and distinguishing whether x comes from the real distribution $P_{data}(x)$ or from the generative distribution $p_g(x)$ is the core goal of D. Using the label $y = 1$ to indicate that the samples are from the true distribution and $y = 0$ to indicate that the samples are from the generative distribution, then having following equations.

$$p(y=1\mid x)=D\left(x;\theta_d\right)\tag{2}$$

$$p(y=0\mid x)=1-D\left(x;\theta_d\right)\tag{3}$$

Discriminator D to minimise the cross-entropy as an objective function, as shown in equation (4).

$$\min_{\theta_d}\left(E_x\left[y\log p(y=1\mid x)+(1-y)\log p(y=0\mid x)\right]\right)\tag{4}$$

3 TTS front-end module design based on NLP and attention mechanism

3.1 Text analysis of English listening instruction based on NLP

The text front-end module in TTS mainly utilises NLP technology to design the preliminary conversion of the original English teaching text into a format and information suitable for subsequent TTS processing, and designs the multi-head long and short-term self-attention (MLSAM) to convert the mixed phoneme hidden sequences into Mel Spectrogram sequences, which are predicted to get the prediction of the English pronunciation rhythms through the prediction of deep learning algorithms. The structure of the designed TTS front-end module is shown in Figure 3.

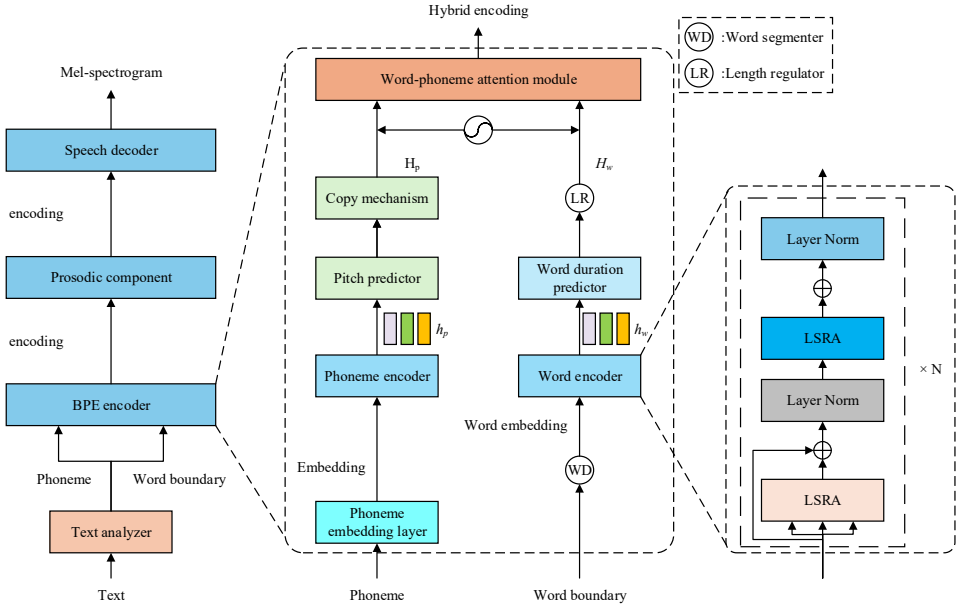
Phonological sequences only provide phonetic information, and for the goal of providing contextual semantic information to the TTS system, the BPE-based phoneme coding algorithm was introduced to encode each word in a text sequence as a sub-phonological sequence (Držík and Forgac, 2024). The phoneme sequences and the expanded subphoneme sequences are aligned and summed to obtain the hybrid phoneme sequences, which are converted to vector representations for processing by the neural network. A commonly used vector representation is one-hot coding (Rodríguez et al., 2018), which is simple to encode and fast to run. Only one bit in the one-hot vector has a value of 1, and the rest of the bits have a value of 0. The One-hot representation of a

sequence of mixed phonemes is fed into the embedding layer to generate the mixed-phonemes embedding, as shown below.

$$M_v = M_p M_e \tag{5}$$

where M_v is the matrix consisting of hybrid phoneme embedding vectors; M_p is the matrix consisting of one-hot vectors; and M_e is the hybrid phoneme embedding matrix.

Figure 3 The structure of the designed TTS front-end module (see online version for colours)



In order to utilise the positional information of each phoneme, it is necessary to add a positional code to the initial input sequence. The position encoding has the same dimension as the input sequence, and by adding the two, a new sequence containing the position information can be obtained, as shown below.

$$pe_{t,2d} = \sin\left(t / 10000^{\frac{2d}{D}}\right) \tag{6}$$

$$pe_{t,2i+1} = \cos\left(t / 10000^{\frac{2d}{D}}\right) \tag{7}$$

where pe is the position coding vector; t is the position; d is the dimension; $pe_{t,2d}$ is the $2i^{\text{th}}$ dimension of the coding vector for the t^{th} position; D is the dimension of the original input vector.

3.2 *Encoders and decoders based on multi-head long- and short-term self-attention*

The input $X^{(0)}$ of the encoder is obtained by adding the position encoding PE to M_v, where e_t is the mixed phoneme embedding vector representation of the t^{th} word and pe_t is the

vector representation of the position. After obtaining the embedded sequence of mixed-phonemes containing PEs, the encoder is responsible for converting it into a hidden sequence of mixed-phonemes. The encoder consists of a stack of four MLSAM modules, where each layer accepts the output of the previous level as input.

$$X^{(0)} = M_v + PE = \left[e_1 + pe_1; ...; e_N + pe_N\right] \tag{8}$$

Although traditional AM can capture the important features in the sequence, the computational complexity is high. Therefore, this paper draws on the existing research to introduce the long-short-term AM (LST), which replaces the original AM by combining the short-term AM and the long-term AM, so as to achieve lower computational complexity, and the detailed information of LST can be found in the literature (Zhu et al., 2021), which will not be repeated here. Since LST cannot handle input sequences with edge length and multiple features, this paper replaces AM in LST with SAM and combines the idea of multi-head AM to innovatively propose MLSAM as shown in equation (9) to equation (12).

$$MLSAM(X) = Concat\left(H_1, H_2, ..., H_i, ..., H_h\right)W^o \tag{9}$$

$$H_{i,t} = Soft\max\left[\frac{X_t W_i^Q \left[LN_s\left(\tilde{X}_t W_i^K\right); LN_L\left(\bar{X}_i\right)\right]^T}{\sqrt{D_k}}\right]\left[LN_S\left(\tilde{X}_t W_i^V\right); LN_L\left(\bar{X}_i\right)\right] \tag{10}$$

$$\bar{X}_i = P_i^T X W_i^K \tag{11}$$

$$P_i = Soft\max\left(XW_i^P\right) \tag{12}$$

where X is the input matrix; H_i is the i^{th} LST; W^o is the output projection matrix; W_i^Q, W_i^K, W_i^V, and W_i^P are the learned projection matrices; $LN_S(\cdot)$ is the layer normalisation; $\tilde{X}_t$ is the input matrix within the local window; $\bar{X}_i$ is the global low-rank projection input matrix; and P_i denotes the dynamic low-rank projection matrix at the i^{th} attention head.

After the MLSAM layer, the output vectors at each phoneme position are obtained through a feed-forward network (FFN) as follows.

$$FFN(x) = W_2 \operatorname{Re} LU\left(W_1 x + b_1\right) + b_2 \tag{13}$$

where W_1, W_2, b_1 and b_2 are network parameters; Relu() is the ReLU activation function.

The linear transformation performed in the FFN is the same at different locations, but with different parameters between layers. Considering the high correlation between the neighbouring hidden states in the phoneme sequences and the Mel Spectrogram sequences, a one-dimensional convolutional layer is used instead of the linear transformation layer in the FFN. The role of the decoder is to convert the hidden sequence of mixed phonemes into a Mel Spectrogram sequence for further synthesis of speech waveforms by TTS. The entire decoder consists of a stack of six MLSAMs, with each position in each layer accepting the output of all positions in the previous layer as input.

3.3 English pronunciation tone prediction

Due to the variability of real tones, it is difficult to accurately predict the pitch values by directly predicting the metrical contours, so the normalised tone contours are converted to pitch spectrograms using the continuous wavelet transform (CWT), as shown below.

$$W(\tau,t)=\tau^{-\frac{1}{2}}\int_{-\infty}^{+\infty}F_0(x)\psi\left(\frac{x-t}{\tau}\right)dx \tag{14}$$

where τ is the scale, t is the translation, $F_0(x)$ is the pitch value, and $\psi(\cdot)$ is the wavelet mother function.

The labels consist of a pitch spectrogram, the mean of the pitch contour and the variance of the pitch contour. The MSE loss is used to quantify the pitch predictor predictions with the true pitch labels as follows.

$$F_0(t)=\int_{-\infty}^{+\infty}\int_{0}^{+\infty}W(\tau,t)\tau^{\frac{5}{2}}\psi\left(\frac{x-t}{\tau}\right)dxd\tau \tag{15}$$

The pitch embedding sequence is obtained by quantising the pitch of each frame into logarithmic scale values and encoding the quantised values. The final pitch embedded sequence is added to the original mixed-phonemes hidden sequence to extend the mixed-phonemes hidden sequence.

4 High-quality TTS based on generative adversarial networks and speech features

4.1 Generative models based on residual networks

After designing the front-end of TTS, this paper designs a TTS method based on GAN and speech features. The residual network-based generative model is responsible for converting the Mel spectrogram into speech waveforms; the grouped convolution-based discriminant model is responsible for distinguishing between real speech and synthesised speech, prompting the generative model to synthesise more realistic speech waveforms. Compared to similar vocoders, this vocoder is faster to train, has fewer parameters, and has higher synthesis quality. The architecture of the proposed TTS is shown in Figure 4.

Unlike the original GAN, the designed generative model does not use random noise, but takes the Mel Spectrogram as input, and finally outputs the corresponding speech waveform of the Mel Spectrogram. The internal structure of the generative model is shown in Figure 5.

In the residual cell MRF, the cascaded network layers include a one-dimensional null convolution layer and a one-dimensional convolution level. Cavity convolution uses the insertion of cavities between neighbouring elements of the convolution kernel to expand the coverage of the convolution kernel, thus increasing the sense field of each output time step while keeping the number of parameters constant and not losing the input information. Assuming that the original size of the convolution kernel is $K \times K$, the effective size of the convolution kernel is $(K + (K - 1) \times (D - 1)) \times (K + (K - 1) \times (D - 1))$ in dilated convolution. Use $[B, C_{in}, L_{in}]$ for the size of the input tensor and $[B, C_{out}, L_{out}]$ for the size of the output tensor. The following relationship exists between the

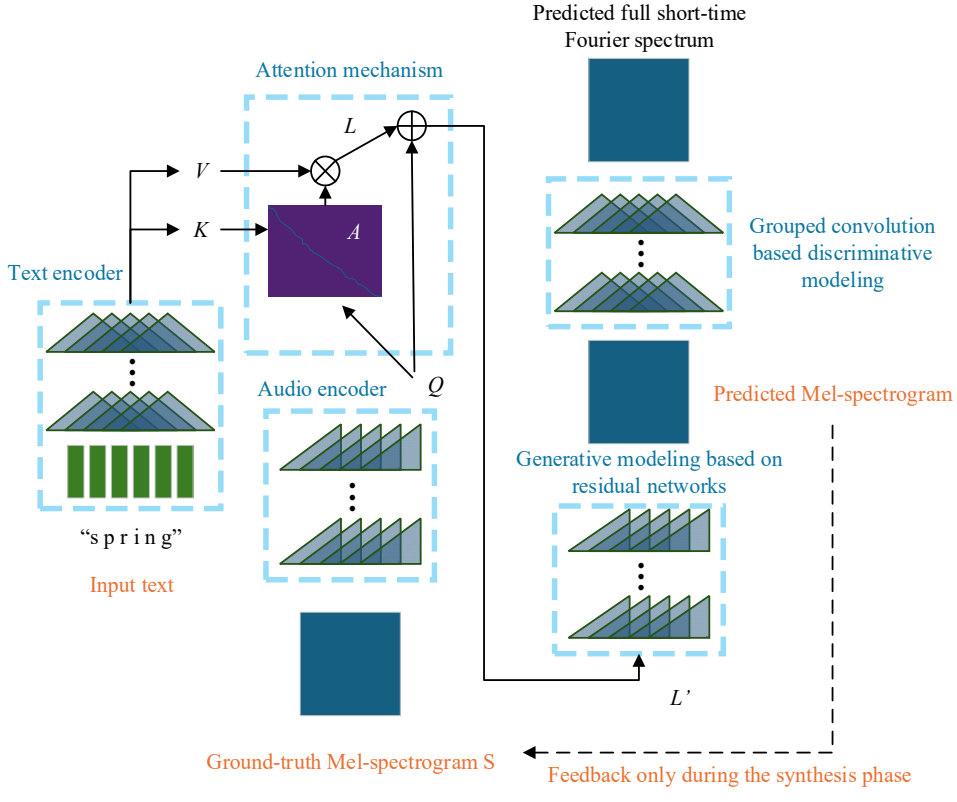
number of columns of the output matrix and the number of columns of the input matrix of a 1D null convolution.

$$L_{out} = \left\lfloor \frac{L_{in} + 2 \times P - D \times (K-1) - 1}{S} + 1 \right\rfloor \tag{16}$$

$$y = \tanh(x_1) \odot \sigma(x_2) \tag{17}$$

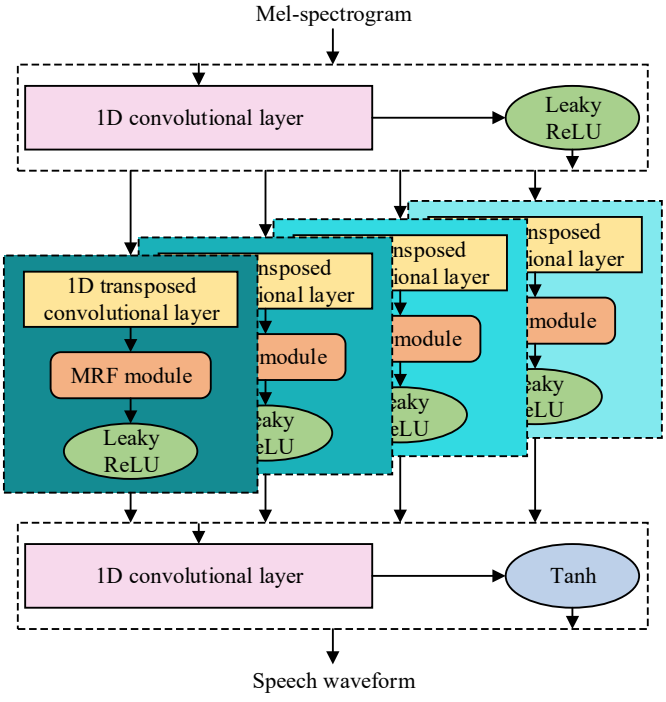
where P is the padding size, D is the dilation factor, K is the size of the convolution kernel, and S is the step size. Two sets of data x_1 and x_2 can be obtained by dividing the output of the 1D null convolutional layer along the feature vector dimension. The input of the 1D convolutional layer is obtained by the following equation, where y is the input of the 1D convolutional layer, $\sigma(\cdot)$ is the logistic function, $\odot$ is the Hadamard multiplication operator, and E is the rule of multiplying the corresponding positional elements.

Figure 4 The architecture of the proposed TTS (see online version for colours)



When training the TTS vocoder, B training samples are taken at each iteration, where B is called the batch size. At each iteration, the generative model is fed with B Mel spectrograms of size $C \times N$, where C represents the Mel frequency dimension of the Mel spectrogram and N represents the number of speech frames. The synthesised speech waveform is output by the generative model designed above.

Figure 5 The internal structure of the generative model (see online version for colours)



4.2 Grouped convolution-based discriminative models

To learn the features within different speech time-frequencies, three sub-time-frequency discriminant models are applied to three sets of speech spectra at different scales. Although discrete wavelet transform (DWT) can also analyse the speech spectrum, fast Fourier transform (FFT) is computationally simple and has better localisation capability compared to DWT, so three sets of speech time-frequencies at different scales are obtained by performing FFT with different parameters on the input speech. The whole sub-time-frequency discrimination model is composed of a series of one-dimensional convolution layers and one-dimensional grouping convolution layers, in which the one dimensional grouping convolution layer adopts convolution kernel with large K value.

For one-dimensional block convolution, the size of each set of inputs is $(C_{in} / g) \times L_{in}$, the size of the convolution kernel is $(C_{in} / g) \times K$, and the number of convolution nuclei is C_{out} / g. The final result can be obtained by concatenating the output tensors of group g, as shown below.

$$X^{k+1} = \left[X_1^k \otimes W_1^k ; X_2^k \otimes W_2^k ; ... ; X_g^k \otimes W_g^k \right] \tag{18}$$

$$X^k \otimes W^k = \left[X^k * w_1^k ; X^k * w_2^k ; ... ; X^k * w_N^k \right] \tag{19}$$

where X^{k+1} is the output of the k^{th} network layer, X_i^k is the i^{th} input of the k^{th} network layer, W_i^k is the i^{th} convolution kernel of the k^{th} network layer, $\otimes$ is the one-dimensional convolution operation, and N is the number of convolution kernels in the group.

Spectral normalisation is a normalisation technique used to stabilise the training of the discriminant model, which has the advantages of simple implementation and small computation, and the basic idea is to ensure that the whole discriminant model meets Lipschitz continuity by constraining the spectral parameter of each weight matrix W. The spectral parameter of W is defined as follows.

$$\sigma(W) = W_2 = \sqrt{\lambda_{\max}\left(W^H W\right)} \tag{20}$$

where $\lambda_{\max}(\cdot)$ is the maximum eigenvalue of the computational input matrix. The spectral normalisation technique uses the following equation to spectrally normalise each of the weight matrices W, such that the new weight matrix W_{SN} satisfies the Lipschitz constraint $\sigma(W_{SN}) = 1$.

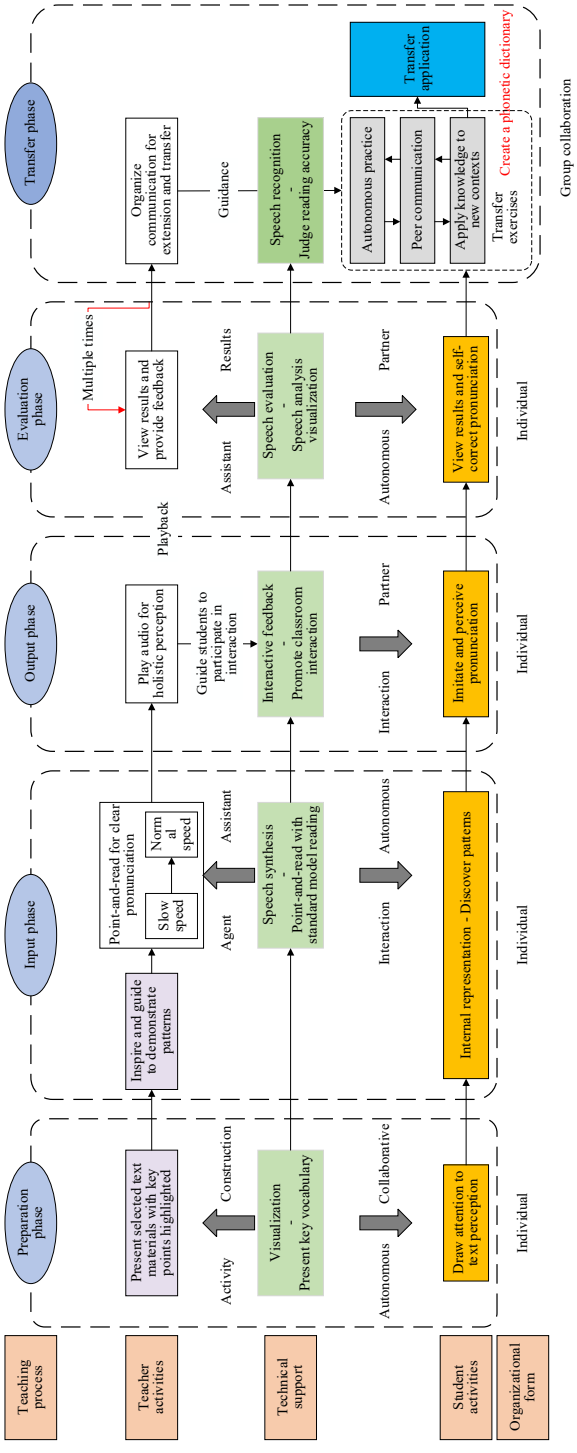
$$W_{SN} = \frac{W}{\sigma(W)} \tag{21}$$

5 Designing personalised English listening instruction based on NLP and TTS

With the proposed TTS technology, teachers can convert relevant knowledge outside the classroom into audio and play it back to the learners to extend their knowledge according to their teaching needs. In addition, the teacher can adjust the speed and timbre of the listening material to take into account the learner's learning characteristics, and read aloud or generate MP3 audio files of any word, sentence or article. The speed of speech can be adjusted, and male and female voices can be selected, thus providing learners with listening materials that are close to their learning and life contexts, in addition to the audio specific to the textbook. According to the psychological process of listening comprehension and information processing mode within learners, combined with the various functional advantages of the proposed TTS technology, this study constructs a personalised application model of TTS technology in English listening teaching as shown in Figure 6.

Through a series of processes such as 'Preparation phase-input phase-output phase-evaluation phase-transfer phase' to build a high-quality English listening teaching environment. In this environment, teachers mainly play the role of the master of intelligent technology advantages and the builder of collaborative activities, while TTS mainly plays the role of agent, assistant and interactor. Due to the collaborative environment constructed by teachers and TTS, learners can give full play to their wisdom and behave more autonomously and cooperatively in the whole teaching process. Assume the role of reading activity collaborator, independent reading trainer and group knowledge builder. Teachers, TTS and learners work together to clarify the pronunciation rules of different letter combinations, so that students can accurately recognise English phonetics and can apply what they have learnt to achieve the effect of learning by example.

Figure 6 The personalised application model of TTS technology in English listening teaching (see online version for colours)



6 Experimental results and analyses

In this paper, the proposed method OURS is used to analyse the listening scores of two classes of students in a city high school, and the pre-test results of the two classes are shown in Table 1. The proposed TTS method is also trained using the LJSpeech dataset for testing. The LJSpeech dataset consists of 13,100 English audio clips of approximately 24 hours duration, with corresponding text. In this experiment, the dataset is divided into two parts: the first 13,000 sentences are used for training, and the last 100 sentences are used for testing. The experiments were conducted on Ubuntu 16.04 and RTX2080Ti graphics card. The headlock model was trained for 100 rounds with Adam optimiser and the learning rate was set to 0.02. The experiments were conducted on Ubuntu16.04 and RTX2080Ti. The hyperparameters in the network model were set as follows: a one-dimensional convolution with a convolution kernel of 3 and a discard rate of 0.5 for pitch, and on-the-fly discard training.

Table 1 Pre-testing of listening scores in two classes

Norm	*Groups*	*Class size*	*Grade point average*	*Standard deviation*	*Standard error mean*
Score	Test group	45	16.04	5.63	1.831
	Comparison group	45	15.99	5.48	1.756

Table 2 Post-testing of listening scores in two classes

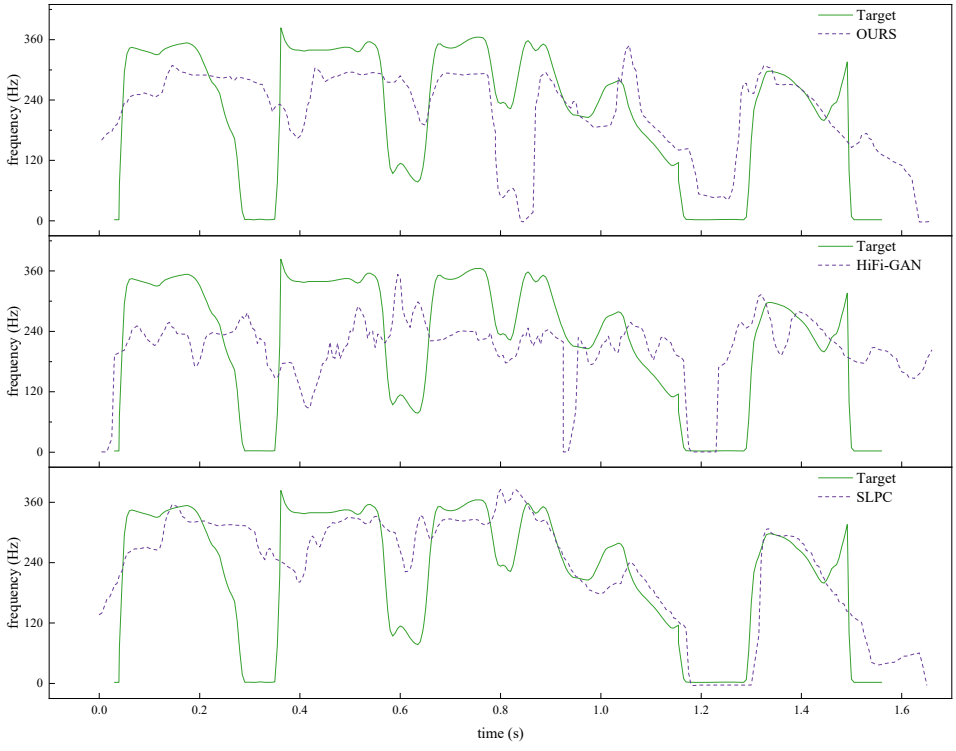
Norm	*Groups*	*Class size*	*Grade point average*	*Standard deviation*	*Standard error mean*
Score	Test group	45	36.58	2.36	1.1294
	Comparison group	45	31.69	2.01	1.0137

After applying the OUS method to the design of the teaching programme, the English listening scores of the students in the two classes were tested again using the LJSpeech dataset, in which the corresponding environments and the time of conducting the test were the same, and the total listening scores after applying the designed TTS method to the test are shown in Table 2. The mean value of the standard error of the pre-test is greater than the mean value of the standard error of the post-test, and from the mean value of the performance, the performance of the test group is 36.58, and the average performance of the control group students is 31.69, which is greater than the control group, which can show that the listening teaching method based on the proposed TTS design has a significant role in promoting the improvement of the students' performance.

In addition to evaluating student performance after the improvement of teaching methods, this paper also conducts comparative experiments on OUS, SLPC (Huang et al., 2020), and HiFi-GAN (Yu et al., 2023) using the Mel cepstral distortion (MCD), the EER, and the mean opinion score (MOS) metrics. The comparison of the speech synthesised by the different methods with the target fundamental frequency curve is shown in Figure 7. It can be seen that both SLPC and Hi-Fi-GAN have loss of realism in speech synthesis, and deviate from the target fundamental frequency curve, compared with the other two methods, the speech contour trend and detail changes of OUS synthesis are closer to the target speaker, especially in the transformation of high frequency features (contour detail trend), which makes the speech contour closer to the

target speaker in details, and has the advantage of generating speech with rhythmic features (e.g., intonation) of the target speaker. It has the advantage of generating speech with rhythmic features (e.g., intonation) of the target speaker. OURS not only performs NLP on English teaching text, innovatively proposes MLSAM to enhance key features and reduce the computational complexity, but also designs residual network-based generative model and group convolution-based discriminative model, which makes the synthesised speech closer to the real scene and improves the synthesis effect.

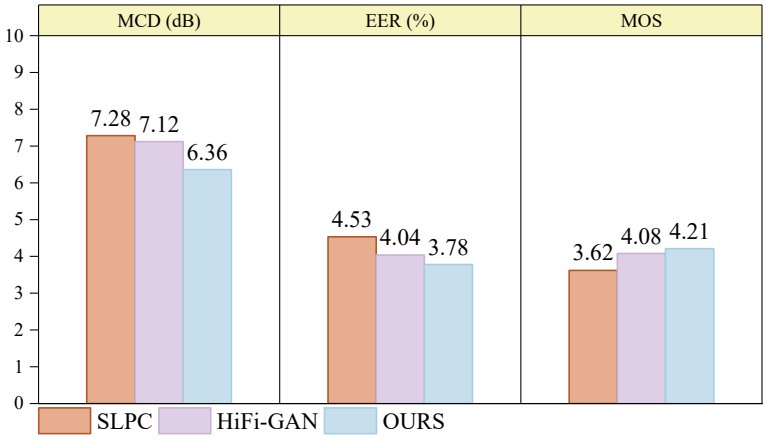
Figure 7 Comparison of synthesised speech and target fundamental frequency curves (see online version for colours)



Comparison of MCD, EER and MOS for different methods is shown in Figure 8. The lower the EER value, the higher the convenience of the TTS. The EER of OURS is 3.78%, which is reduced by 0.75% and 0.26% compared to SLPC and Hi-Fi-GAN, respectively, indicating that OUS accurately identifies target speakers and rejects non-target speakers. MCD is a metric for evaluating the difference between two Mel Spectrum cepstrum sequences, and lower MCD values usually imply better speech synthesis quality. The MCDs of SLPC, Hi-Fi-GAN and OURS are 7.28 dB, 7.12 dB and 6.36 dB, respectively, indicating that the closer the cepstrum coefficient of the OUS clock synthesised speech to the real speech, the more natural the synthesised speech is. In this paper, the MOS value of each method is obtained by collecting subjective evaluations from individual students. The MOS value of OURS is 4.21, which is 0.59 and 0.13 higher than SLPC and Hi-Fi-GAN respectively. Upon comparison, OUS is able to generate speech quality that is highly close to that of human beings, which significantly improves

the naturalness of synthesised speech, and thus contributes even further to the improvement of the quality of English listening teaching.

Figure 8 Comparison of MCD, EER and MOS for different methods (see online version for colours)



7 Conclusions

TTS technology is widely used in the field of English listening teaching, but the poor quality of speech synthesis of existing TTS methods makes it impossible to significantly improve the quality of English listening teaching. To this end, this paper first analyses the ELT text using NLP to convert the original ELT text into a format and information suitable for subsequent TTS processing, and designs an encoder and decoder for MLSAM to convert the mixed phoneme hidden sequences into Mel spectrogram sequences. Then the TTS method based on GAN and speech time-frequency features is designed. The residual network-based generative model is responsible for converting the Meier spectrogram into speech waveforms, and the frequency domain discriminant model based on packet convolution distinguishes between real speech and synthesised speech, which together motivate the generative model to synthesise more lifelike speech waveforms. Finally, based on the psychological process of listening comprehension and information processing mode within learners, a personalised application model in English listening teaching is constructed by combining the proposed TTS method. The experimental results show that the proposed method can not only synthesise high-quality speech, but also has an obvious promotion effect on the improvement of students' performance.

Declarations

All authors declare that they have no conflicts of interest.

References

Adam, E.E.B. (2020) 'Deep learning based NLP techniques in text to speech synthesis for communication recognition', *Journal of Soft Computing Paradigm (JSCP)*, Vol. 2, No. 4, pp.209–215.

Adiga, N. and Prasanna, S. (2019) 'Acoustic features modelling for statistical parametric speech synthesis: a review', *IETE Technical Review*, Vol. 36, No. 2, pp.130–149.

Aji, M.P.P. (2017) 'English listening blended learning: the implementation of blended learning in teaching listening to university students', *Kajian Linguistik Dan Sastra*, Vol. 2, No. 1, pp.25–32.

Držík, D. and Forgac, F. (2024) 'Slovak morphological tokenizer using the byte-pair encoding algorithm', *PeerJ Computer Science*, Vol. 10, p.e2465.

Fang, S. (2021) 'Talk about the problems of English listening in teaching and the problems in the process of students' learning', *Advances in Educational Technology and Psychology*, Vol. 5, No. 10, pp.90–96.

Fitria, T.N. (2023) 'Using natural reader: a free text-to-speech online with AI-powered voices in teaching listening TOEFL', *ELTALL: English Language Teaching, Applied Linguistic and Literature*, Vol. 4, No. 2, pp.1–17.

Gilakjani, A.P. and Ahmadi, M.R. (2011) 'A study of factors affecting EFL learners' English listening comprehension and the strategies for improvement', Journal of Language Teaching and Research, Vol. 2, No. 5, pp. 977-988.

Huang, L., Zhuang, S. and Wang, K. (2020) 'A text normalization method for speech synthesis based on local attention mechanism', *IEEE Access*, Vol. 8, pp.36202–36209.

Kayte, S., Mundada, M. and Gujrathi, J. (2015) 'Hidden Markov model based speech synthesis: a review', *International Journal of Computer Applications*, Vol. 130, No. 3, pp.35–39.

Khan, R.A. and Chitode, J.S. (2016) 'Concatenative speech synthesis: a review', *International Journal of Computer Applications*, Vol. 136, No. 3, pp.1–6.

Kong, J., Kim, J. and Bae, J. (2020) 'Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis', *Advances in Neural Information Processing Systems*, Vol. 33, pp.17022–17033.

Mattheyses, W. and Verhelst, W. (2015) 'Audiovisual speech synthesis: an overview of the state-of-the-art', *Speech Communication*, Vol. 66, pp.182–217.

Nakai, A. (2019) 'Use and effects of an online text-to-speech resource to improve English listening for the TOEIC test', *Journal of Tourism Studies*, Vol. 11, p.13.

Reddy, M.K. and Rao, K.S. (2020) 'Excitation modelling using epoch features for statistical parametric speech synthesis', *Computer Speech & Language*, Vol. 60, p.101029.

Rodríguez, P., Bautista, M.A., Gonzalez, J. and Escalera, S. (2018) 'Beyond one-hot encoding: Lower dimensional target embedding', *Image and Vision Computing*, Vol. 75, pp.21–31.

Valin, J-M. and Mustafa, A. (2024) 'Very low complexity speech synthesis using framewise autoregressive GAN (FARGAN) with pitch prediction', *IEEE Signal Processing Letters*, Vol. 31, pp.2115–2119.

Valizada, A., Jafarova, S., Sultanov, E. and Rustamov, S. (2021) 'Development and evaluation of speech synthesis system based on deep learning models', *Symmetry*, Vol. 13, No. 5, p.819.

Wajdi, M., Sanjaya, I.N.S. and Sumartana, I.M. (2021) 'Developing a listening English learning model using text2speech application', *Journal of Applied Studies in Language*, Vol. 5, No. 2, pp.274–281.

Yan, X. (2024) 'A teaching mode of college English listening in intelligent phonetic environments', *International Journal of e-Collaboration (IJEC)*, Vol. 20, No. 1, pp.1–17.

Yang, X. (2019) 'On the obstacles and strategies in English listening teaching', *Theory and Practice in Language Studies*, Vol. 9, No. 8, pp.1030–1034.

Yasuda, Y., Wang, X. and Yamagishi, J. (2021) 'Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis', *Computer Speech & Language*, Vol. 67, p.101183.

Yu, C., Wu, L., Li, J. and Li, S. (2023) 'English listening teaching mode under artificial intelligence speech synthesis technology', *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 20, No. 6, pp.1–6.

Zhang, Y. and Xiao, W. (2018) 'Keyphrase generation based on deep seq2seq model', *IEEE Access*, Vol. 6, pp.46047–46057.

Zhu, C., Ping, W., Xiao, C., Shoeybi, M., Goldstein, T., Anandkumar, A. and Catanzaro, B. (2021) 'Long-short transformer: efficient transformers for language and vision', *Advances in Neural Information Processing Systems*, Vol. 34, pp.17723–17736.