Applied Intelligence (2022) 52:15193–15209
https://doi.org/10.1007/s10489-021-03075-x

Limited text speech synthesis with electroglottograph based on Bi-LSTM and modified Tacotron-2

Lijiang Chen[1] · Jie Ren[1] · Pengfei Chen[1] · Xia Mao[1] · Qi Zhao[1]

Accepted: 2 December 2021 / Published online: 12 March 2022

Abstract
This paper proposes a framework of applying only the EGG signal for speech synthesis in the limited categories of contents scenario. EGG is a sort of physiological signal which can reflect the trends of the vocal cord movement. Note that EGG's different acquisition method contrasted with speech signals, we exploit its application in speech synthesis under the following two scenarios. (1) To synthesize speeches under high noise circumstances, where clean speech signals are unavailable. (2) To enable dumb people who retain vocal cord vibration to speak again. Our study consists of two stages, EGG to text and text to speech. The first is a text content recognition model based on Bi-LSTM, which converts each EGG signal sample into the corresponding text with a limited class of contents. This model achieves 91.12% accuracy on the validation set in a 20-class content recognition experiment. Then the second step synthesizes speeches with the corresponding text and the EGG signal. Based on modified Tacotron-2, our model gains the Mel cepstral distortion (MCD) of 5.877 and the mean opinion score (MOS) of 3.87, which is comparable with the state-of-the-art performance and achieves an improvement by 0.42 and a relatively smaller model size than the origin Tacotron-2. Considering to introduce the characteristics of speakers contained in EGG to the final synthesized speech, we put forward a fine-grained fundamental frequency modification method, which adjusts the fundamental frequency according to EGG signals and achieves a lower MCD of 5.781 and a higher MOS of 3.94 than that without modification.

Keywords Electroglottograph (EGG) · Speech Synthesis · Bi-LSTM · Tacotron

1 Introduction

In 1970, Fant [1] set up the Source-Filter model, a classical acoustic modeling method, which provided a promising approach to conduct speech synthesis researches. The Source-Filter model represented speeches as the combination of a source and a linear acoustic filter, corresponding to the vocal cords and the vocal tract (soft palate, tongue, nasal cavity, oral cavity, etc.), respectively.

Electroglottograph (EGG) records electrical impedance in the glottis collected by electrodes situated on the throat and can reflect the vocal cord movement. When the vocal cord closes, the contact area between the two cords reaches its maximum, which leads to the lowest resistance and highest collected voltage in EGG. Conversely, when the vocal cord opens, the lowest collected voltage will be collected [2]. Figure 1 illustrates the waveform of a EGG signal sample. Phase 1, 2, 3, 4 represent the closing phase, maximum contact, opening phase and open but no contact phase, respectively. Depending on the periodical change of the amplitude of EGG signals during speaking, we can mark pitching and obtain the source information, which has been researched by Hussein [3].

In respect that EGG is directly collected from throat, there are two definite superiority of the speech analysis based on the EGG: 1. It is not affected by mechanical vibration and noise so it is suitable to apply in the ultra-high noise environment; 2. It can accurately reflect the vocal cord vibration state information.

✉ Qi Zhao
zhaoqi@buaa.edu.cn

Lijiang Chen
chenlijiang@buaa.edu.cn

Jie Ren
rj980728@buaa.edu.cn

Pengfei Chen
chenpengfei0104@buaa.edu.cn

Xia Mao
moukyou@buaa.edu.cn

[1] Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing, China

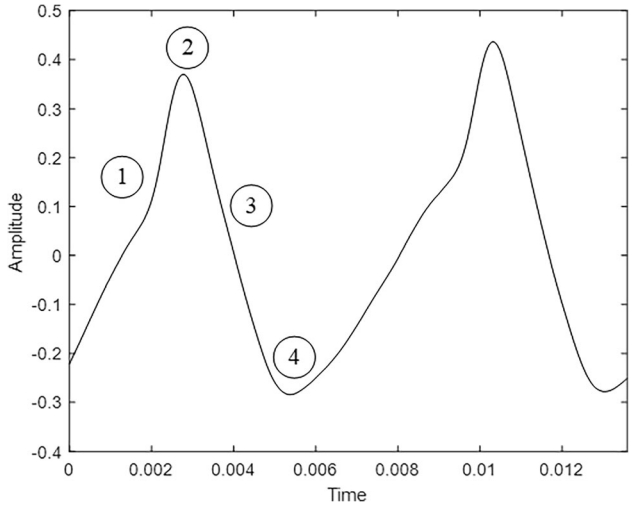


Fig. 1 the waveform of EGG. Phase 1: the closing phase; Phase 2: maximum contact; Phase 3: opening phase; Phase 4: open but no contact phase

As the EGG signal highly corresponds to speaking, lots of researches have been carried out about EGG. Aiming at exploring the characteristics of EGG signals, Paul figured out that some features, such as gender, vowel, and phonatory registers can be extracted from the EGG signal [4]. Lu discussed the relationship between EGG and emotions [5]. Alberto evaluated EGG signal variability by amplitude-speed combined analysis [6]. As for utilizing EGG signals, Chen proposed a speech emotional features extraction method based on EGG [7]. Michal Borsky utilized EGG signal as a feature type to investigate its performance for voice quality classification task [8]. Sunil Kumar put forward a robust method to detect glottal activity using the phase of the EGG signal [9]. Liu compared the parametrization methods of EGG signals in distinguishing between phonation types [10]. Lebacq analyzed the dynamics of vocal onset through the shape of EGG signals [11]. Filipa discovered the immediate effects of using a flow ball device for voice exercises, which can assist voice training [12].

Focusing on speech synthesis researches, traditional speech synthesis aims at utilizing the information of the raw speech to recover it, which consists of waveform synthesis method, rule-based synthesis method, and synthesis method based on parameters. Waveform synthesis mainly refers to edit and joint waveform, which has limited performance. The rule-based synthesis method produces speeches through phonetic rules. Belonging to it, PSOLA is a representative algorithm for waveform splicing and prosody control [13]. However, this method requires a large volume of sound libraries, making it difficult to be applied to portable devices. The parameter-based synthesis method mainly refers to speech synthesis with acoustic features. Lots of well-known synthesis systems implemented this method such as Klatt series-parallel formant synthesizer [14], LPC [15], LSP [16], and LMA [17]. Additionally, char2wav [18], straight [19], WORLD [20], vocaine [21], Mel spectrum [22] and other models have also achieved good results.

Speech synthesis with the deep learning method seeks to convert text to speech, which is mainly achieved by extracting deep features. Oord [23] proposed a deep neural network model named WaveNet for generating original audio waveform signals. In 2017, Baidu put forward deep voice [24], which replaced the traditional method with the neural network at different levels, and applied the WaveNet model to the final speech synthesis module. Another widely-used model named Tacotron [25] was proposed by Google, which is an end-to-end generative text to speech model and achieves to directly learn the mapping between the text and speech pair. Later, Baidu launched deep voice2 [26] and deep voice3 [27] to improve and modify the previous generation model. Then, based on Tacotron, several improvements have been put forward to tackle different problems, in particular focus on specific characteristics when dealing with other languages. In Japanese text-to-speech (TTS) tasks, Yasuda [28] included self-attention to Tacotron to capture long-term dependencies of pitch accents. Liu [29] designed a distillation loss function to modify the feature loss function and proposed a teacher-student training scheme based on Tacotron to solve the exposure bias problem. To improve the naturalness and tackle the prosodic problems in Mandarin TTS, several solutions have also been figured out. Yang [30] proposed SAG-Tacotron which replaced the CBHG encoder of Tacotron with the self-attention-based one and utilized learnable Gaussian bias to enhance localness modeling and overcome the problem of self-attention's dispersing the distributions of attention. Another popular direction is to design an extra front-end, realized by Lu [31] who proposed a text enhancing method and tried to leverage previous phrasing models and larger text database at the same time, and Pan [32] who set up a unified front-end to solve polyphone disambiguation and prosody word prediction. However, as all the speech synthesis methods mentioned above are text-to-speech, it is not suitable to generate personalized speech in our application scenario.

Attending to the inherent superiority of EGG, we have made efforts to collect EGG and speech signals simultaneously ant built up a database named Chinese Dual-mode Emotional Speech Database (CDESD [33]), which provides a basis for the research of EGG, especially in the Mandarin speech research. In our previous study, we have proved that EGG signals can be used in text content recognition [34]. Two long Chinese sentences with different contents often vary in the vocal cord movement, which can be reflected by the EGG signal. Thus, it is reasonable to convert the EGG signal into one sentence under the condition of limited class of contents. We extracted the fundamental frequency (F_0), the relative difference of F_0 ($diffF_0$) and log short-term energy ($logE$) from the EGG signal in every frame to combine into a feature

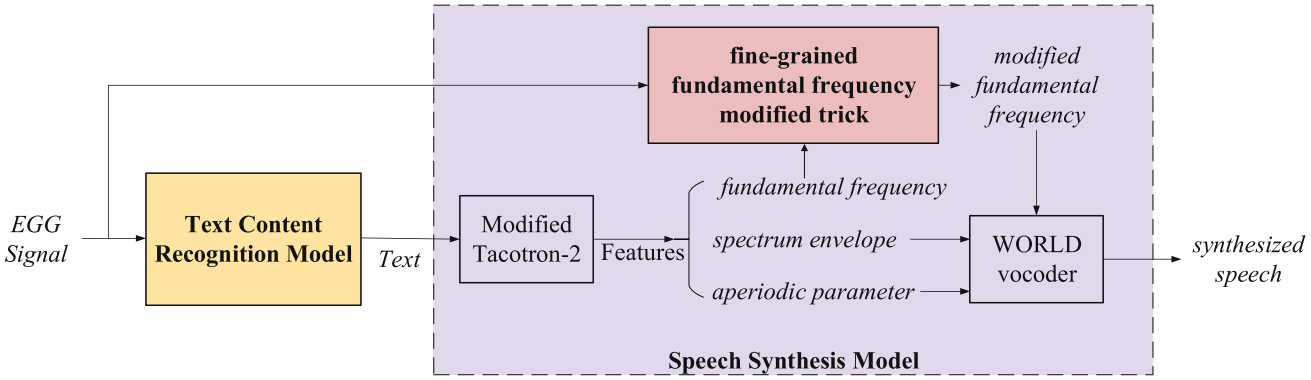


Fig. 2 The flow chart of our framework

vector sequence, and fed in a 3-layer bidirectional LSTM (Long Short-Term Memory [35]) to convert the sequence into one of 20-class of sentences with different contents. This research archived to recognize the text from some specific classes of sentences through the EGG signal and provided support on our task of speech synthesis with EGG signals.

Based on our previous study, which has proved that EGG signal can be classified into the text in the scenario of limited contents, we now propose a framework to synthesize personalized speeches utilizing EGG signals. Compared to speech signals, EGG signals have the following two superiority under our application scenarios. (1) clean EGG signals can be collected in high noise circumstances, where clean speech signals can not. (2) To enable patients who retain vocal cord vibration but lose the ability to produce voice to speak again. It must be highlighted that to our knowledge, this is the first attempt to utilize EGG signals into Tacotron-2.

This paper is organized as follows. In Section 2, our methods and materials are introduced. In Section 3, we discuss the results of our models and the comparison experiments we have conducted. Section 4 makes conclusions and points out the expected future works.

2 Methods and materials

2.1 Methods

This paper proposes a framework of applying only the EGG signal for speech synthesis in the limited categories of contents scenario. Our framework consists of a text content recognition model with the EGG signal and a speech synthesis model with the text and EGG signal. Figure 2 is the overall structure of our entire framework. The function of the text content recognition model is to get text recognition results of the EGG signal input, which is essential for speech synthesis. Then, the speech synthesis model produces speech with the EGG signal and its text recognition result. Based on Modified Tacotron-2, three features (fundamental frequency, spectrum envelope, and aperiodic parameter) are extracted from the text recognition result. To utilize the information contained in EGG signals and synthesize personalized speeches, we choose WORLD as the vocoder and put forward a fine-grained fundamental frequency modification method to obtain the modified fundamental frequency. Finally, speech is synthesized by the WORLD vocoder according to three features.

2.1.1 Text content recognition model with the EGG signal

In our previous study, we have set up a text content recognition model for getting the text recognition result from the EGG signal, proposed in the article [34]. So we just briefly introduce our method here.

Figure 3 is the structure of our text content recognition model, which consists of an EGG feature extraction module to obtain the feature vector sequence from the whole EGG signal and a recognition network to get the recognition result of the feature vector sequence.

As Fig. 3 depicts, the EGG feature extraction module consists of three parts, voiced segments extraction part, feature extraction part, and smoothing part. When setting about EGG signals, we firstly extract voiced frames from the EGG signal to avoid the unvoiced segments' influence. Considering that the change of the pitch varies greatly from each other for two long sentences, we choose three parameters as features: the fundamental frequency (F_0), the relative first-order difference of F_0 ($diffF_0$) and the log short-term energy ($logE$). Between them, F_0 is a commonly-used parameter to characterize vocal cord vibration. The extraction of F_0 is based on the periodical change of the amplitude of EGG signals, which is estimated by the auto-correlation method as follows:

$$F_0 = \frac{f_s}{\frac{f_s}{f_{max}} \le k \le \frac{f_s}{f_{min}} \arg\max \sum_{m=0}^{N-1-k} x_{EGG}(m) x_{EGG}(m+k)} \quad (1)$$

where f_s is the sampling rate. f_{max} and f_{min} are the maximum and minimum of the F_0, respectively.

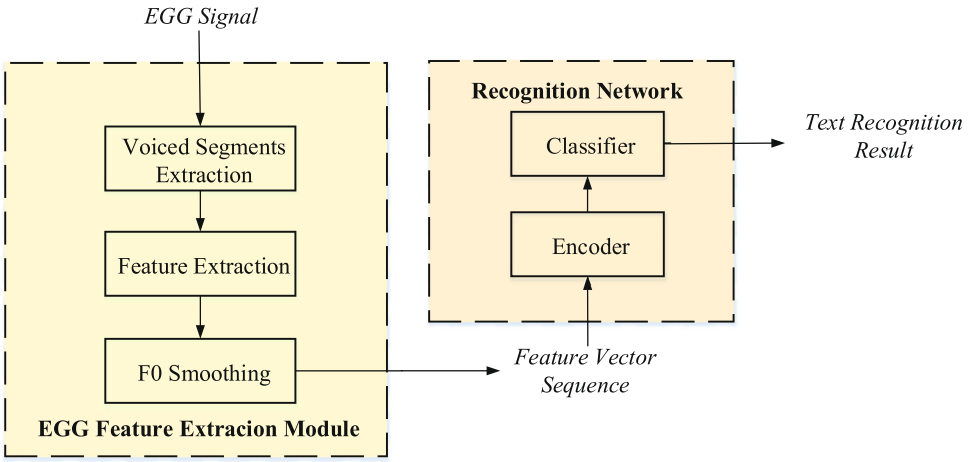


Fig. 3 The structure of our text content recognition model

$diffF_0$ indicates the change of F_0 over time, which is naturally calculated as equation (2). The log short-term energy ($logE$) is included to characterize the stress distribute of the EGG signal.

$$diffF_0(i) = \frac{F_0(i+1) - F_0(i)}{F_0(i)} \tag{2}$$

where $F_0(i)$ and $diffF_0(i)$ are respectively F_0 and the relative first-order difference of F_0 at the frame i.

As the method of F_0 extraction will cause erroneous values [36], F_0 smoothing is required. We adopt the smoothing method with bidirectional searching proposed by Jun et al [37], which combined interpolation and mean filtering according to a normalized F_0 in every segment. Compared with the traditional median filter, bidirectional searching achieves better performance on F_0 smoothing according to our experiment.

Through feature extraction module, a feature vector sequence is prepared, which contains 504 frames and 3 features at every time step. Next, we feed the feature vector sequence in our recognition network. Figure 4 illustrates the structure of our recognition network.

As Fig. 4 shows, our recognition network consists of an encoder and a classifier. The encoder extracts contextual information from the feature vector sequence and generates a contextual vector. Then the classifier converts the contextual vector into an index that can search the sentence from 20-class of contents dictionary. For reasons of the superior performance of LSTM in sequence processing tasks, we select Three-layer Bi-LSTM [38] as our encoder, which has proved effective according to our comparative experiment. Through the encoder, the forward and backward output are concatenated to obtain the last encoded vector. Then the classifier generates a probability vector by feeding the contextual vector into the fully connected layer and the softmax operator.

2.1.2 Speech synthesis model with the EGG signal and the text

After the text content recognition model, we have obtained corresponding texts from EGG signals. The next step is to synthesize speeches utilizing the texts and EGG signals. Our speech synthesis model consists of the same EGG feature extraction module to extract F_0 and a Chinese speech synthesis model to synthesize speeches. Besides, to contain the personal characteristic into synthesized speech utilizing EGG signals, we propose a fine-grained fundamental frequency modification method. Figure 5 is the overall flow chart of our speech synthesis model.

The principle and details of the EGG feature extraction module have been discussed in 2.1.1, we reuse it here to obtain F_0 to work in the fine-grained fundamental frequency modification method. As for the Chinese speech synthesis module, it consists of three parts, text frontend, acoustic model and vocoder. Next we introduce these parts as well as our proposed F_0 modification method as follows.

Text Frontend The function of the text frontend is to classify the input characters and encoder into a limited number of classes to reduce the difficulty in training the acoustic model. When focusing on the Mandarin TTS task, it means to convert a sequence of Chinese characters into smaller text modeling units, like a sequence of pinyin or phones.

To explore which text modeling unit is optimal between pinyin and phone, we conduct comparative experiments based on pinyin and phone, respectively. Specifically, pypinyin library is used to convert Chinese characters into pinyin and a pinyin-to-phones dictionary is set up through the mapping between pinyin and phone sequence in the Chinese speech synthesis dataset. Compared with 4103 classes of Chinese characters, we squeeze the number into only 1609 of pinyin and 263 of phones after the text frontend.

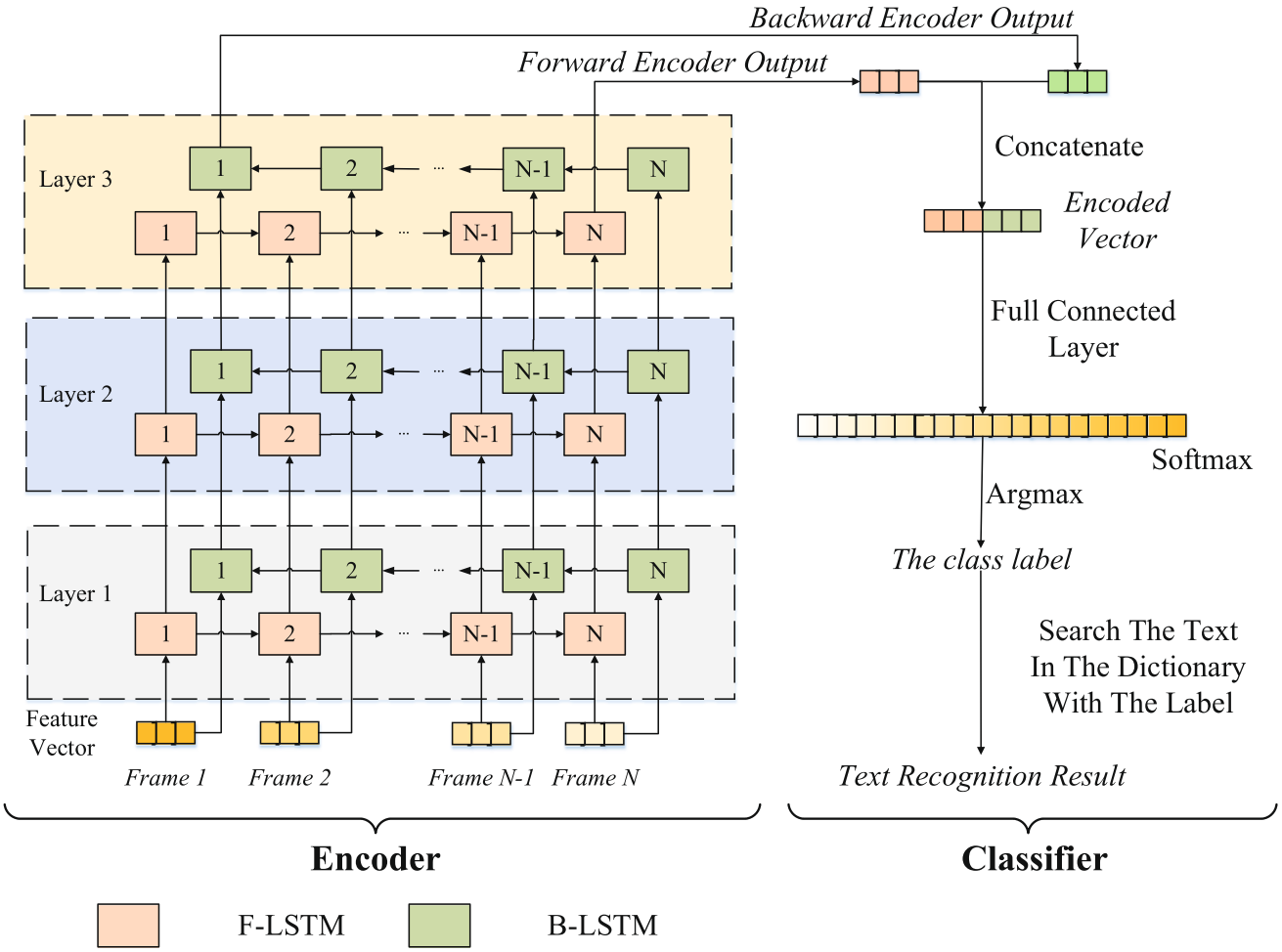
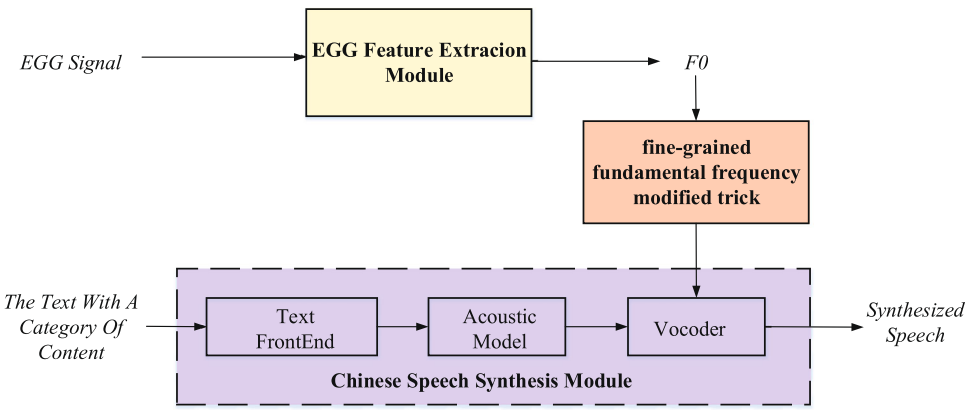


Fig. 4 The structure of the recognition network

Fig. 5 The structure of the speech synthesis model



Acoustic Model The function of the acoustic model is to establish the mapping between text modeling units. Our acoustic model is based on Tacotron-2 [39] and modified with depthwise separable convolution to decrease the model size. Our model replaces all the convolutions in the Tacotron-2 model with depthwise separable convolutions.

Tacotron-2 is an end-to-end TTS model. The core of Tacotron-2 is the acoustic model, shown in Fig. 6, which comprises an encoder, a decoder and a postprocessing module (PostNet). In the encoder, the text modeling unit sequence is converted into a dense vector through character embedding. This dense vector is input into a 3-layer 1-dimensional convolutional layer to simulate the language model and then feed in a 2-layer Bi-LSTM to obtain an encoded vector.

In terms of the decoder, an attention weight vector is calculated based on the encoder output. Two types of attention,

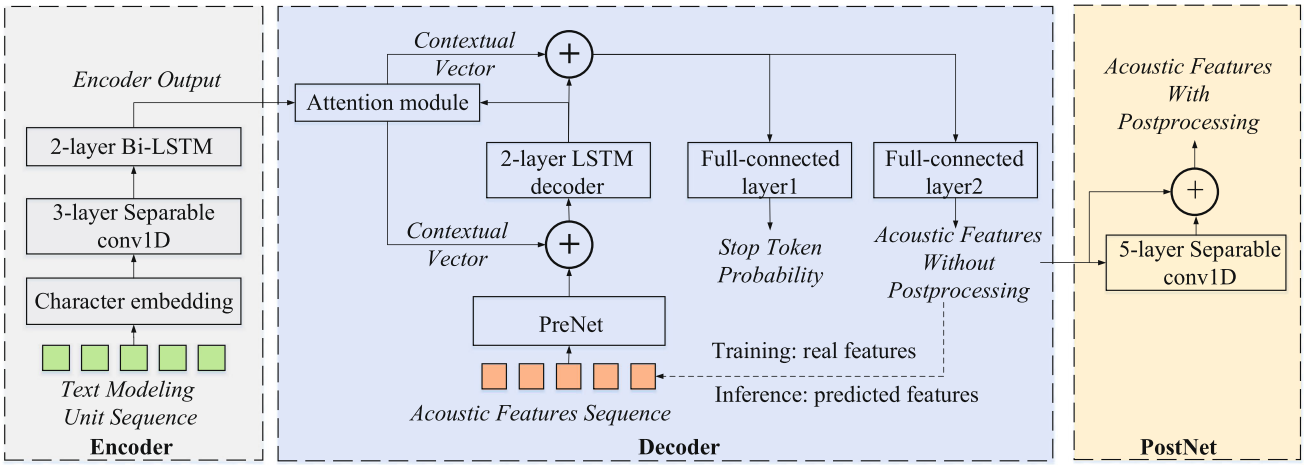


Fig. 6 The structure of the acoustic model

content attention and position attention, are applied here. Content attention focuses on the correlation between the decoder's hidden vector at a certain time step and the encoder's at each time step, while position attention figures out the correlation between the decoder's attention weight vector at a certain time step and the encoder's hidden vector at each time step. Attention scores are calculated by fully connected layers, defined in the following equation:

$$e_{ij} = score(s_i, ca_{i-1}, h_j) = v_a^T tanh(Ws_i + Vh_j + Uf_i + b) \tag{3}$$

where W, V, U and b are the parameters that fully connected layers learn. s_i is the hidden state of the decoder at the current time step i. h_j is the hidden state of the encoder at time step j. ca_i indicates the accumulation of attention weight vector a_j calculated by equation (4) and f_i is convoluted by ca_i, shown in equation (5). The attention weight vector a_j in equation 4 is a composition of the attention weight coefficient, $a_j = [a_{j1}, a_{j2}, \cdots, a_{jS}]$. By calculating ca_i, the attention weight network can acquire the attention information which has been learned, so that the model could avoid repeating the unexpected speech.

$$ca_i = \sum_{j=1}^{i-1} a_j \tag{4}$$

$$f_i = F * ca_i \tag{5}$$

After calculating the attention score (e_{ij}), the attention weight coefficient a_{ij} can be figured out by softmax, as equation (6). Finally, the output of attention module, the context vector c_i can be generated by accumulating the product of a_{ij} and the hidden state of the encoder h_j, shown in equation (7).

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{s} exp(e_{ik})} \tag{6}$$

$$c_i = \sum_{j=1}^{S} a_{ij} h_j \tag{7}$$

The input of preprocessing network (PreNet) are acoustic features, and the teacher forcing criterion works in the training phase. The output of PreNet and the context vector calculated from the last decoding time step are input into the 2-layer LSTM decoder. Meanwhile, the context vector is generated by the decoder's output combined with the attention weight of the last decoding time step. This process forms a cycle. The final output is predicted by the linear projection of the concatenated vector of decoder output and the context vector. There are two forms of output, one is the acoustic feature, the other is the stop token probability, of which the latter is a binary recognition task, determining whether the decoding process ends. Besides, the acoustic features of p frame ($p > 1$) are predicted at each time step to speed up calculation and reduce memory consumption.

Considering PostNet, 5-layer convolutional layers and residual connections are combined to refine the predicted acoustic features.

We design the loss of our acoustic model to include the following four parts: (a) The mean square error between target acoustic features $y_{target,i}$ and predicted ones without post-processing $y_{prev,i}$. (b) The mean square error between target acoustic features and predicted features with post-processing $y_{post,i}$. (c) The cross-entropy loss between the one-hot vector of the target stop token S_{target} and the probability vector of the predicted stop token $S_{prediction}$. (d) The L_2 regularization loss ($\lambda = 10^{-6}$). The loss function is defined as equation (8):

$$\begin{aligned} loss = \; & MSE(y_{target,i}, y_{prev,i}) + MSE(y_{target,i}, y_{post,i}) \\ & + CE(S_{target}, S_{prediction}) + \lambda \sum_{j=1}^{p} w_j^2 \end{aligned} \tag{8}$$

Additionally, to reduce the size of our model, we modify the baseline Tacotron-2 introduced above and replace the regular convolution with the depthwise separable convolution in our model. The depthwise separable convolution originated from Xception [40] and MobileNet [41]. This

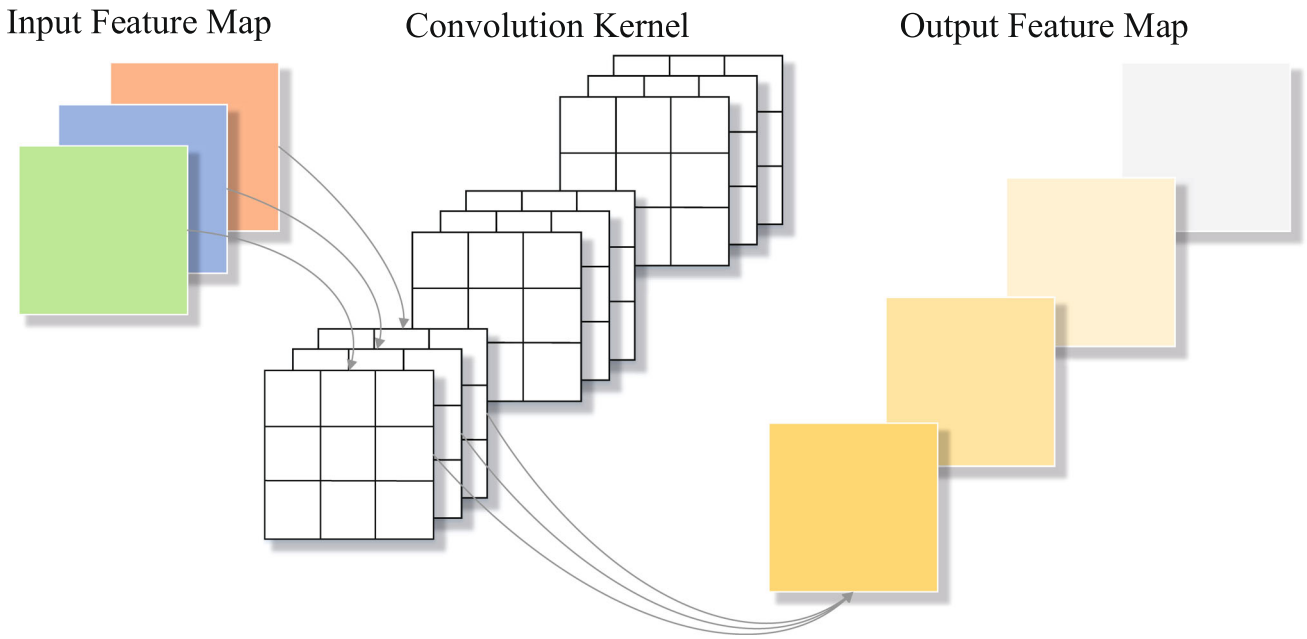


Fig. 7 The principle of the regular convolution

method is widely used to substitute the regular convolution to reduce the number of parameters as well as the model size, sometimes enables the model to converge faster [42].

Figure 7 illustrates the principle of the regular convolution. Concretely, for the 2-dimensional regular convolution, each convolution kernel squeezes all channels of the input feature map into a one-channel output feature map. Then all one-channel output feature maps are concatenated into an output with a given number of channels.

The 2-dimensional depthwise separable convolution comprises the depthwise convolution and pointwise convolution. On the depthwise convolution phase, each convolution kernel goes through a channel of the input feature map to generate C_M (channel multiplier, $C_M \geq 1$) channels of output. On the pointwise convolution phase, all the channels of the depthwise convolution output are feed in a 1×1 pointwise convolution to generate one-channel output. Finally, all one-channel output feature maps are combined to produce an output with a given number of channels. Figure 8 depicts the principle of the depthwise separable convolution.

According to Fig. 8, the number of parameters of the depthwise separable convolution is calculated as the following equation.

$$N_{sc} = k_h \times k_w \times C_{in} \times C_M + C_{in} \times C_{out} \times C_M \tag{9}$$

The ratio of the number of parameters between these two methods is shown in equation (10), which proves that the depthwise separable convolution has much fewer parameters than the regular.

$$\frac{N_{sc}}{N_{rc}} = \frac{C_M}{C_{out}} + \frac{C_M}{k_h \times k_w} \tag{10}$$

Considering that all the convolutions in Tacotron-2 are 1-dimensional. To utilize depthwise separable convolution, we expand all the feature maps with a height dimension. Concretely, we change the shape from 2-dimensional normal structure $T \times W : T \times d$ to 3-dimensional normal structure $H \times W \times C : 1 \times T \times d$. Afterward, we set k_h to 1, k_w to the number of time steps, and C_{out} to the dimension of features at each time step. Finally, we remove the height dimension to obtain the 2-dimensional normal structure ($H \times W \times C : 1 \times T \times d \rightarrow T \times W : T \times d$).

Vocoder The function of the vocoder is to generate the speech according to acoustic features. Wavenet, the default vocoder in Tacotron-2, is utilized in the process of synthesizing the original speech. But to synthesize the personalized speech with the aid of the EGG signal, we require a vocoder that utilizes the F_0, so we choose WORLD in our framework. WORLD vocoder generates the speech according to the F_0, spectrum envelope, and aperiodic parameter. Among them, F_0 is used as the periodic excitation and the aperiodic parameter as the aperiodic excitation to constitute the mixed excitation signal $e(n)$. The spectrum envelope simulates the resonance part of the vocal tract through the minimum phase response $h(n)$. The synthesized speech signal is figured out by the convolution of these two signals. Figure 9 is the principle of the WORLD vocoder.

Fine-grained fundamental frequency modification method To utilize the speaker's characteristics contained in the EGG signal and synthesize personalized speech, a fine-grained fundamental frequency modification method is proposed. We design both the paralleled path, which means the EGG signal corresponds to the text, and the unparalleled path, which

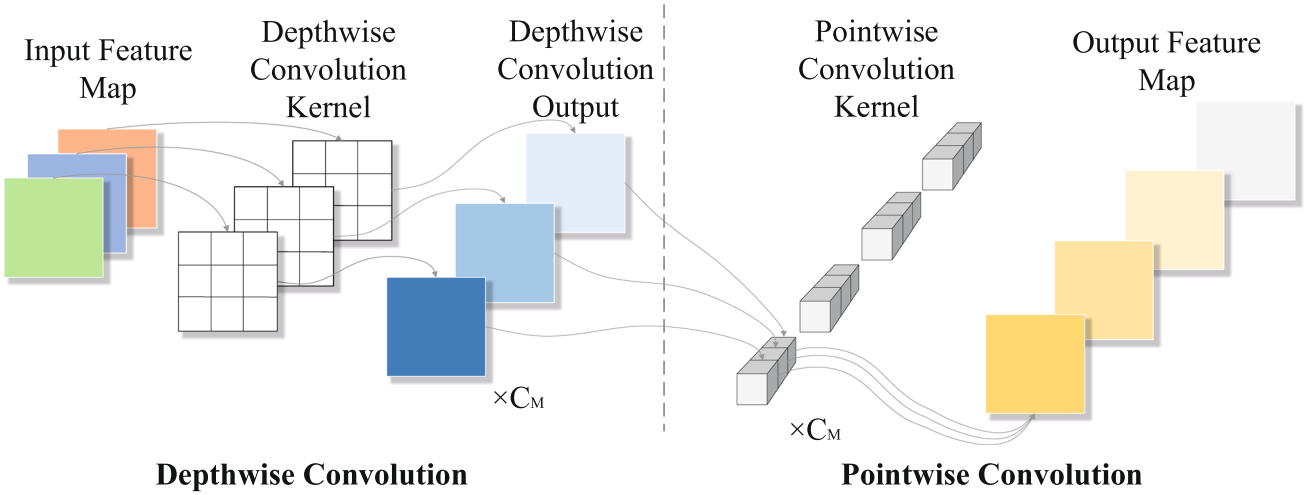


Fig. 8 The principle of the depthwise separable convolution

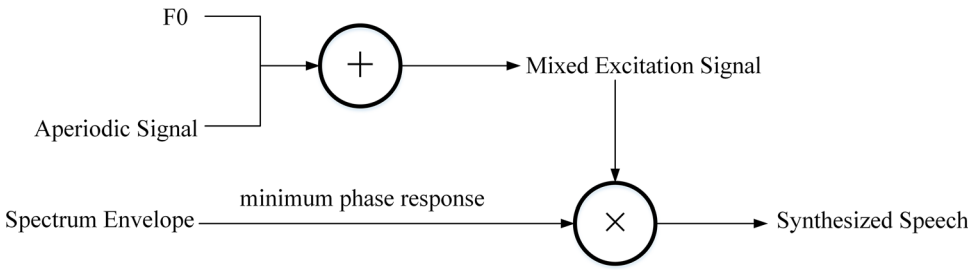


Fig. 9 The principle of WORLD vocoder

means the EGG signal indicates a different context from the text. Figure 10 depicts the principle of our fine-grained fundamental frequency modification method.

For the unparalleled path, as the waveform between F_{0EGG} and $F_{0Speech,old}$ is different, we apply the coarse-grained fundamental frequency modification method to synthesize personalized speech. As the average value of the F_0 describes the pitch of the speaker, we calculate the ratio of the average value of the F_0 of the EGG signal and the original F_0 feature as equation (11) shows and set this ratio as the modified scale to adjust the F_0 and spectrum envelope of the acoustic features predicted by Tacotron-2 point by point. The adjustment equations are defined as follows.

$$R = \frac{F_{0EGG}}{F_{0feature}} \tag{11}$$

$$F_{0personalized}(i) = R \times F_{0feature}(i) \tag{12}$$

$$Spec_{personalized}(i,k) = Spec_{feature}(i, [\frac{k}{R}]) \tag{13}$$

where R is the coarse-grained adjustment ratio. $F_{0EGG}, F_{0feature}$ is the average value of the F_0 of the EGG signal and the original synthesized speech, respectively. $F_{0feature}(i), F_{0personalized}(i)$ is the $i-th$ frame of the F_0 of the original synthesized speech and the newly synthesized speech with adjustments, respectively. $Spec_{feature}(i,k)$, $Spec_{personalized}(i,k)$ is the $i-th$ frame and the $k-th$ frequency sampling point of the spectrum envelope of the original synthesized speech and the newly synthesized speech with adjustments, respectively.

By adjusting the overall range of F_0 and re-sampling the frequency axis to adjust the spectrum envelope, we can synthesize personalized speech contained speaker's characteristics by the WORLD vocoder.

For the paralleled path, as F_{0EGG} and $F_{0feature}$ correspond to the same context, their waveforms are similar to each other and can be aligned. So we put forward the fine-grained fundamental frequency modification method to more detailed adjust $F_{0feature}$ and synthesize personalized speech. Due to the obvious difference in sampling rate and duration between the EGG signal and the original synthesized speech. F_{0EGG} often mismatches $F_{0feature}$. So firstly, we apply dynamic time wrapping to F_{0EGG} to obtain $F_{0EGG,aligned}$ which shares the same length and zero segments as $F_{0feature}$. Then, we apply the coarse-grained fundamental frequency modification method to gain $F_{0coarse-grained}$. For further adjusting $F_{0coarse-grained}$ to imitate the changes of F_{0EGG} over time, we conduct F_0 fine-grained modification, which generates a specific ratio $r(i)$ to indicate the relationship between this time step and the overall range, and modified $F_{0feature}$ at every time step, defined as follows.

$$r(i) = \frac{F_{0EGG}(i)}{F_{0EGG}} \tag{14}$$

$$F_{0personalized}(i) = r(i) \times F_{0coarse-grained}(i) \tag{15}$$

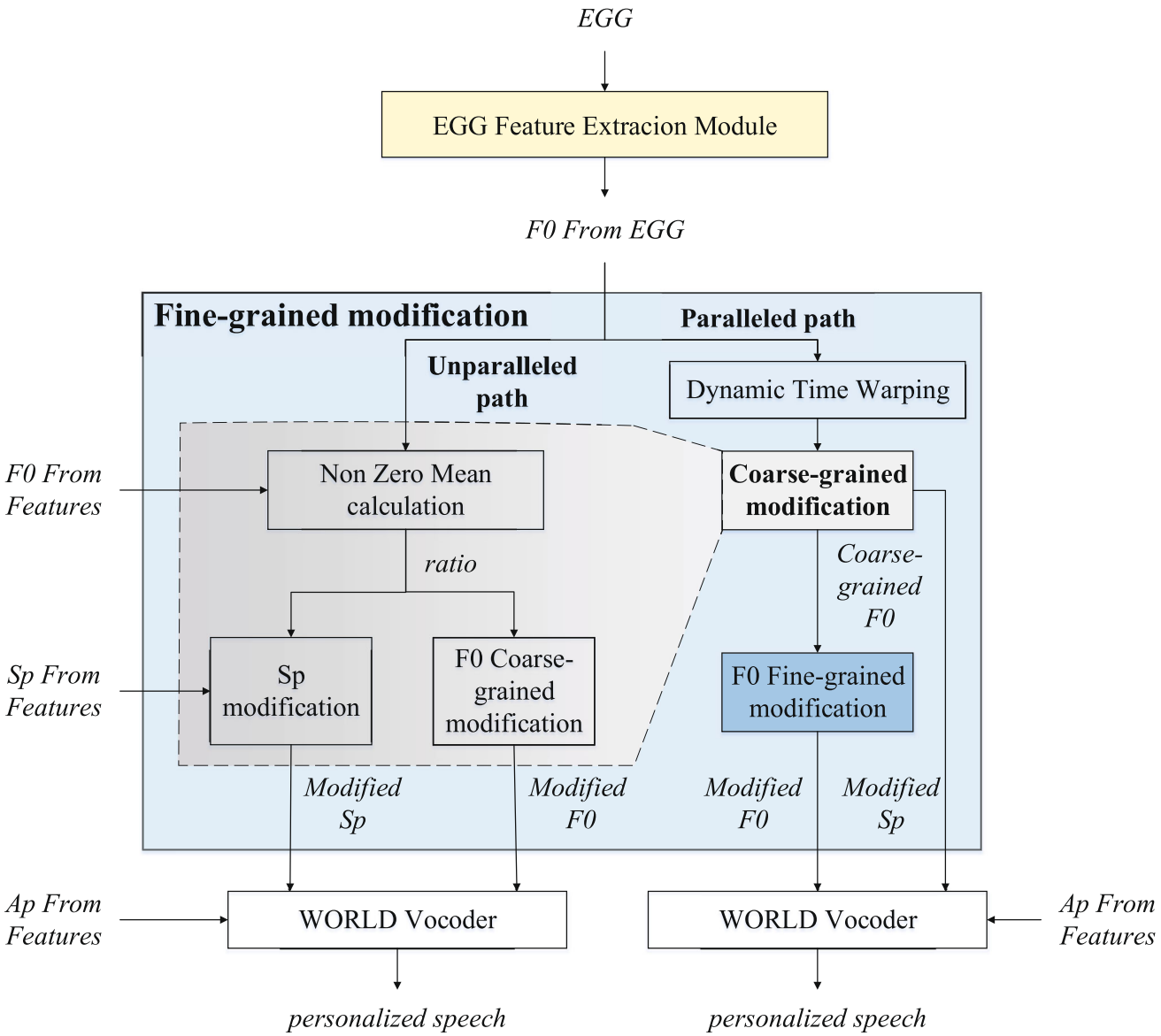


Fig. 10 The principle of fine-grained fundamental frequency modification method

where $r(i)$ is the fine-grained adjustment ratio. F_{0EGG} is the average value of the F_0 of the EGG signal. $F_{0_{coarse-grained}}(i)$, $F_{0_{personalized}}(i)$ is the $i-th$ frame of $F_{0_{coarse-grained}}(i)$ and the F_0 of the newly synthesized speech with fine-grained adjustments, respectively.

By applying fine-grained modification, we rely much deeply on F_{0EGG}. We obtain personalized F_0 which not only shares the same overall range as F_{0EGG} but also imitates the range changes over time, which promises to contain more personalized characteristics such as tone and stress.

2.2 Materials

The dataset for our text content recognition model is the Chinese Dual-mode Emotional Speech Database (CDESD [33]). This dataset is built by the pattern recognition and human intelligence laboratory affiliated with the Department of Electronics and information engineering at Beihang University and collected from 20 speakers aged 21 to 23 (13 men, 7 women). The dataset contains 11366 speeches and corresponding EGG samples, and there are 20 classes of sentences with different contents in this dataset, which is the output of the classifier. In the experiment, 0.8 of the total dataset are chosen as the training set and the others as the validation test.

The dataset for the Chinese speech synthesis is Biaobei Chinese female voice dataset[1], which is widely used for Mandarin TTS task. The dataset is recorded by a 20-year-old woman, whose voice is active and intelligent. The total duration is about 12 hours and the sampling rate is 48 kHz.

[1] https://www.data-baker.com/data/index/source

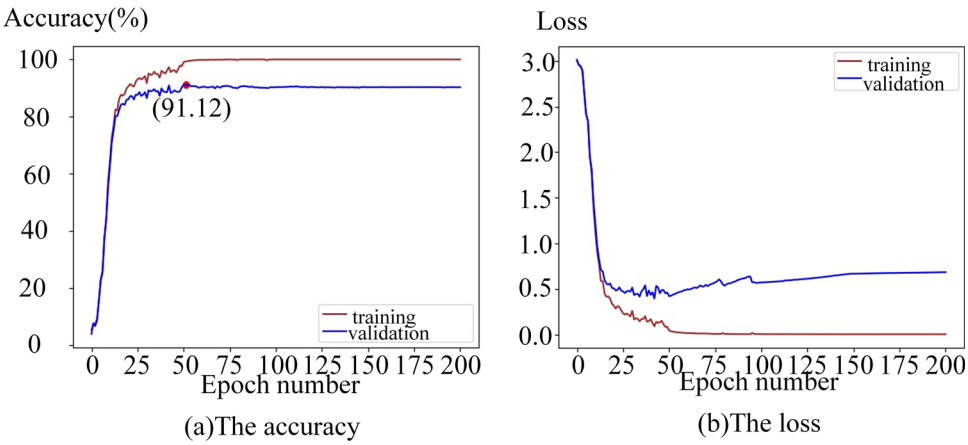


Fig. 11 Loss and accuracy of recognition model

3 Experiments, results and discussions

3.1 Text content classification model

Figure 11(a) and (b) show the loss and accuracy of every epoch fo our text content classification model. The best result of the validation set occurs at epoch 52, whose accuracy reaches 91.12 %. This result is based on the following conditions: (a) choosing the 3-layer Bi-LSTM as the encoder, (b) including all of the three features, the F_0, the relative first order difference of F_0 and the log short-term energy *logE*, (c) choosing bidirectional smoothing as the smoothing method. The promising recognition accuracy provides strong support for speech synthesis based on the classified text.

To figure out the best conditions of our text content classification model, we design three series of comparative experiments. The first comparative experiment explores which encoder is the most effective in extracting contextual information. We choose commonly-used encoders as baseline, including CNN, Bi-GRU and LSTM, to highlight the superior of Bi-LSTM. Additionally, to figure out how many layers of Bi-LSTM perform best, we also conduct experiments under different numbers of layers. The result is listed in Table 1 and suggests that the 3-layer Bi-LSTM is the best encoder compared to the others. This result proves the effectiveness of our encoder and provides the guidance to choose an encoder with appropriate numbers of parameters.

The second comparative experiment explores whether every feature we select contributes to improve the accuracy of the recognition network and which combination is the best. We try out different units of features with the orginal EGG signal as baseline. In Table 2, the result indicates that using all the three features is more effective than other combinations. That is, all of these three features work for improving recognition. Among these features, F_0 proves to have an essential influence on the result, which accords with our expectation that F_0 directly reflects the characteristic of the vocal cord vibration of speakers.

The third comparative experiment explores which smoothing method is optimal. We set F_0 without any smoothing as baseline and compare bidirectional smoothing method with traditional median filter. The result shown in Table 3 suggests that bidirectional smoothing achieves a better result than other methods.

The experiment with the best result in this section shows the satisfying performance of our text content recognition model with the EGG signal, which lands a strong basis for the research of speech synthesis with the EGG signal and the text. Besides, concluded from the comparative experiments, the combination of three conditions, encoder, feature and smoothing method selection, contributes to the best results of our model.

3.2 Speech synthesis model

Figure 12(a), (b), (c) and (d) shows the total loss and the former three types of losses defined in equation (8) of every iteration of the acoustic model.

In our experiment, we set the batch size as 32, the total training iterations as 2M, the initial learning rate as 1e-3 and the final as 1e-5 and exponentially decay every 4000 iterations. The optimizer is Adam [45]. After about 200k iteration, the loss curve converges at a very low value, which proves a satisfying performance of the model. Compared with Fig. 11 (b) and (c), the loss of the acoustic features with postprocessing is much less than that without postprocessing, which proves the effectiveness of the postprocessing module [46]. Figure 11 (d) proves that the model has learned how many time steps should stop generating the predicted acoustic features.

To evaluate the sound quality of synthesized speeches, objective and subject test are conducted. We choose Mel cepstral distortion (MCD) [47] in objective test, for MCD is regarded

Table 1 The comparative experiments among different encoders

Network	Acc-Train(%)	Acc-Val(%)
CNN [43]	72.20	68.78
Bi-GRU [44]	92.64	86.08
LSTM (1 layer)	90.80	82.94
Bi-LSTM (1 layer)	90.47	85.05
Bi-LSTM (2 layer)	98.98	90.19
Bi-LSTM (3 layer)	**99.25**	**91.12**
Bi-LSTM (>3 layer)	Not converged	Not converged

Table 2 The Acc results among different feature selection strategies

Network	Acc-Train (%)	Acc-Val (%)
Original EGG signal	70.45	61.25
F_0	97.04	86.06
$diffF_0$	89.45	80.56
$logE$	86.41	74.80
$F_0 + diffF_0$	96.76	86.47
$F_0 + logE$	97.75	88.52
$diffF_0 + logE$	85.45	78.54
$F_0 + diffF_0 + logE$	**99.25**	**91.12**

Table 3 The Acc results among different F_0 smoothing methods

Network	Acc-Train (%)	Acc-Val (%)
Without any process	97.33	89.49
5-Median filter	91.00	82.19
7-Median filter	97.38	85.79
9-Median filter	97.85	86.80
Bidirectional Smoothing	**99.25**	**91.12**

as a target to indicate spectral performance. When dealing with subject test, the mean opinion score (MOS) is figured out. We choose SAG-Tacotron [30] to a representative state-of-the-art performance, for it improves the naturalness of speeches without complex front-end, corresponding to our target.

3.2.1 objective test

In the objective test, we set the original Tacotron-2 as the baseline, and compare our model with SAG-Tacotron. We test our model under the condition with and without our fine-grained fundamental frequency modification method to explore the influence of EGG in speech synthesis. Dynamic time warping (DTW) is applied to align the frames of the predicted Mel spectrums with the ground truth.

Table 4 shows the results of different methods judged by MCD. As lower MCD indicates better spectral performance, it can be figured out that our model outperforms the Tacotron-2 with a decrease of 0.14 and our fine-grained fundamental frequency modification method works in improving the quality of the speech. The performance of our model, Tacotron-2+ DSC with modification, is comparable with the state-of-the-art performance.

3.2.2 Subjective test

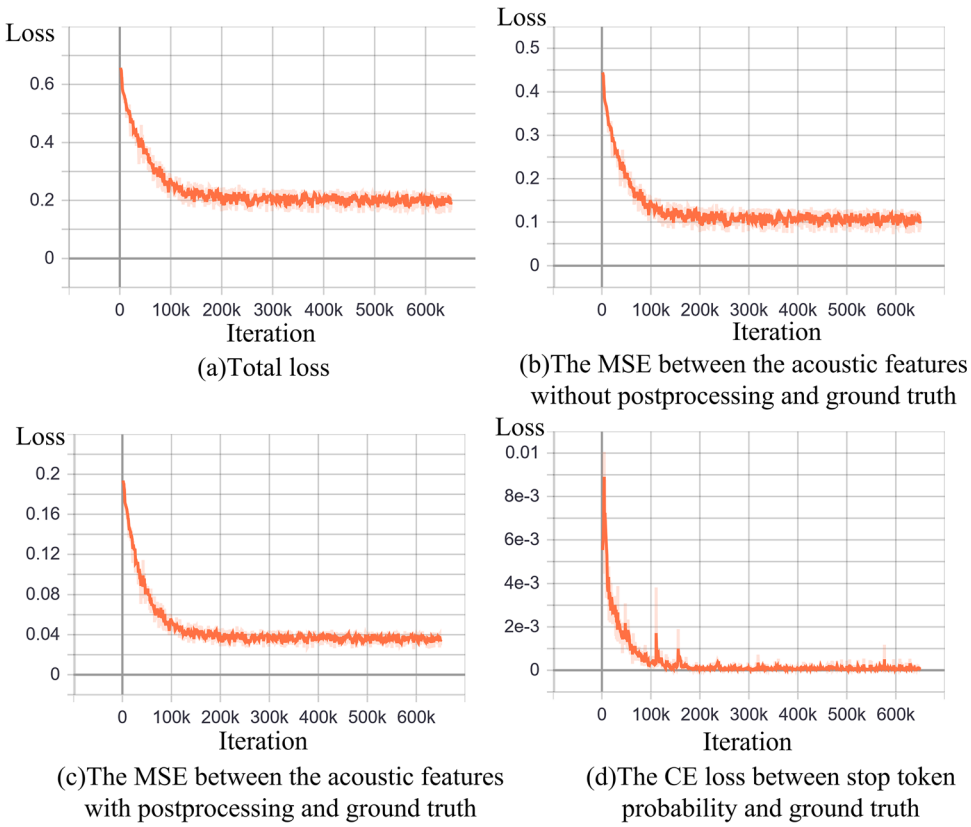
For the mean opinion score (MOS) measurement, we set up 5 series of evaluation sets in which 5 different sentences are included and invite 20 listeners, 10 men and 10 women, aged 18 to 40, to randomly choose and rate the quality on a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. Table 5 shows the performance of different methods. As participants' feedback, our model improves the performance of the original Tacotron-2 with a gain of 0.42 and achieves a comparable score with SAG-Tacotron. When associating with the fine-grained fundamental frequency modification method, we can get a higher score of 3.94, which proves our modification is effective. Besides, the low variance indicates that the robustness of our model.

3.2.3 Comparative experiments

We conduct two series of comparative experiments to figure out the best acoustic model, focusing on text modeling unit and the selection of depthwise separable convolution parameters. For Mandarin TTS task, the choice of text modeling unit is essential. So we explores both pinyin and phone to figure out which text modeling unit is optimal. Figure 13 shows the alignment between the encoder and decoder. Comparing Fig. 13 (a) with (b), the alignment curve of phone modeling is nearly a straight line, while that of pinyin modeling is messy. This phenomenon proves that for this dataset, choosing phone as the text modeling unit is much better than pinyin because the number of the classes of phone is much smaller than pinyin, let alone Chinese characters.

Table 6 shows the comparative results of MOS under two text modeling units. The result proves the better performance of the phone modeling once again. However, the quality of synthesized speech is worse than the ground truth. It may be because that phone modeling regards the transition of two consecutive phones from two different Chinese characters as the same as that in a Chinese character, which will cause the synthesized speech to be not fluent and natural enough.

The other comparative experiment explores whether the modification on the acoustic model works. Selecting the original Tacotron and Tacotron-2 as baselines, we seek the best C_M on our Tacotron-2 revised by depthwise separable convolution. Table 7 shows the comparative results under two



(a)Total loss

(b)The MSE between the acoustic features without postprocessing and ground truth

(c)The MSE between the acoustic features with postprocessing and ground truth

(d)The CE loss between stop token probability and ground truth

Fig. 12 Losses of the acoustic models

Table 4 The MCD evaluation of different acoustic models

Methods	MCD(dB)
Baseline	6.017
Tacotron-2+DSC[1]	5.877
Tacotron-2+DSC with modification	**5.781**
SAG-Tacotron [30]	5.775

DSC is short for depthwise separable convolution

Table 5 Mean option scores(MOS) with 95% confidence intervals

Method	MOS
Tacotron [25]	3.40±0.14
Tacotron-2 [39]	3.45±0.13
Tacotron-2+DSC	**3.87±0.15**
Tacotron-2+DSC with modification	**3.94±0.14**
SAG-Tacotron	3.87±0.13
Ground Truth	4.50

aspects: (a)The MOS of the synthesized speech. (b)The model size of the acoustic model. The conclusion is that Tacotron-2 is much better than Tacotron. Besides, the quality of the speech synthesized by the modified Tacotron, which is revised by depthwise separable convolution structure achieves better-synthesized speech than that by the original Tacotron. Meanwhile, the model size of the revised model is much smaller than the original one.

Both aiming to achieve natural prosody on an end-to-end speech synthesis system for Mandarin, the state-of-the-art performance is realized by SAG-Tacotron [30]. Shown in Tables 4 and 5, our model achieves a comparable performance with SAG-Tacotron in the objective and subjective test. And as Table 7 shows, the trade-off between the quality of the synthesized speech and the model size gets the balance.

3.3 Fine-grained fundamental frequency modified method

Figure 14 shows the F_0 extracted from the EGG signal. For unparalleled path, Fig. 15(a) shows the F_0 extracted from the features of Tacotron-2. As the F_{0EGG} signal contains personalized features of the speech, it can be utilized to adjust the original F_0 and synthesize personalized speech. By equation (11), the adjustment ratio R is figured out. Figure 15(c) is the spectrum of the original synthesized speech. Figure 15(b) and (d) is the F_0 and the spectrum of the adjusted synthesized speech, respectively. Figure 15(b) shows the average F_0 is more similar to that of the EGG signal, which means the pitch of newly

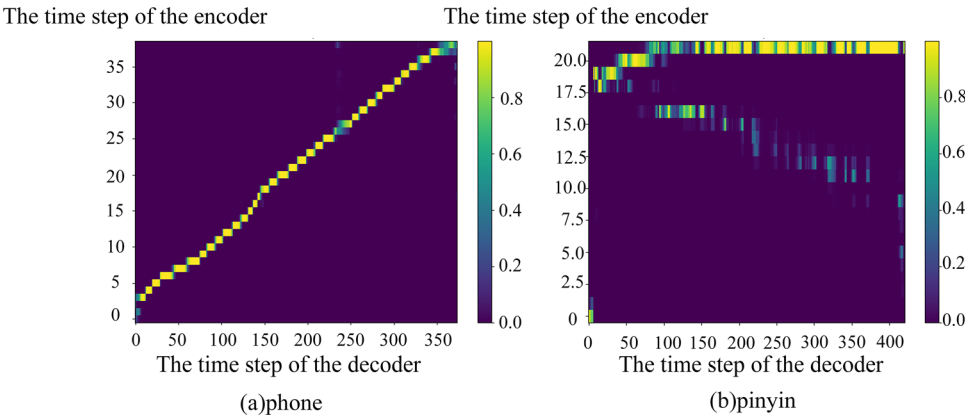


Fig. 13 The alignment between the encoder and decoder with different text modeling units

Table 6 The MOS under different text modeling units

Text Modeling Unit	MOS
Pinyin	1.05
Phone	**3.87**
Ground Truth	4.50

Table 7 The MOS and model size of different acoustic models

Acoustic Model	MOS	Model Size
Tacotron [25]	3.40	278.3MB
Tacotron-2 [39]	3.45	320.3MB
Tacotron-2+DSC($C_M = 1$)	3.34	232.8MB
Tacotron-2+DSC ($C_M = 2$)	**3.87**	255.0MB
Tacotron-2+DSC ($C_M = 3$)	2.98	277.1MB
Tacotron-2+DSC ($C_M = 4$)	3.67	299.3MB
Ground Truth	4.50	

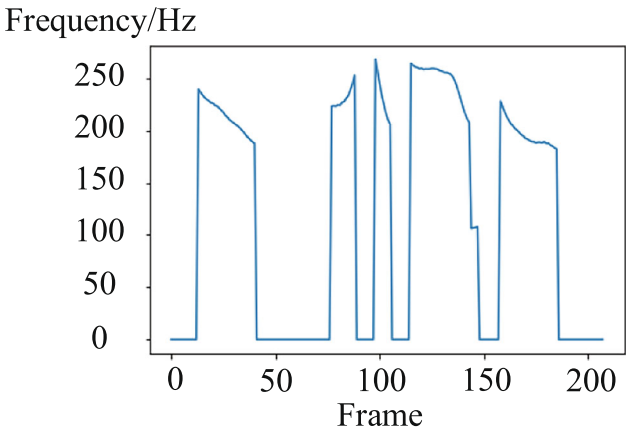


Fig. 14 The F_0 extracted from EGG

synthesized speech is similar to the speaker. Figure 15(d) shows the frequency axis resampling, which means the reasonable change of the fundamental frequency and the harmonic frequency.

For paralleled path, the F_{0EGG} is first aligned to the F_0 from features by conducting the dynamic time wrapping method. Then fine-grained fundamental frequency modification method is conducted to adjust F_0 in more detailed levels. The comparison of F_0 between the original and personalized speech is shown as Fig. 16. It indicates that not only the average F_0 has been modified to fit the speaker's tone, but also the trend of F_0 has been adjusted according to the F_{0EGG}, which includes the stress information into the final personalized speech.

To figure out the voice quality of personalized speeches, a series of subjective evaluation is conducted. Table 8 shows that for unparalleled path, coarse-grained modified speech gains the mean opinion score (MOS) of 3.94, which proves that EGG signal contributes to improving the naturalness of F_0 and synthesizing personalized speech. Compared with the state-of-the-art performance of Mandarin TTS named SAG-Tacotron [30], our method achieves better result of MOS. For paralleled path, the MOS of fine-grained modified speech is slightly lower than the original speech. It may because there still be something mess when conducting the alignment. However, it must be pointed out that the fine-grained modified speech includes the stress of the speaker, as listeners feed back. So fine-grained fundamental frequency modification method still proves to add more detailed information into the final speech and its good performance is promising when solving the alignment problem.

The results of our experiment indicate that utilizing EGG signals enables the personalized synthesized speech to be more consistent with the speaker's characteristics. For unparalleled path, it has been proved that coarse-grained fundamental frequency modification method can get a higher MOS of 3.94 than the original speech. For paralleled path, it is also proved that fine-grained fundamental frequency modification method makes the final speech include stress information of the speaker.

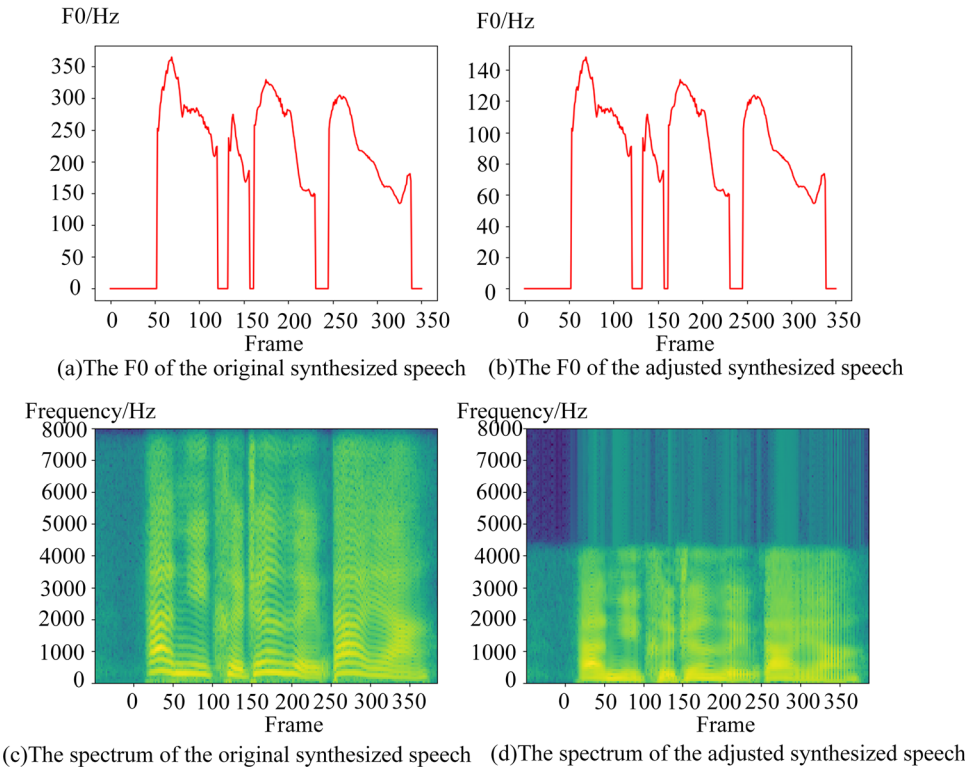


Fig. 15 The F_0 and spectrum of the original and personalized speech

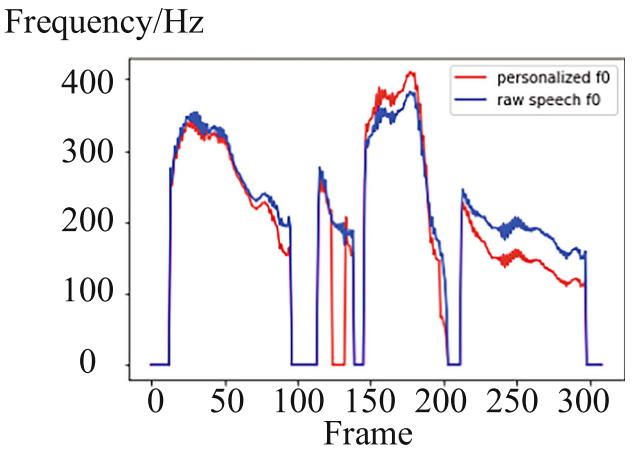


Fig. 16 The comparison of F_0 between the original and personalized speech

Table 8 The MOS of personalized speeches

Speech	MOS
our original synthesized speech	3.87
coarse-grained modified speech	**3.94**
fine-grained modified speech	3.72
SAG-Tacotron [30]	3.87
Ground Truth	4.50

4 Conclusions

In this paper, a speech synthesis framework with EGG signals based on the modified Tacotron-2 is proposed to utilize in some extreme environments where speech signals can hardly be collected. This framework consists of a text content recognition model and a speech synthesis model. To synthesize personalized speech, we propose a fine-grained fundamental frequency modification method.

The text content recognition model is to convert each EGG signal sample into the corresponding text with a category of content. This model achieves 91.12% accuracy on the validation set in a 20-class content recognition experiment. The comparative experiments show the following results: (1) The 3-layer Bi-LSTM gains higher accuracy than other recognition models we choose. (2) All of the three features contributes to the result and the combination of three features is more effective. (3) The smoothing method with bidirectional searching achieves better results than traditional methods.

The speech synthesis model is to synthesize the personalized speech with the corresponding text and EGG signals. Our model achieves a comparable result to the state-of-the-art performance according to both MCD and MOS. This model gains the mean opinion score (MOS) of 3.87 with relatively small model size and synthesizes the personalized speech with the MOS of 3.94, which is more consistent with the speaker's characteristics, with the aid of EGG signals. From the

comparative experiments, it can be proved that: (1) In terms of the text modeling units, phone is much better than pinyin. (2) Tacotron-2+depthwise separable convolution (channel multiplier=2) is better than other acoustic models considering the quality of synthesized speech and model size.

The expected future works are listed as follows. For the text content recognition model, the dataset will be expanded for more classes of contents to obtain a more general result. Considering the speech synthesis procedure, a better acoustic model will be explored to increase the speech quality and for other applications. For instance, to utilize it in portable devices, other modifications can be explored. Spiking neural networks (SNNs) [48, 49], as the third generation of neural network, comprise of spiking neurons. Addition of the temporal dimension for information encoding in SNNs yields new insight into the dynamics of the human brain and makes it potential to result in compact representations of large neural networks [50]. As such, SNNs have great potential for solving complicated time-dependent pattern recognition problems defined by time series. So it is a fascinating direction to apply SNN in speech synthesis in the future. As the development of spiking neural networks (SNNs) controlling mobile robots is one of the modern challenges in computational neuroscience and artificial intelligence [51], more motivations may arouse when associating the TTS task with neuromorphic computing [52–54]. For example, when dealing with TTS tasks based on large-scale datasets, to enhance the biological realism of neuromorphic systems and further understand the computational power of neurons, multicompartment emulation is an essential step to discuss [55]. Finally, for the fine-grained fundamental frequency modification method, a more proper alignment method will be explored.

Acknowledgements This research is supported in part by the National Science Foundation for Young Scientists of China under grant 61603013, National Natural Science Foundation of China under grant 62072021, and the Fundamental Research Funds for the Central Universities(No. YWF-21-BJ-J-534).

References

1. Fant G (1971) Acoustic Theory of Speech Production. De Gruyter Mouton. https://doi.org/10.1515/9783110873429
2. Tronchin L, Kob M, Guarnaccia C (2018) Spatial information on voice generation from a multi-channel electroglottograph. Applied Sciences 8(9) https://doi.org/10.3390/app8091560
3. Hussein H, Jokisch O (2007) Hybrid electroglottograph and speech signal based algorithm for pitch marking. In: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, ISCA, pp 1653–1656
4. Paul N, Kumar S, Chatterjee I, Mukherjee B (2011) Electroglottographic parameterization of the effects of gender, vowel and phonatory registers on vocal fold vibratory patterns An indian perspective. Indian Journal of Otolaryngology and Head & Neck Surgery 63(1):27–31. https://doi.org/10.1007/s12070-010-0099-0
5. Hui L, Ting LH, See SL, Chan PY (2015) Use of electroglottograph (egg) to find a relationship between pitch, emotion and personality. Procedia Manufacturing pp 1926–1931 https://doi.org/10.1016/j.promfg.2015.07.236
6. Macerata A, Nacci A, Manti M, Cianchetti M, Matteucci J, Romeo SO, Fattori B, Berrettini S, Laschi C, Ursino F (2017) Evaluation of the electroglottographic signal variability by amplitude-speed combined analysis. Biomedical Signal Processing and Control pp 61–68 https://doi.org/10.1016/j.bspc.2016.10.003
7. Chen L, Mao X, Wei P, Compare Angelo (2013) Speech emotional features extraction based on electroglottograph. Neural Computation 25:3294–3317. https://doi.org/10.1162/neco_a_00523
8. Borsky M, Mehta DD, Van Stan JH, Gudnason J (2017) Modal and nonmodal voice quality classification using acoustic and electroglottographic features. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(12):2281–2291. https://doi.org/10.1109/TASLP.2017.2759002
9. Sunil Kumar SB, Mandal T, Sreenivasa Rao K (2017) Robust glottal activity detection using the phase of an electroglottographic signal. Biomedical Signal Processing and Control 36:27–38. https://doi.org/10.1016/j.bspc.2017.03.007
10. Liu D, Kankare E, Laukkanen AM, Alku P (2017) Comparison of parametrization methods of electroglottographic and inverse filtered acoustic speech pressure signals in distinguishing between phonation types. Biomedical Signal Processing and Control 36(Jul.):183–193 https://doi.org/10.1016/j.bspc.2017.04.001
11. Lebacq J, Dejonckere PH (2019) The dynamics of vocal onset. Biomedical Signal Processing and Control 49:528–539. https://doi.org/10.1016/j.bspc.2019.01.004
12. Filipa MBL, Ternstrm S (2020) Flow ball-assisted voice training Immediate effects on vocal fold contacting. Biomedical Signal Processing and Control 62:102064. https://doi.org/10.1016/j.bspc.2020.102064
13. Niimi Y (2002) A chinese text to speech system based on td-psola. In: IEEE Region 10 Conference on Computers https://doi.org/10.1109/tencon.2002.1181250
14. Klatt Dennis H (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. The Journal of the Acoustical Society of America 87(2):820–857. https://doi.org/10.1121/1.398894
15. Atal BS (1982) A new model of lpc excitation for producing natural-sounding speech at low bit rates. Proc ICASSP. https://doi.org/10.1109/icassp.1982.1171649
16. Itakura F (1975) Line spectrum representation of linear predictive coefficients of speech signals. Journal of Acoustic Society of America 57:S35. https://doi.org/10.1121/1.1995189
17. Qingfeng L, Renhua W (1998) A new speech synthesis method based on the lma vocal tract model. Chinese Journal of Acoustics 02:153–162
18. Sotelo J, Mehri S, Kumar K, Santos JF, Kastner K, Courville AC, Bengio Y (2017) Char2wav End-to-end speech synthesis. In: 5th

International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings, OpenReview.net
19. Kawahara H (1999) Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction Possible role of a repetitive structure in sounds. Speech Communication 27. https://doi.org/10.1016/S0167-6393(98)00085-5
20. Morise M, Yokomori F, Ozawa K (2016) World A vocoder-based high-quality speech synthesis system for real-time applications. Ice Transactions on Information & Systems 99(7):1877–1884. https://doi.org/10.1587/transinf.2015edp7457
21. Agiomyrgiannakis, Y(2015) Vocaine the vocoder and applications in speech synthesis. In: IEEE International Conference on Acoustics(ICASSP), pp 4230–4234 https://doi.org/10.1109/icassp.2015.7178768
22. J S, R P, J WR, et al (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 4779–4783 https://doi.org/10.1109/icassp.2018.8461368
23. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior AW, Kavukcuoglu K (2016) Wavenet A generative model for raw audio. In: The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016, ISCA, p 125
24. Arik SÖ, Chrzanowski M, Coates A, Diamos GF, Gibiansky A, Kang Y, Li X, Miller J, Ng AY, Raiman J, Sengupta S, Shoeybi M (2017) Deep voice Real-time neural text-to-speech. In: Precup D, Teh YW (eds) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, PMLR, Proceedings of Machine Learning Research, vol 70, pp 195–204
25. Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le Q, Agiomyrgiannakis Y, Clark R, Saurous RA (2017) Tacotron Towards end-to-end speech synthesis. In: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, ISCA, pp 4006–4010 https://doi.org/10.21437/interspeech.2017-1452
26. Gibiansky A, Arik SÖ, Diamos GF, Miller J, Peng K, Ping W, Raiman J, Zhou Y (2017) Deep voice 2 Multi-speaker neural text-to-speech. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in Neural Information Processing Systems 30 Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp 2962–2970
27. Ping W, Peng K, Gibiansky A, Arik SÖ, Kannan A, Narang S, Raiman J, Miller J (2018) Deep voice 3 Scaling text-to-speech with convolutional sequence learning. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net
28. Yasuda Y, Wang X, Takaki S, Yamagishi J (2019) Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 6905–6909 https://doi.org/10.1109/ICASSP.2019.8682353
29. Liu R, Sisman B, Li J, Bao F, Gao G, Li H (2020) Teacher-student training for robust tacotron-based TTS. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, IEEE, pp 6274–6278 https://doi.org/10.1109/ICASSP40776.2020.9054681
30. Yang F, Yang S, Zhu P, Yan P, Xie L (2019) Improving mandarin end-to-end speech synthesis by self-attention and learnable gaussian bias. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp 208–213 https://doi.org/10.1109/ASRU46091.2019.9003949
31. Lu Y, Dong M, Chen Y (2019) Implementing prosodic phrasing in chinese end-to-end speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019, IEEE, pp 7050–7054 https://doi.org/10.1109/ICASSP.2019.8682368
32. Pan J, Yin X, Zhang Z, Liu S, Zhang Y, Ma Z, Wang Y (2020) A unified sequence-to-sequence front-end model for mandarin text-to-speech synthesis. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, IEEE, pp 6689–6693 https://doi.org/10.1109/ICASSP40776.2020.9053390
33. Jing S, Mao X, Chen L et al (2015) Annotation and consistency detection of chinese dual-mode emotional speech database. Journal of Beijing University of Aeronautics and Astronautics 41(10):1925–1934. https://doi.org/10.13700/j.bh.1001-5965.2014.0771
34. Chen P, Chen L, Mao X (2020) Content classification with electroglottograph. Journal of Physics Conference Series 1544:012191. https://doi.org/10.1088/1742-6596/1544/1/012191
35. Irie K, Tuske Z, Alkhouli T, Schluter R, Ney H (2016) Lstm, gru, highway and a bit of attention An empirical overview for language modeling in speech recognition. In: Interspeech 2016 https://doi.org/10.21437/interspeech.2016-491
36. Prukkanon N, Chamnongthai K, Miyanaga Y (2016) F0 contour approximation model for a one-stream tonal word recognition system. AEUE - International Journal of Electronics and Communications 70(5):681–688. https://doi.org/10.1016/j.aeue.2016.02.006
37. Xiao Z (2001) An approach of fundamental frequencies smoothing for chinese tone recognition. Journal of Chinese Information Processing 15:45–50. https://doi.org/10.3969/j.issn.1003-0077.2001.02.007
38. Chiu JPC, Nichols E (2015) Named entity recognition with bidirectional lstm-cnns. Computer Science. https://doi.org/10.1162/tacl_a_00104
39. Shen J, Pang R, Weiss R, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerry-Ryan R, Saurous R, Agiomyrgiannakis Y, Wu Y (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 4779–4783 https://doi.org/10.1109/icassp.2018.8461368
40. Chollet F (2017) Xception Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) https://doi.org/10.1109/cvpr.2017.195
41. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets Efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861, 1704.04861
42. Wang J, Xiong H, Wang H, Nian X (2020) Adscnet asymmetric depthwise separable convolution for semantic segmentation in real-time. Appl Intell 50(4):1045–1056. https://doi.org/10.1007/s10489-019-01587-1
43. Wang Z, Yan W, Oates T (2017) Time series classification from scratch with deep neural networks A strong baseline. 2017 International Joint Conference on Neural Networks (IJCNN) https://doi.org/10.1109/ijcnn.2017.7966039
44. Jing L, Gulcehre C, Peurifoy J, Shen Y, Tegmark M, Soljačić Bengio Y (2017) Gated orthogonal recurrent units On learning to forget. Neural Computation 31:765–783. https://doi.org/10.1162/neco_a_01174
45. Kingma DP, Ba J (2015) Adam A method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings

46. Hu X, Jing L, Sehar U (2021) Joint pyramid attention network for real-time semantic segmentation of urban scenes. Applied Intelligence. https://doi.org/10.1007/s10489-021-02446-8
47. Kubichek R (1993) Mel-cepstral distance measure for objective speech quality assessment. Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing 1:125–128 vol.1
48. Yang S, Gao T, Wang J, Deng B, Linares-Barranco B (2021) Efficient Spike-Driven Learning With Dendritic Event-Based Processing. Frontiers in Neuroscience 15. https://doi.org/10.3389/fnins.2021.601109
49. Yang S, Wang J, Deng B, Azghadi MR, Linares-Barranco B (2021b) Neuromorphic Context-Dependent Learning Framework With Fault-Tolerant Spike Routing. IEEE Transactions on Neural Networks and Learning Systems pp 1–15 https://doi.org/10.1109/TNNLS.2021.3084250
50. Ghosh-Dastidar S, Adeli H (2009) Spiking neural networks. International Journal of Neural Systems 19(04):295–308. https://doi.org/10.1142/S0129065709002002
51. Lobov SA, Mikhaylov AN, Kazantsev VB (2020) Spatial Properties of STDP in a Self-Learning Spiking Neural Network Enable Controlling a Mobile Robot. Frontiers in Neuroscience 14:88. https://doi.org/10.3389/fnins.2020.00088
52. Yang S, Wang J, Deng B, Liu C, Li H, Fietkiewicz C, Loparo KA (2019) Real-Time Neuromorphic System for Large-Scale Conductance-Based Spiking Neural Networks. IEEE Transactions on Cybernetics 49(7):2490–2503. https://doi.org/10.1109/TCYB.2018.2823730
53. Yang S, Wang J, Hao X, Li H, Wei X, deng B, Loparo K (2021a) Bicoss Toward large-scale cognition brain with multigranular neuromorphic architecture. IEEE Transactions on Neural Networks and Learning Systems PP:1–15 https://doi.org/10.1109/TNNLS.2020.3045492
54. Yang S, Wang J, Zhang N, Deng B, Pang Y, Azghadi MR (2021b) CerebelluMorphic Large-Scale Neuromorphic Model and Architecture for Supervised Motor Learning. IEEE Transactions on Neural Networks and Learning Systems pp 1–15 https://doi.org/10.1109/TNNLS.2021.3057070
55. Yang S, Deng B, Wang J, Li H, Lu M, Che Y, Wei X, Loparo KA (2020) Scalable Digital Neuromorphic Architecture for Large-Scale Biophysically Meaningful Neural Network With Multi-Compartment Neurons. IEEE Transactions on Neural Networks and Learning Systems 31(1):148–162. https://doi.org/10.1109/TNNLS.2019.2899936