Hindawi
Scientific Programming
Volume 2022, Article ID 3638136, 9 pages
https://doi.org/10.1155/2022/3638136

Research Article

Research on Discourse Transfer Analysis Based on Deep Learning of Cross-language Transfer

Yu Shen [1,2]

[1]*Huanghe Science and Technology University, Foreign Languages School, Zhengzhou 450000, China*
[2]*University of Málaga, Department of Linguistic, Literature and Translation, Málaga, Spain*

Correspondence should be addressed to Yu Shen; shenyu@hhstu.edu.cn

Received 15 April 2022; Revised 4 May 2022; Accepted 11 May 2022; Published 27 May 2022

Academic Editor: Jie Liu

With the current exchange and communication between different countries becoming more and more frequent, the language conversion of different countries has become a difficult problem. The analysis of a series of problems in cross-language discourse conversion, the study of the discourse conversion path, and innovation motivation based on the deep learning theory of cross-language transfer, it has theoretical and practical significance. This paper aims at the technical difficulties in speech conversion methods to effectively utilize the local mode information of signal time spectrum and the long-term correlation of speech signal. A discourse conversion method based on convolutional recurrent neural network model is proposed. In the model, the extended convolutional neural network is used to model the long-term correlation of speech signals. In the part of speech fundamental frequency estimation, the prosodic information generated by the decomposition of the fundamental frequency by continuous wavelet transform is used as the training target of the fundamental frequency estimation model. The experimental results show that the speech transformation method based on the convolutional cyclic network model proposed in this paper has better quality and intelligibility than the speech transformed by the contrast method.

1. Introduction

In the face of the diversity of social values and cultural diversity, coupled with the development of new media technology, traditional ideological and political education discourse is facing inevitable challenges. In the process of dealing with challenges, its disadvantages and problems are constantly exposed. For example, most of the discourse content is still confined to the propaganda of documentary language and policy discourse, which lacks effective connection with the daily life of the audience [1]. In the form of discourse, one-way indoctrination is more than two-way interaction; there are more grand narratives and less elaborate descriptions; more empirical life language, less rigorous academic discourse; and there are more referential words but less original ones extracted from practice. There are more empty and stale words than up-to-date ones; there are many words of conformity, but few words of independent thinking [2]; and the lack of ideological discourse power in the field of network. In this discourse field, therefore, how to use by educators to understand, trust, and open discourse established the ideological and political education position, the spread of the ideological and political education content, firmly grasp the ideological and political education, in turn, say, raise new era the pertinence, effectiveness and validity of ideological and political education, and ideological and political education has become the key problem facing. [3–6].

With the rise of deep learning and artificial intelligence, traditional speech conversion methods based on statistical models can no longer meet the requirements of large amounts of corpus data involved in training, and the performance of traditional language conversion models deteriorates rapidly in the case of large amounts of corpus data involved in training. DNN [7] has strong data fitting ability and can better explain various complex data features, which is very suitable for the scenario where a large amount of corpus data participates in training. Deep belief network (DBM) maps speakers' spectrum features to higher order eigenspace, thus realizing the transformation between speakers' spectrum features.

Based on dnnvc (deep bidirectional long-term and short-term memory recursive neural network), it has recursive neural network, Dblstm RNN) to construct the discourse transformation model. Because dblstm-rnn can capture the forward and backward time relationship of the speaker's speech spectrum characteristics, the performance of the conversion model is significantly improved [8]. The proposal of convolutional neural network (CNN) [9] is a milestone in the field of deep neural network, which greatly promotes the development of deep learning and artificial intelligence. The depth generation network model based on convolutional neural network has been proposed and applied to the field of discourse conversion. Conditional variational automatic encoder network (CVAE) is used to unlock the content and timbre characteristics of the input speech spectrum [10]. Completely nonparallel many-to-many discourse transformation [11]. Star creative adversity network VC (stargan VC) method makes use of the advantages of the cvae-vc method and the cyclegan VC method, respectively. Multi-speaker multi to multidiscourse conversion is realized by using speaker identity tag thermal vector, which is the best in the current nonparallel multi to multi discourse conversion methods. The improved methods of VAE series and Gan Series in deep generation neural network model have been recognized and affirmed by many scholars, and a series of novel methods have been proposed, such as vawgan VC [12], vqvae-vc [13], cdvae VC [14], acvae-vc [15], adagan VC [15], cyclegan-vc2 [16], and stargan-vc2 [17].

Based on the above research, this paper innovatively uses the deep neural network model of cross-language transfer to solve the discourse conversion problem. In order to solve the problem that existing speech conversion methods cannot effectively utilize the acoustic mode information in the speech time spectrum, and it is difficult to effectively model the long-term correlation of speech signals, a novel convolutional recurrent neural networks based on convolutional recurrent neural networks is proposed. CRNN, which uses extended convolutional network to describe the pattern information of the discourse spectrum and model the long-term correlation of signals, and BiLSTM conduct the time sequence modeling. The performance of this method is better than that of BiLSTM.

2. Model Theory

In order to effectively describe the acoustic pattern information of speech in the time-frequency domain, model the long-term correlation of signals, and improve the naturalness of translated speech, a convolutional recurrent neural network with continuous wavelet transform is proposed in this paper. This CRNN model combined with the advantages of neural network, signal processing theory, and depth can use signal processing methods to obtain more suitable for the acoustic characteristics of the task and to make full use of the depth of the neural network nonlinear description ability to the words the local characteristics of spectrum and long correlation model, so as to achieve better performance of discourse transformation [18].

2.1. Discourse Conversion Model of Convolution Recurrent Neural Network with Continuous Wavelet Transform. Continuous wavelet transform (CWT) is a commonly used time-frequency analysis tool [19]. The traditional fixed-window transform algorithm (such as Fourier transform) determines the size and shape of the time-frequency window after selecting the window function and has the same ability to analyze both high and low frequencies. However, in practical signal analysis, we usually expect the algorithm to have different time-frequency resolution in different frequency bands. Continuous wavelet transform is an algorithm to solve this kind of problem, and its calculation process is shown in formula (1):

$$WT_f(a,\tau) = \prec f(t),$$
$$\psi_{a,\tau}(t) \succ = a^{-1/2}\int_R f(t)\overline{\psi\left[\frac{(t-\tau)}{a}\right]}\mathrm{dt}. \tag{1}$$

In the formula, $F(t)$ represents the original signal, A represents the scale factor in the wavelet transform, τ represents the translation factor, and the wavelet basis function ψ increases with the increase of the scale factor, the time window function also increases, and the frequency resolution of the unit increases correspondingly, otherwise, the time resolution increases. When the wavelet basis function meets the admissible condition, the algorithm has contravariant transformation, and the Morlet wavelet basis satisfying the condition is adopted in this paper. The fundamental frequency component predicted by the model can be reconstructed into the fundamental frequency feature by inverse wavelet transform. The inverse wavelet transform formula is as follows:

$$x(t) = \frac{1}{C_\psi}\int_0^{+\infty}\frac{\mathrm{da}}{a^2}\int_{-\infty}^{+\infty}\mathrm{WT}_x(a,\tau)\psi_{a,\tau}(t)\mathrm{d}\tau,$$
$$= \frac{1}{C_\psi}\int_0^{+\infty}\frac{\mathrm{da}}{a^2}\int_{-\infty}^{+\infty}\mathrm{WT}_x(a,\tau)a^{-1/2}\psi\left(\frac{t-\tau}{a}\right)\mathrm{d}\tau. \tag{2}$$

$X(t)$ represents the reconstructed signal, where the calculation method of admissible conditions is given by formula (3):

$$C\psi = \int_0^{\infty}\frac{|\psi(a\omega)|}{a}\mathrm{da} \prec \infty. \tag{3}$$

2.2. CNN Model. CNN is a commonly used neural network structure. Different from fully connected networks, the neurons of CNN are usually arranged in three dimensions. In the field of audio processing, 2d convolutional neural networks are usually used. In 2d convolutional kernels, the height and width correspond to the size of the time-frequency window of the convolution kernels, that is, the time-frequency range of each convolution of the convolution kernels. The depth of the convolution kernel corresponds to the number of channels of features after convolution. Usually, the depth of the convolution kernel used is gradually increased at the beginning of the model to improve the

fitting ability of the model, while the depth of the convolution kernel is gradually reduced at the output end of the model to map features to the target dimension. Figure 1 shows a schematic diagram of a two-dimensional convolution kernel.

Assuming that the input feature of this example is C^l, the eigenvalue of the output of the convolutional network is C^{l+t}, and the target feature is C^{ture}, the convolution operation can be expressed by formula (4):

$$C^{l+1}(i,j) = \left[C^l \otimes w^{l+1}\right](i,j) + b^{l+1} = \sum_{k+1}^{Kl}\sum_{x=1}^{f}\sum_{y=1}^{f} \left[C_k^l(s*i+x, s*j+y) * w^{l+1}(x,y)\right] + b^{l+1},$$
$$\left(i,j \in \{0,1,\ldots,L_{l+1}\}, L_{l+1} = \frac{L_l + 2*z - f}{s} + 1\right). \tag{4}$$

In the formula, w and B, respectively, represent the weight matrix and bias of the convolution kernel; I and j represent the number of pixels of the feature graph; f, z and s correspond to the size, filling number, and step size of the convolution kernel. The training of convolutional network requires the setting of loss function, and the commonly used MSE loss function can be expressed by formula (5):

$$\text{MSE} = \left[\frac{1}{D}\sum_{i=1}^{D}\left(\frac{1}{2}\left(C^{\text{ture}} - C^{l+1}\right)^2\right)\right], \tag{5}$$

where D represents the corresponding feature dimension.

The extended convolutional neural network is a special convolutional network whose filter is discontinuous. Studies have found that such network structure with spacing between filters can make the convolution kernel have a large receptive field with minimal precision loss. The following figure shows the schematic diagram of the receptive field range of an ordinary $3 * 3$ convolution kernel and an extended $3 * 3$ convolution kernel.

The rectangular block in Figure 2 represents the feature graph, and the deepened part represents the convolution region of the convolution kernel filter. As can be seen from the figure, in the case of the same convolution kernel size, the receptive field of the extended convolutional network is larger than that of the ordinary convolutional network. This feature enables the model to have a larger receptive field under the same conditions and enables the network model to be capable of modeling longer context information.

2.3. CRNN Model. CRNN is mainly used to recognize text sequences of indeterminate length end-to-end, without cutting a single text first, but transforming text recognition into a sequence-dependent sequence learning problem, which is image-based sequence recognition. Figure 3 shows the structure diagram of the CRNN model used in this paper. After the acoustic features of ear discourse are input into the model, the feature extraction module is used to obtain the local features of the discourse spectrum.

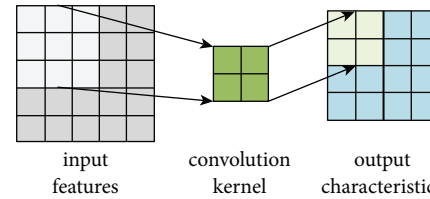

Figure 1: Schematic diagram of two-dimensional convolution kernel.

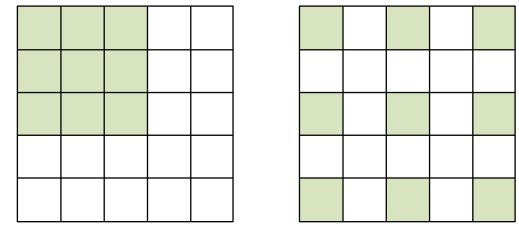

Figure 2: Comparison of receptive fields of different convolution structures. (a) Ordinary convolution, (b) Expansion of the convolution.

Feature extraction module is composed of two sets of two-dimensional dilated convolution. One set of convolution layer uses a convolution kernel with a size of 3×3. The first dimension of the convolution kernel corresponds to the time direction of the discourse feature sequence and makes the convolution layer perform dilation in the time domain direction, which is called the time domain dilated convolution layer. Another set of convolution layers performs frequency domain expansion using convolution kernels of the same size.

The characteristic graph output by the time-frequency expansion module is connected and reconstructed into one-dimensional features and then input into the time-domain modeling module. The time-domain modeling module consists of a group of time-domain expansion blocks, whose structure is shown in Figure 3. To model discourse long-term correlation, one-dimensional dilated convolution was used in each dilated block and Gated Linear Units (GLUs) were used to improve the stability of the model during training. The calculation process of GLUs is shown in formula (6):

$$y = \sigma(x * W_1 + b_1) \otimes (x * W_2 + b_2), \tag{6}$$

where W_1 and W_2 represent the weight of the convolution layer, b_1 and b_2 represent the corresponding bias term, σ represents the sigmoID activation function, and $\otimes$ represents the element-by-element multiplication symbol. The calculation process of the MISH activation function used in the expansion block can be expressed by formula (7):

$$\text{MISH} = x * (\tanh(\text{softplus}(x))). \tag{7}$$

In the formula, TANH and Softplus represent corresponding activation functions, respectively, and the calculation process of softPLu function is described in formula (8).

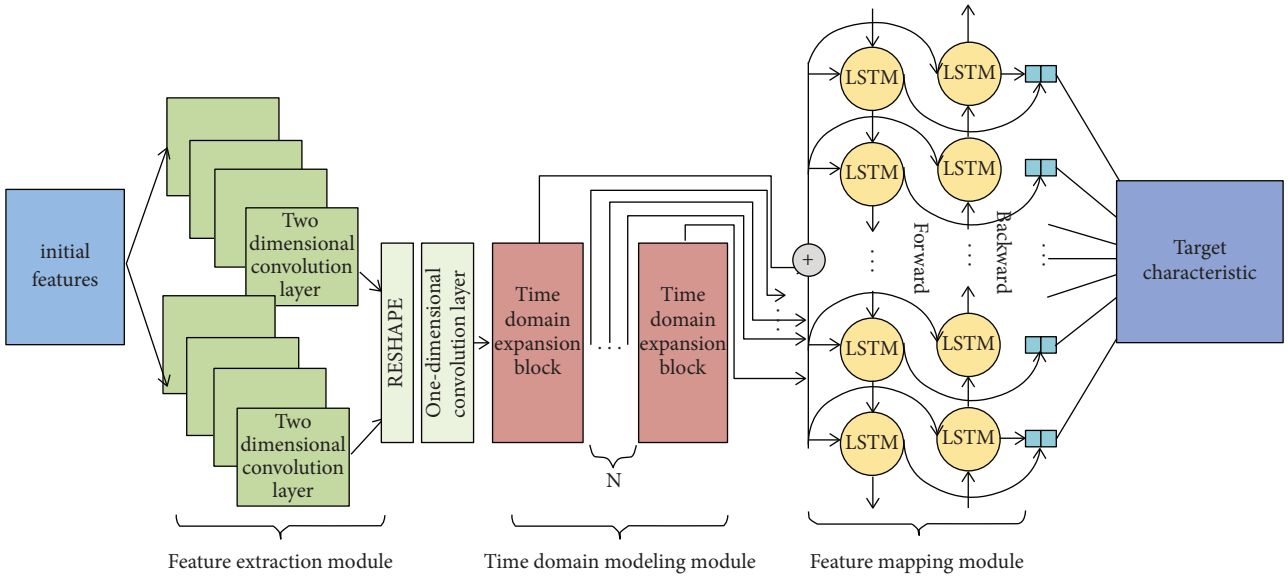

Figure 3: Structure of utterance conversion model based on CRNN.

$$\text{softplus} = \log\left(1 + e^{x}\right). \tag{8}$$

It can be seen from formula (8) that the function has a small number of negative intervals, which provides an additional flow interval for the gradient flow, thus alleviating the gradient problem of the network. The input of adjacent time domain expansion block is the output A of the previous expansion block, and the input of feature mapping module is obtained by adding the output B of each expansion block element by element.

The output of the feature mapping module is calculated by the two groups of memory cells with opposite directions, and the calculation process can be expressed by the following formula:

$$\begin{aligned} \overrightarrow{h}_t &= \text{lstm}\left(x_t\right), \\ \overleftarrow{h}_t &= \text{lstm}\left(x_t\right), \\ y_t &= W_{\underset{hy}{\rightarrow}} \overrightarrow{h}_t + W_{\underset{hy}{\leftarrow}} \overleftarrow{h}_t + b_y. \end{aligned} \tag{9}$$

The calculation process of LSTM in the above formula can be expressed by the following formula:

$$\begin{aligned} i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i i_t\right), \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{ci}c_{t-1} + b_f\right), \\ c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1}b_c\right), \\ o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c + b_o\right), \\ h_t &= o_t \tanh\left(c_t\right). \end{aligned} \tag{10}$$

In the above formula, I, F, O, and C correspond to the input gate, forgetting gate, output gate, and cell state in the cell structure, respectively. O represents the commonly used SigmoID activation function, and W and B represent the weights and bias items to be learned during network training. Because the time-domain modeling module uses a large number of extended convolutional neural networks, the feature graph input by each neuron in the circular layer of the feature mapping module contains the whole discourse context information of the input model, which is beneficial to the model to describe the long-term correlation of signals.

2.4. Proposed Ear Discourse Conversion. The proposed ear discourse conversion method based on the CRNN model is shown in Figure 4. During the model training, the STRAIGHT model was used to extract the characteristic parameters of the two kinds of discourse, respectively. As mentioned above, the STRAIGHT model is a classical parametric vocoder, which has been widely used in speech analysis and synthesis tasks. After extracting relevant parameters, DTW algorithm is used to align feature sequences. Then, the spectral envelope features are converted to MCC features, and the normal speech fundamental frequency is decomposed by continuous wavelet transform. Finally, the MCC feature estimation model (CRNN_mcc) was trained using MCC features of ear speech and normal speech.

In the transformation stage, the extracted ear speech spectrum envelope is converted into MCC features, then the MCC features are input to the two transformation models after training to obtain the MCC features and nonperiodic components estimated by the model, and then the MCC features estimated by the model are input to the CRNN_f0 model to obtain the estimated fundamental frequency components. Then, the inverse of the estimated MCC feature is transformed into a spectral envelope, and the obtained fundamental frequency component is reconstructed into the speech fundamental frequency by inverse wavelet transform. Finally, the spectral envelope, aperiodic component, and fundamental frequency predicted by the model are reconstructed into transformed discourse by the STRAIGHT model.

Table 1 shows that the input and output parameters of two-dimensional convolution are frame number, frequency channel, and characteristic image channel in turn. The

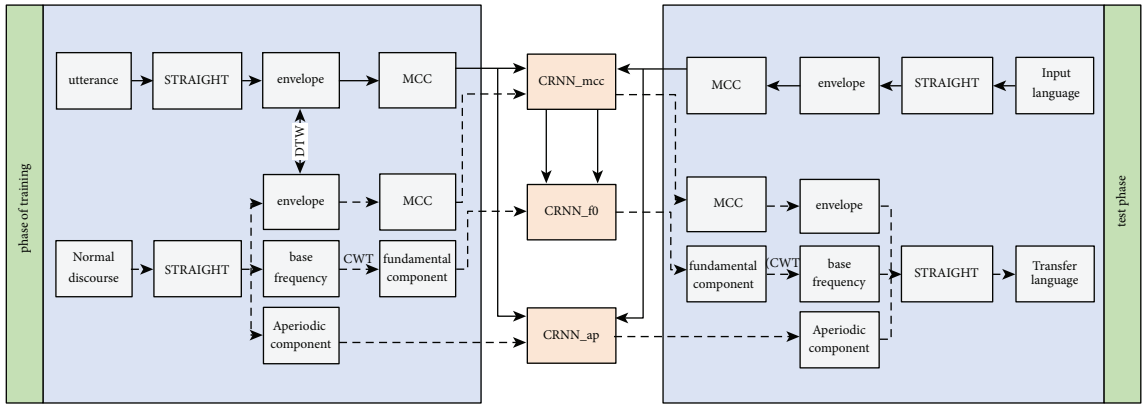

Figure 4: Flow chart of CRNN-based utterance conversion method.

Table 1: Parameter configuration of the CRNN utterance conversion model.

Network layer	Input size	Super parameter	Output size
Expand	(150 × 30)	—	(150 × 30 × 1)
Conv2d_1.1 Conv2da_1.2	(150 × 30 × 1)	(3 × 3, (1, 1), 16) (3 × 3, (1, 1), 16)	(150 × 30 × 1) (150 × 30 × 16)
Conv2d_2.1 Conv2d_2.2	(150 × 30 × 16)	(3 × 3, (1, 1), 16) (3 × 3, (1, 1), 16)	(150 × 30 × 16)
Conv2d_31 Conv2d_3.2	(150 × 30 × 16)	(3 × 3, (2, 1), 32) (3 × 3, (1, 2), 32)	(150 × 30 × 32)
Conv2d_4.1 Conv2d_4.2	(150 × 30 × 32)	(3 × 3, (4, 1), 32) (3 × 3, (1, 4), 32)	(150 × 30 × 32)
Concatenate	(150 × 30 × 32) (150 × 30 × 32)	—	(150 × 30 × 64)
Reshape	(150 × 30 × 64)		(150 × 1920)
Conv1d_1	(150 × 1920)	(1, (1), 512)	(150 × 512)
TD block		(1, (1), 256) (3, (1), 128) (1, (1), 512)	
TD block		(1, (1), 256) (3, (2), 128) (1, (1), 512)	
TD block	(150 × 512)	(1, (1), 256) (3, (4), 128) (1, (1), 512)	(150 × 512)
TD block		(1, (1), 256) (3, (8), 128) (1, (1), 512)	
TD block		(1, (1), 256) (3, (16), 128) (1, (1), 512)	
BiLSTM	(150 × 512)		(150 × 1024)
Dense	(150 × 1024)	(30/30/513)	(150 × 30/30/513)

parameters of convolution layer represent the size, expansion rate, and number of convolution kernels, respectively. The input and output parameters of one-dimensional convolution are tonnage and frequency channel in turn. The convolution layer parameters have the same meaning as two-dimensional convolution. In order to keep the temporal characteristics of discourse unchanged, zeroing is applied to all convolution layers to maintain the consistency of input and output dimensions. Only one set of time domain block parameters is shown in the table, and three sets of time domain expansion blocks with the same parameters are stacked in the model. The TD block represents the time domain extension block. The output of recurrent neural network is the symbiosis of the output of two groups of neurons. Therefore, this paper splices the output of two groups of LSTM and uses the full connection layer to map the feature map to the target dimension.

This method uses the function shown in formula (11) as the training error function in the training process:

Table 2: Impact of time-frequency dilated convolution on model performance.

Convolution kernels	CD	PESQ	STOI
3×3	4.5826	1.2679	0.6003
Time-frequency dilated convolution	4.5163	1.3201	0.6104

Table 3: Impact of time domain dilated block on model performance.

Model	CD	PESQ	STOI
CRNN_nt	4.6532	1.2895	0.5765
CRNN_ot	4.5885	1.3111	0.6004
The model in this paper	4.5163	1.3201	0.6104

$$\text{loss} = \frac{1}{N}\sum_{t=1}^{N}\frac{10}{\log 10}\sqrt{2\sum_{i=1}^{m}\left(y_i - Y_i\right)^2}. \quad (11)$$

In the above formula, y_i and y_i represent target feature and prediction feature, respectively.

3. Experimental Simulation and Result Analysis

3.1. Experimental Data and Evaluation Indicators. To further evaluate the performance of the proposed method in the auditory speech conversion task, 348 auditory utterance and corresponding target sounds from the wTIMIT discourse database were selected as experimental data. The selected corpus has a sampling rate of 8000 Hz and is stored in 16 bit PCM format. When extracting speech features, the frame length is 40 ms, the frame offset is 5 ms, and 1024 point fast Fourier transform is used for each frame of speech. In total, 313 auditory utterances and their corresponding normal utterances were randomly selected as the training set, and the other 35 corpora were used as the test set. The relevant test set has strong adaptability.

All the above methods use the straight algorithm to analyze the reconstructed discourse. The GMM method and the DNN method in the comparison method are limited by the model structure and cannot be modeled by using the dynamic correlation between frames of discourse. In order to improve the algorithm performance of the comparison method and make the effectiveness of the proposed method more convincing, the dynamic characteristics of speech frames are taken as the training parameters of the two methods. The calculation formula of dynamic characteristics is given by formula (12).

$$\text{sp\_dy}_k = \frac{\left(-2 * sp_{k-2} - \text{sp}_{k-1} + \text{sp}_{k+1} + 2 * \text{sp}_{k+2}\right)}{3}, \quad (12)$$

sp_dy_k represents the corresponding dynamic feature.

The specific parameter configuration of the comparison method is described as follows: in the gMM-based ear speech conversion method, three models, GMM_mcc, GMM_ap, and GMM_f0, are, respectively, trained to estimate the MCC, aperiodicity and fundamental frequency of normal sounds. The Gaussian component number of GMM MCC and GMM_f0 is set to 32, and the Gaussian component number of GMM_ap is set to 16. In the DNN ear speech conversion method, three DNN models are trained to estimate the MCC feature, nonperiodic component and fundamental frequency of target speech. The structure of THE DNN model is 30×30-900-1024-2048-1024-1024-900/7710/30. The Dropout technology is used for the hidden layers of the model to improve the model and reduce overfitting. The Dropout parameter value is set to 0.9, and the three dimensions of the output layer correspond to the three different characteristics of the prediction. For BiLSTM, three BiLSTM models are also trained, respectively, to estimate the acoustic characteristics of transformed discourse. The BiLSTM used contains two hidden layers with 512 units. All comparison methods adopted MSE objective function, and Adam algorithm was used to optimize model parameters, with a learning rate of 0.0001.

Table 4: Quality evaluation of converted speech by different methods.

Transfer approach	CD	PESQ	STOI
GMM	5.4415	1.0121	0.4603
DNN	5.1732	1.0901	0.5062
BiLSTM	4.8611	1.2523	0.5559
The model in this paper	4.5163	1.3201	0.6104

Table 5: RMSEs of fundamental frequency of different methods.

Model	GMM	DNN	BiLSTM	CRNN	The model in this paper
RMSE (HZ)	121.09	88.76	81.14	69.27	66.93

3.2. Model Parameter Selection. In order to evaluate the influence of extended convolution in the time-frequency domain of the feature extraction module on the translated speech quality, the traditional 3×3 single-size convolution kernel and the extended convolution in the time-frequency domain used in this paper were used to conduct the speech conversion experiment. It is obvious from Table 2 that the time-frequency expansion convolution adopted in this paper is conducive to improving the discourse conversion performance of the model. The specific comparison of discourse quality after transformation is shown in Table 2.

In order to explore whether the time domain expansion block in the time domain modeling module can effectively improve the performance of the discourse conversion method, the transformed discourse of CRNN discourse conversion method in three cases is compared. Table 3 shows the comparison of discourse quality after transformation under the three conditions, where CRNN_nt represents the CRNN model without time domain dilators and CRNN_ot represents only one group of time domain dilators. As can be seen from Table 3, the CRNN model without the time-domain expansion block has the worst performance of discourse conversion. The prediction accuracy of the CRNN model that only uses a group of time-domain dilators is lower than that of the method in this paper, because the CRNN model that only contains a group of time-domain dilators is difficult to use the context information of the

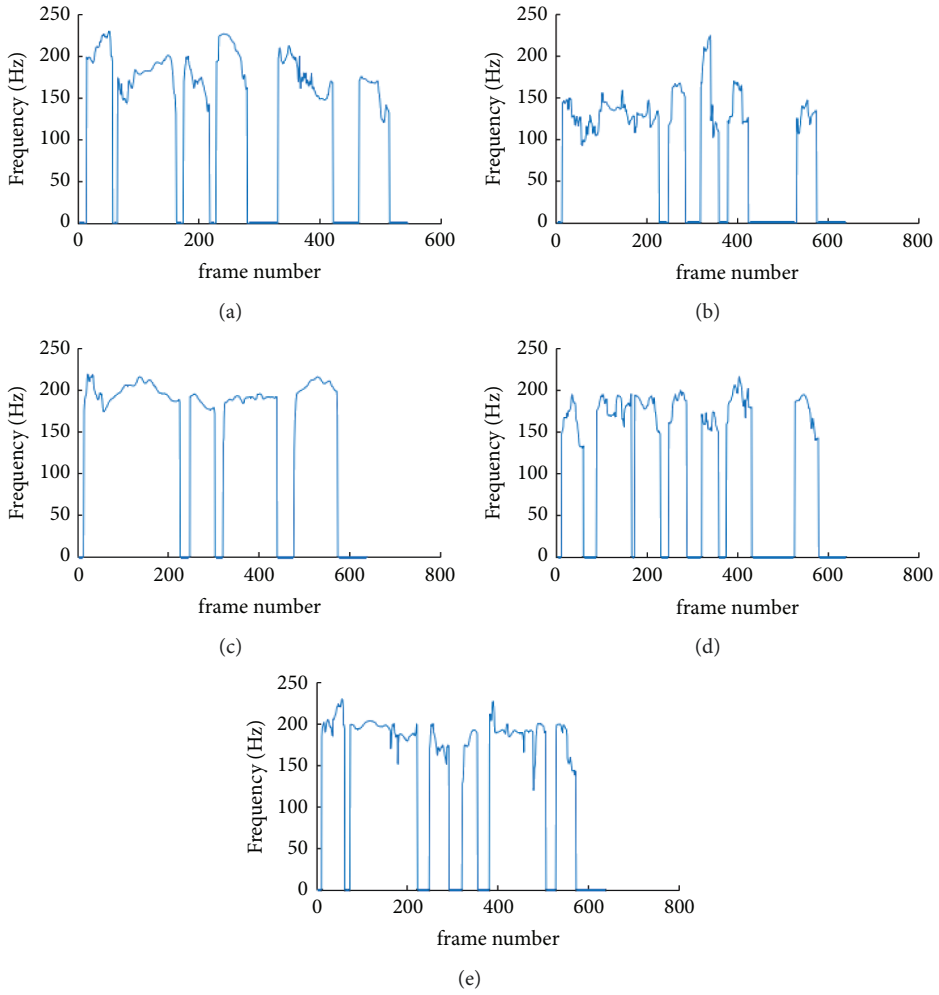

Figure 5: Fundamental frequency curves obtained by different conversion methods. (a) Enter the words, (b) GMM, (c) DNN, (d) BiLSTM, (e) CRNN.

whole input discourse, and the model has a small receptive field, which makes it impossible to effectively model the long-term correlation of discourse. Therefore, this paper finally sets up three groups of time domain expansion blocks in the CRNN network.

3.3. Comparative Analysis of Experimental Results. To demonstrate the effect of CRNN discourse conversion model, GMM, DNN, and BiLSTM are used as comparison models. Table 4 shows the evaluation results of transformed discourse. The performance of the GMM model is poor because the modeling ability of GMM is weaker than that of the neural network model. Although the DNN method can well represent the nonlinear mapping relationship, it cannot model the long-term correlation of discourse, and the effect of discourse conversion is not ideal. Compared with the DNN method, the BiLSTM method can make better use of the interspeech correlation. When the time step is large, the BiLSTM method can also model the long-term correlation of discourse, so the conversion effect is better than the GMM method and the DNN method. However, BiLSTM is difficult to effectively utilize the local features in the time-frequency domain of discourse, resulting in some spectral errors in the transformed discourse. As can be seen from Table 4, compared with the comparison method, the effect of the CRNN speech conversion model under the quality assessment of different conversion speech methods is 4.5163 s in the disc time; 1.3201 s in the ticker; and 1.3201 s in the station time 0.6104 s, the utterances transformed by this method in this paper have the best inhomogeneous quality and intelligibility.

Table 6: MOS of converted speech by different methods.

Model	GMM	DNN	BiLSTM	The model in this paper
MOS	2.35	2.51	2.82	2.90

Table 5 shows RMSE values of fundamental and target tones predicted by four conversion methods, where, CRNN indicates that CWT is not used to convert the discourse fundamental frequency in the training process. As can be seen from Table 5, GMM is difficult to accurately estimate the fundamental frequency characteristics of utterances, and BiLSTM has better fundamental frequency estimation

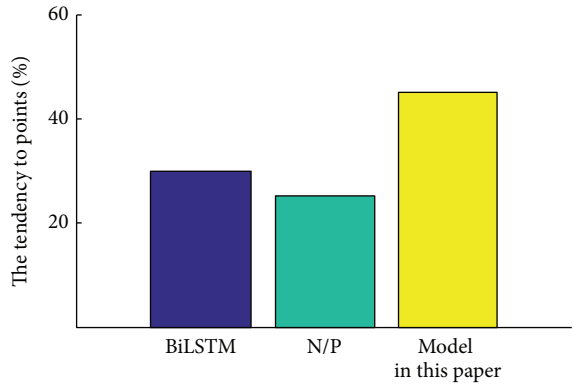

Figure 6: ABX test.

performance than DNN. In the process of model training, CWT decomposition of fundamental frequency can improve the prediction accuracy of the model to a certain extent. A horizontal comparison of the five methods shows that the difference between the fundamental frequency predicted by the proposed method and the target fundamental frequency is the smallest, which proves that the proposed method has higher fundamental frequency prediction accuracy compared with the comparison method.

It can be seen from Figure 5 that the GMM speech conversion method is difficult to fit the fundamental frequency curve of target speech effectively, and the fundamental frequency of transformed speech is greatly different from that of the target speech. The DNN method can only estimate unvoiced speech conversion, but cannot accurately predict the fundamental frequency curve. The fundamental frequency curve estimated by the BiLSTM method has a certain similarity with the target curve, but there is still a great difference with the expected target in details such as 170–190 frames and 230–270 tons. However, the overall trend of the speech fundamental frequency curve estimated by the proposed method is close to that of the target fundamental frequency, which indicates that the proposed method has better fundamental frequency estimation performance.

Table 6 shows MOS scores of discourse obtained after four methods of transformation. As can be seen from Table 6, the comfort level of discourse listening sensation after GMM conversion is poor, which is not suitable for discourse conversion task. Because BiLSTM can effectively make use of the dynamic interframe correlation of utterances, the transformed utterances have stronger continuity and better comprehensibility, thus achieving a better subjective score. The method in this paper can effectively use the acoustic model information to establish a long-term correlation model of utterances and use the prosodic features of utterances as the learning objective of the model. Therefore, the naturalness of statements transformed by the method in this paper is high, and the opinions are emotional, thus achieving the highest subjective score.

In this paper, ABX test is used to further evaluate and compare BiLSTM discourse conversion method with this method, which has better subjective score. Figure 6 shows the results of the ABX test method. After several rounds of listening tests, the auditioners generally believe that the transformed utterances in this paper are closer to the target utterances.

4. Conclusion

This paper mainly introduces the discourse conversion method based on convolution recurrent neural network with continuous wavelet transform. Compared with the existing statistical model-based discourse conversion methods, the following conclusions can be drawn:

(1) The existing discourse transformation methods usually only consider the differences between discourse spectra and rarely consider the characteristics of discourse itself from the perspective of the internal characteristics of discourse. This paper uses the local connection feature of CNN network to effectively extract the local features of discourse.

(2) Discourse signals have long-term correlation, and existing discourse conversion methods are limited by model structure, so it is difficult to model the long-term correlation of discourse. Inspired by the extended convolutional neural network in the task of discourse synthesis, the method in this paper stacks multiple one-dimensional extended convolutional network layers in the model, so that the feature mapping module of the model can use the whole discourse context information for modeling, so as to describe the long-term relevance of discourse more effectively.

(3) Due to its special motivation source and vocal form, the overall listening sensation of the utterance lacks of tonal change and the naturalness of the listening sensation of the utterance is poor. The converted utterances have better listening comfort, and continuous wavelet transforms are used to decompose the fundamental frequency features instead of the original declarations when training the model. The decomposed fundamental frequency can represent the prosodic characteristics of utterances. Taking the decomposed essential frequency component as the training target can give the transformed speech a better subjective hearing evaluation. At the end of this paper, a number of experimental results show that the discourse conversion method proposed in this paper has better discourse conversion performance compared with the contrast method, and the transformed discourse has better performance in both subjective and objective evaluation.

Data Availability

The data set can be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] X. L. Li, *New Discourse Analysis of Ideological and Political Education in Colleges and Universities in Micro Era and Research on the Frontier Issues of Development*, Xinhua Publishing House, Beijing, 2017.

[2] X. Shi, *Discourse Research in Contemporary China*, Higher Education Press, Beijing, China, vol. 8, 2018.

[3] D. S. Park, W. Chan, Y. Zhang et al., "Specaugment: a simple data augmentation method for automatic speech recognition," 2019, https://arxiv.org/abs/1904.08779.

[4] Y. Jia, Y. Zhang, R. J. Weiss et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," 2018, https://arxiv.org/abs/1806.04558.

[5] T. Tereza, J. Rusz, V. Jan, B. Serena, S. Alessandro, and T. P. Maria, "Speech disorder and vocal tremor in postural instability/gait difficulty and tremor dominant subtypes of Parkinson's disease," *Jounal of Neural Transmission*, vol. V127, no. 4, pp. 328–339, 2020.

[6] A. Gautam, J. G. Naples, and S. J. Eliades, "Control of speech and voice in cochlear implant patients," *The Laryngoscope*, vol. 129, no. 9, pp. 124–136, 2019.

[7] O. Perrotin and I. V. Mcloughlin, "Glottal flow synthesis for whisper-to-speech conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 889–900, 2020.

[8] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-Sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[9] H. Lian, Y. Hu, W. Yu, J. Zhou, and W. Zheng, "Whisper to normal speech conversion using sequence-to-sequence mapping model with auditory attention," *IEEE Access*, vol. 7, pp. 130495–130504, 2019.

[10] D. J. Cates, M. J. Magnetta, L. J. Smith, and C. A. Rosen, "Novel, anatomically appropriate balloon dilation technique of the glottis to treat posterior glottic stenosis in a 3D-printed model," *The Laryngoscope*, vol. 129, no. 10, pp. 2239–2243, 2019.

[11] B. Msa, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Communication*, vol. 11, pp. 29–43, 2019.

[12] L. Czap, "Automated speech production assessment of hard of hearing children," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 380–389, 2020.

[13] J. Zhou, Y. Hu, H. Lian, H. Wang, L. Tao, and H. K. Kwan, "Multimodal voice conversion under adverse environment using a deep convolutional neural network," *IEEE Access*, vol. 7, pp. 170878–170887, 2019.

[14] P. Olivier and V. M. Lan, "Glottal flow synthesis for whisper-to-speech conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 889–900, 2020.

[15] J. B. Salyers, Y. Dong, and Gai, "Continuous wavelet transform for decoding finger movements from single-channel EEG," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1588–1597, 2019.

[16] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform F0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, 2019.

[17] C. Heejin and H. Minsoo, "Sequence-to-Sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42674–42687, 2021.

[18] S. Vekkot, D. Gupta, M. Zakariah, and Y. A. Alotaibi, "Emotional voice conversion using a hybrid framework with speaker-adaptive DNN and particle-swarm-optimized neural network," *IEEE Access*, vol. 8, pp. 74627–74647, 2020.

[19] K. Tan, J. Chen, and Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 189–198, 2019.