Hindawi
Complexity
Volume 2024, Article ID 9781498, 1 page
https://doi.org/10.1155/2024/9781498

Retraction

Retracted: Articulatory-to-Acoustic Conversion Using BiLSTM-CNN Word-Attention-Based Method

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] G. Ren, G. Shao, and J. Fu, "Articulatory-to-Acoustic Conversion Using BiLSTM-CNN Word-Attention-Based Method," *Complexity*, vol. 2020, Article ID 4356981, 10 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 4356981, 10 pages
https://doi.org/10.1155/2020/4356981

Research Article

Articulatory-to-Acoustic Conversion Using BiLSTM-CNN Word-Attention-Based Method

Guofeng Ren, Guicheng Shao, and Jianmei Fu

Department of Electronics, Xinzhou Teachers University, Xinzhou 034000, China

Correspondence should be addressed to Guofeng Ren; renguofeng926@sina.com

Received 6 July 2020; Revised 27 August 2020; Accepted 12 September 2020; Published 26 September 2020

Academic Editor: Zhihan Lv

In the recent years, along with the development of artificial intelligence (AI) and man-machine interaction technology, speech recognition and production have been asked to adapt to the rapid development of AI and man-machine technology, which need to improve recognition accuracy through adding novel features, fusing the feature, and improving recognition methods. Aiming at developing novel recognition feature and application to speech recognition, this paper presents a new method for articulatory-to-acoustic conversion. In the study, we have converted articulatory features (i.e., velocities of tongue and motion of lips) into acoustic features (i.e., the second formant and Mel-Cepstra). By considering the graphical representation of the articulators' motion, this study combined Bidirectional Long Short-Term Memory (BiLSTM) with convolution neural network (CNN) and adopted the idea of word attention in Mandarin to extract semantic features. In this paper, we used the electromagnetic articulography (EMA) database designed by Taiyuan University of Technology, which contains ten speakers' 299 disyllables and sentences of Mandarin, and extracted 8-dimensional articulatory features and 1-dimensional semantic feature relying on the word-attention layer; we then trained 200 samples and tested 99 samples for the articulatory-to-acoustic conversion. Finally, Root Mean Square Error (RMSE), Mean Mel-Cepstral Distortion (MMCD), and correlation coefficient have been used to evaluate the conversion effect and for comparison with Gaussian Mixture Model (GMM) and BiLSTM of recurrent neural network (BiLSTM-RNN). The results illustrated that the MMCD of Mel-Frequency Cepstrum Coefficient (MFCC) was 1.467 dB, and the RMSE of F2 was 22.10 Hz. The research results of this study can be used in the features fusion and speech recognition to improve the accuracy of recognition.

1. Introduction

Along with the popularity of artificial intelligence, man-machine interaction technology has put forward higher requirements for speech processing technology, and it is hoped that intelligent products, such as computers and mobile phones, will have the ability to communicate harmoniously with human beings and the ability to express emotions. The existing technology of emotional speech processing inevitably took advantage of the human pronunciation mechanism, and then human speech is pronounced successfully by the systematic movements through the muscle's contraction of the vocal organs, such as the tongue, lips, and jaw. This relationship between articulatory and acoustic data has been formed through the accumulation of a great deal of articulatory experience.

Although people have adopted a variety of technologies to collect the motion information of articulators, such as X-ray [1], real-time Magnetic Resonance Imaging (rMRI) [2], Ultrasound [3], EPG [4], and EMA [5], most data acquisition environments were not perfect, and the collected data were of poor natural degree or were easily disturbed by external noise [6]. Among them, due to the EMA technology using sensors placed on the pronunciation organs such as the surface of the lip, contact area is only 3 mm^2; at the same time, the sensors' working theory is simple and with stable performance, which has been widely used in pronunciation organs' trajectory tracking and data collection.

For more than a decade, researchers have been studying the acoustic-to-articulatory inversion. Ouni and Laprie [7] first proposed the codebook method in 2005, which used vector quantization to encode the acoustic vectors of speech

and calculate the minimal Euclidean distance between the acoustic vectors and the articulatory vectors, so as to construct the inversion system. The drawback of this method is that it requires a large amount of data to achieve the accurate conversion effect.

King and Wrench [8] implemented a dynamic system to train EMA data using Kalman filter in 1999. They defined the acoustic and articulatory features of speech as linear relationship based on the physical model of speech production. However, there was no strict linear relationship between the acoustic and articulatory features.

Furthermore, in 2000, Dusan and Deng [9] used an extended Kalman filter to train the acoustic-articulatory data to establish a more realistic inversion relationship. By combining this model with Kalman smoothing filter, the movement trajectory of the articulator would be simulated, and the RMSE between the simulated trajectory and the original trajectory was realized to be 2 mm.

Korin Richmond and Yamagishi [10] used neural network to realize the acoustic-to-articulatory inversion firstly in 2002. They used the data of two subjects in MOCHA-TIMIT and achieved the inversion result with a RMSE as low as 1.40 mm. At the same time, Toda et al. [11] proposed a feature inversion method based on Gaussian Mixture Model (GMM), which used maximum likelihood estimation method to analyze the parallel acoustic data stream and the EMA data stream, and established the joint probability density function. Different quantities of Gaussian mixture elements had been used to achieve higher inversion accuracy.

Hiroya and Honda [12], Lin et al. [13], and Ling et al. [14] successively used and improved HMM and finally achieved the integrated RMS of 1.076 mm, which is also the highest inversion accuracy achieved by using HMM model so far.

In recent years, deep learning has attracted great attention for its ability to model nonlinear mapping relations and has been applied to the inversion of articulatory and acoustic features. Leonardo Badino et al. [15, 16] realized acoustic-to-articulatory inversion using Deep Belief Network (DBN) and Hidden Markov Model (HMM) and applied it to speech recognition, resulting in a 16.6% reduction in the recognition relative bit error rate. At the early stage, convolutional neural network (CNN) [17] has been widely used in the field of image signal processing, which had obvious advantages in the analysis of local features; meanwhile the articulatory features could be seen as the visual features of speech. Sun et al. [18] from Yunnan University showed that CNN could be applied to the emotion classification of speech and achieved good results. They were the first to introduce word-attention mechanism to emotion classification and reveal the influence of semantics on classification effect.

However, most researchers only focus on the acoustic-to-articulatory inversion, and the research on the articulatory-to-acoustic conversion is less and started relatively late. Yet the articulatory-to-acoustic conversion is helpful to the study of pronunciation mechanism and the development of speaker recognition and emotion recognition. Liu et al. [19, 20] of the University of Science and Technology of China used Cascade Resonance Network and BiLSTM-RNN to convert articulatory features into spectral energy and fundamental frequency features in 2016 and 2018, respectively, and achieved a good conversion effect.

At present, the conversion focuses on the frame or phoneme level, with emphasis on the pronunciation rules and acoustic characteristics of phonemes. However, in the tonal languages like Mandarin, the interaction between syllables must hide certain acoustic-pronunciation information. Meanwhile, word-attention mechanism has been widely applied in the field of text processing and emotion classification. Wang and Chen [21] proposed an LSTM emotion classification method based on attention mechanism and realized emotion classification through feature screening of short- and long-text features combined with attention mechanism. Wang et al. [22] proposed a word-attention convolution model with the combination of CNN and attention mechanism, aiming at word feature extraction.

Relying on deep learning with nonlinear and attention mechanism, BiLSTM-CNN method and word-attention mechanisms were used to realize the articulatory-to-acoustic conversion in this paper. The paper is organized as follows. First, we review related work on articulatory-to-acoustic conversion, as well as CNN and word-attention mechanism in Section 2. Next, the detailed method we proposed is described in Section 3, and Section 4 reports our experiments and their results. Section 5 provides the discussion and conclusion of the work.

2. Related Work

To explore the articulatory-to-acoustic conversion and improve the conversion effect, lots of researches have been carried out in the past decades, and several methods have been proposed to model the conversion, including Gaussian Mixture Model (GMM), recurrent neural network (RNN), Long Short-Term Memory (LSTM), BiLSTM, and CNN. We will give a brief introduction in this section.

2.1. GMM-Based Articulatory-to-Acoustic Conversion. GMM is a classical feature conversion method [23], which used the joint probability density function of acoustic-articulatory features to realize the conversion. The description of the transformation model is

$$y_t = \sum_{i=1}^{M} p\left(\lambda_i \mid t_i\right) p\left(y_i \mid x_i, \lambda_i\right), \tag{1}$$

Here, M is used to represent the number of Gaussian mixture elements, $p(\lambda_i \mid x_i)$ denotes the probability of acoustic feature vector x_i, and $p(y_i \mid x_i, \lambda_i)$ represents full covariance matrix of conditional Gaussian distribution.

$x = (x_1, x_2, \ldots, x_M)$ and $y = (y_1, y_2, \ldots, y_M)$ have been defined as articulatory and acoustic features, respectively, where, M is the number of frames. Considering that the articulatory features of frame i are known, the first-order dynamic features are as follows:

$$\Delta x_i = -0.5x_{i-1} + 0.5x_{i+1}. \tag{2}$$

The articulatory features and the first-order dynamic features are spliced as the input feature vector $X_i = [x_i^M, \Delta x_i^M]^M$, and then output vector $Y_i = [y_i^M, \Delta y_i^M]^M$ can be obtained. Thus, the joint probability distribution of input and output vectors can be described as follows:

$$\begin{aligned} P\left(Z_i \mid \vartheta\right) = P\left(X_i, Y_i \mid \Theta\right) &= \sum_{j=1}^{N} \partial_j \mathbb{N}\left(Z_i, \mu_j^2, \sum_j^2\right), \\ Z_i &= \begin{bmatrix} X_i \\ Y_i \end{bmatrix}, \\ \mu_j^Z &= \begin{bmatrix} \mu_j^X \\ \mu_j^Y \end{bmatrix}, \\ \sum_j^Z &= \begin{bmatrix} \sum_j^{XX} & \sum_j^{XY} \\ \sum_j^{YX} & \sum_j^{YY} \end{bmatrix}, \end{aligned} \tag{3}$$

where $Z_i = [X_i^M, Y_i^M]$ is the joint vector of articulatory and acoustic features, N is the number of Gaussian elements, $\theta = \alpha_j, \mu_j^Z, \Sigma_j^Z \mid j = 1, 2, \ldots, N$, denotes the model parameters of GMM, and α_j, μ_j^Z, and Σ_j^Z are weight, mean, and covariance of Gaussian element j, respectively. Among them, model parameter Θ will be estimated by Maximum Likelihood Estimation Algorithm (MLEA) [24]. When the dimension between articulatory and acoustic features is different, the covariance matrix Σ_j^Z is full-rank matrix.

During the conversion, input articulatory features are supposed to be $X = [x_1, x_2, \ldots, x_M]$, and output acoustic features are supposed to be $Y = [y_1, y_2, \ldots, y_M]$; y^* can be calculated relying on the MLE as follows:

$$y^* = \arg\max_y p(Y \mid X, \Theta), \quad \text{s.t. } Y = Wy, \tag{4}$$

Here, W is dynamic window coefficient matrix. In formula (4), conditional probability distribution can be rewritten as follows:

$$p(Y \mid X, \Theta) = \sum_{j=1}^{N} p(j \mid X, \Theta) p(Y \mid j, Y, \Theta). \tag{5}$$

If we only refer to a Gaussian element, it can be calculated by Maximum Posterior Probability, which is shown as follows:

$$j^* = \arg\max p(j \mid X, \Theta). \tag{6}$$

If the frames are independent of each other, for input of frame i, X_i exist as Formula (7); meanwhile, the output of frame i, Y_i exist as Formula (8):

$$p\left(j \mid X_i, \Theta\right) = \frac{p(j \mid \Theta) p\left(X_i \mid j, \Theta\right)}{p\left(X_i \mid \Theta\right)} = \frac{\alpha_j \mathbb{N}\left(X_i \mid \mu_j^X, \sum_j^{XX}\right)}{\sum_n^N \alpha_n \mathbb{N}\left(X_i \mid \mu_n^X, \sum_n^{XX}\right)}, \tag{7}$$

$$p\left(Y_i \mid j^*, X_i, \Theta\right) = \mathbb{N}\left(Y_i \mid \mu_{j^*,i}^{y \mid x}, \sum_{j^*}^{y \mid x}\right). \tag{8}$$

Here, $\mu_{j^*,i}^{y|x}$ and $\Sigma_{j^*}^{y|x}$ are mean and covariance matrix, respectively, which are calculated using the following two formulas:

$$\mu_{j^*}^{y|x} = \mu_{j^*}^Y + \sum_{j^*}^{YX} \sum_{j^*}^{YX^{-1}} \left(X_i - \mu_{j^*}^X\right), \tag{9}$$

$$\sum_{j^*}^{y|x} = \sum_{j^*}^{YY} - \sum_{j^*}^{YX} \sum_{j^*}^{XX} \sum_{j^*}^{XY}. \tag{10}$$

On this basis, we can obtain the output sequence using maximum likelihood criterion, as shown in formula (11), where Σ^y is square matrix and μ^y can be calculated through $\mu_{j^*,t}^{y|x}$ connecting nose to tail:

$$y^* = \left(W^T \sum^{y-1} W\right)^{-1} W^T \sum^{y-1} \mu^y. \tag{11}$$

2.2. LSTM of RNN. Recurrent neural network (RNN) is a kind of neural network that takes sequence data as input data and recourses along the time domain direction of the sequence [20]. All nodes in this network are connected in a chain. RNN has the advantages of memorability, parameter sharing, and Turing completeness and is obviously superior to GMM in learning nonlinear features. The network has been widely used in speech recognition, speech modeling, feature conversion, and other fields.

The core of RNN is the directed graph, and the loop unit is fully connected. Input sequence is given as $X = \{X_1, X_2, \ldots, X_M\}$, and spread length is given as τ. For time-step t, the recurrent unit should be taken as

$$h^{\langle t \rangle} = f\left(s^{\langle t-1 \rangle}, X^{\langle t \rangle}, \theta\right), \tag{12}$$

where h denotes systematic state of RNN, s denotes inner state calculated by $s = s(h, X, y)$, and f represents activation function, such as logistic and hyperbolic tangent function, or represents a kind of feedforward neural network. The excitation function corresponds to the simple recurrent network, and the feedforward neural network corresponds to some depth algorithms. θ is weight coefficient in the recurrent unit.

We take the example of an RNN containing a hidden layer; the hidden layer vector sequence $H = \{h_1, h_2, \ldots, h_M\}$ can be obtained by

$$h^{\langle t \rangle} = f\left(W_{hh} h^{\langle t-1 \rangle} + W_{xh} X^{\langle t \rangle} + \theta_h\right). \tag{13}$$

Then, output sequence $Y = \{Y_1, Y_2, \ldots, Y_M\}$ can be shown as follows:

$$Y^{\langle t \rangle} = W_{hy} h^{\langle t \rangle} + \theta_y. \tag{14}$$

Initially, inverse error transfer algorithm on the time axis has been adopted to update the parameters, which would produce some inverse transfer error. So gradient erasing and explosion would occur, which seriously affected the training effect of RNN. In order to reduce the above problems, Li et al. [25] put forward Long Short-Term Memory (LSTM), including nonlinear transform and gate-structure affection function. Through the development of LSTM, the structure brought forward by Aviles and Kouki [26] is consisting of

input gate, output gate and forgetting gate. Among them, input gate is used to control the conversion processing from accepted information to memory sequence, which is shown as follows:

$$i^{\langle t\rangle} = \sigma\left(W_{ix}X^{\langle t\rangle} + W_{ih}h^{\langle t-1\rangle} + W_{ic}c^{\langle t\rangle} + \theta_i\right), \quad (15)$$

Here, σ is sigmoid function and c is memory sequence. Forgetting gate is used to control how much of the current memory information should be discarded, the implementation method of which is

$$f^{\langle t\rangle} = \sigma\left(W_{fx}X^{\langle t\rangle} + W_{fh}h^{\langle t-1\rangle} + W_{fc}c^{\langle t\rangle} + \theta_f\right). \quad (16)$$

The memory sequence can be updated as follows relying on input and output gates:

$$c^{\langle t\rangle} = i_t * \tan h\left(W_{cx}X^{\langle t\rangle} + W_{ch}h^{\langle t-1\rangle} + \theta_c\right) + f^{\langle t\rangle} * c^{\langle t-1\rangle}. \quad (17)$$

The output gate can be used to scale output sequence, and the detailed method is as follows:

$$o^{\langle t\rangle} = \sigma\left(W_{ox}X^{\langle t\rangle} + W_{oh}h^{\langle t-1\rangle} + W_{oc}c^{\langle t\rangle} + \theta_o\right). \quad (18)$$

Finally, we can obtain

$$h^{\langle t\rangle} = o^{\langle t\rangle} * \tan h\left(c^{\langle t\rangle}\right), \quad (19)$$

and the result can be transferred into RNN.

2.3. BiLSTM. Bidirectional Long Short-Term Memory (BiLSTM) [18] is a variant of traditional neural network and combination of forward LSTM and backward LSTM. Output of the model can be represented as

$$\begin{aligned} \overrightarrow{h}^{\langle t\rangle} &= \overrightarrow{\mathrm{LSTM}}\left(h^{\langle t-1\rangle}, X^t, c^{\langle t-1\rangle}\right), \\ \overleftarrow{h}^{\langle t\rangle} &= \overleftarrow{\mathrm{LSTM}}\left(h^{\langle t+1\rangle}, X^t, c^{\langle t+1\rangle}\right), \\ h^{\langle t\rangle} &= \left[\overrightarrow{h}^{\langle t\rangle}, \overleftarrow{h}^{\langle t\rangle}\right]. \end{aligned} \quad (20)$$

Let us take the mean of $h^{\langle t\rangle}$ as the output; that is to say, the output is mean$(h^{\langle t\rangle})$. Until the long-short sequence has arrived at BiLSTM layer, gate structure began to carry adoption and releasing of the information through sigmoid, and the output is between 0 and 1 (1 means complete adoption, and 0 means complete discarding). The ideal structure of BiLSTM is shown in Figure 1.

2.4. CNN. Convolutional neural network (CNN) [18] is a feedforward neural network containing convolution operation, and its model structure generally includes input layer, convolution layer, pooling layer, full-connection layer, and output layer. The convolution layer, pooling layer, and full-connection layer can all be seen as hidden layers. Among them, the role of the convolution layer is to carry out feature extraction, and the feature extraction of input layer data can be realized by using the set filter. The specific method is shown as follows:

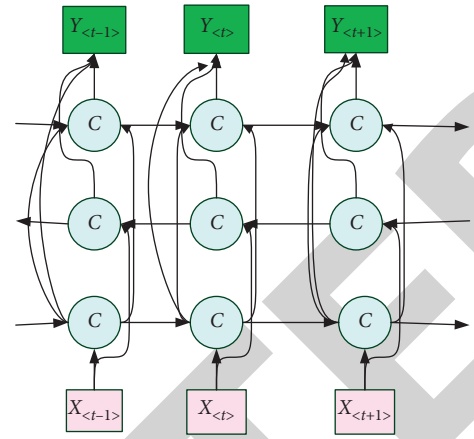

Figure 1: Structure of BiLSTM.

$$J_i = f\left(\omega * X_{i:\, i+g-1} + \theta\right). \quad (21)$$

Here, ω denotes convolution kernel, g denotes the size of convolution kernel, $X_{i:\, i+g-1}$ denotes articulatory feature vector from frame i to frame $i + g - 1$, and θ denotes bias value. Thus, we can obtain the feature matrix $J = [c_1, c_2, \ldots, c_{n-g+1}]$ through the convolution layer calculation.

Using max pooling technology, pooling layer can downsample the feature matrix and achieve optimal solution of the local value. Full-connection layer is located in the last layer of the hidden layer and can expand the feature diagram with topological structure to activate function. Output layer uses logical function or Softmax function to output classified label and predicted value.

3. Methods

3.1. Speaker Normalization Based on Prussian Transformation. Because speakers' articulatory characteristics are easily influenced by the speakers themselves, including their vocal tract characteristics, height, and sitting position; these factors are inherent differences between speakers. In order to eliminate these inherent differences and better quantify the kinematic characteristics of speech, we used the Prussian Transformation to normalize the articulatory characteristics of different speakers. The specific processing is shown in Figure 2.

The algorithm realized the linear geometric transformation from the original multipoint object to the target multipoint object, including scale transformation, translation transformation, and rotation transformation. It is supposed that the raw articulatory data was D_1; then the normalization of D_1 was D_3, and the target speaker's articulatory data was D_2. Using hybrid transform consisting of scale transform and rotation transform, we can take the relation between D_1 and D_3 as follows:

$$D_3 = HD_1a + b, \quad (22)$$

where the normalizing parameter $\{H, a, b\}$ can be optimized relying on minimized Root Mean Square Error between target data D_2 and the normalized data of raw speaker's articulatory D_3.

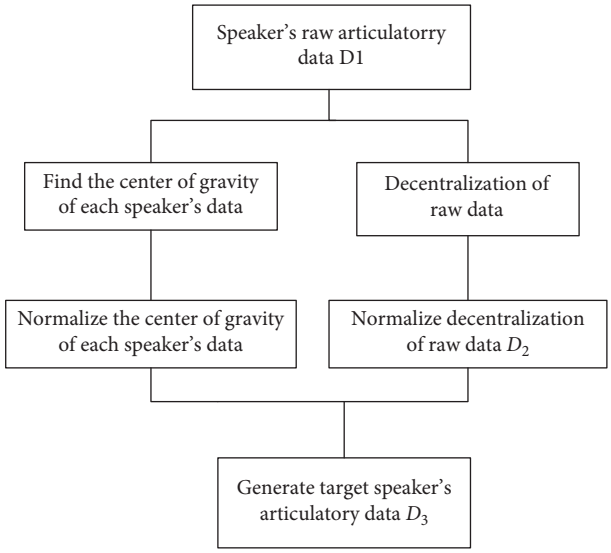

Figure 2: Diagram of speaker normalization algorithm based on Prussian Transformation.

To be specific, rotation matrix can be calculated using singular value decomposition, which is shown as follows:

$$(D_1)^T D_2 = U \sum V^T, \tag{23}$$

Here, Σ is the diagonal matrix, U and V are separate orthogonal matrices, and A is the diagonal matrix in which the absolute value of the diagonal elements is 1.

3.2. BiLSTM-CNN-Based Articulatory-to-Acoustic Conversion. According to Sections 2.2 and 2.3, CNN has a good ability to extract local features, and BiLSTM network has a good performance on the coherence of previous frames and semantic features based on word-attention mechanism [27]. This paper combined CNN and BiLSTM and used the theory of word attention to achieve articulatory-to-acoustic conversion, where BiLSTM used context information to analyze the articulatory features and train continuous frames, and word-attention layer used word-attention mechanism to extract semantic features and send them to the BiLSTM for training. In the later stage, the CNN was mainly composed of convolutional layer, pooling layer, and full-connection layer. Finally, acoustic features are output by regression layer. The specific model structure is shown in Figure 3.

As illustrated in Figure 3, the LSTM cells at each layer in the BiLSTM-CNN were divided into two parts to capture the forward and backward dependency, respectively. In this case, the forward and backward articulation feature sequences were both 10 frames and the feature vector of each frame was 8 dimensions, and the semantic feature was 1 dimension. Thus, the input feature dimension of the feature fusion layer was 169 dimensions. In the CNN part, we used 4 full-connection layers, the convolution layer with size of 128 dimensions, and the regression layer.

4. Experiments and Results

4.1. Materials

4.1.1. Participants. In the study, ten participants (5 males and 5 females) were recruited; all of them were aged between 25 and 40 years (average of ages is 27.1, and the STD is 1.94) with no professional language training and no orofacial surgery history [28]. Before collecting the data, all subjects were told the processes for collecting data and signed informed consent. The study was approved by the Health Sciences Research Ethics Board at Institute of Psychology of the Chinese Academy of Sciences (No. H16012).

4.1.2. Textual Material. Disyllable words and sentences of neural affect were chosen as textual material. Sentences of neural textual material were chosen as the spoken material, including "Xia4 yu3 le1." (It is raining.), "Jin1 tian1 shi4 xing1 qi1 yi1." (Today is Monday.), "Wo3 xiang3 gei3 ta1 yi2 ge4 jing1 xi3." (I want to give him a surprise.), "Ni3 yuan2 lai2 zai4 zhe4 li3" (So you here.), "Wo3 cuo4 le1." (I am wrong.), "Ni3 xue2 de1 zhen1 kuai4." (You learn fast.), and "Wo3 men2 shi4 lao3 tong2 xue2." (We are old classmates.).

Disyllable words were chosen as the spoken material, including "Mama" (Mum), "Zaijian" (Good-bye), "Tiantian" (Everyday), "Daqi" (Encourage), and "Nihao" (Hello).

4.1.3. Data Collection. All articulatory data and acoustic data were collected using the AG501 [29] EMA device of Carstens [29] (Lenglern, Germany) as shown in Figure 4, which has 24 articulatory channels and one audio channel with 250 Hz and 48 kHz sampling rate. AG501 is widely used in electromagnetic articulography, which allows the collecting in 3D of the movements of the articulators with high precision.

We have glued 6 sensors (2 mm ∗ 3 mm) with thin wires to the left and right mastoids, nose bridge, and the bite plane to carry head collection and 9 sensors to the upper and lower lips, left and right lip corners, upper and lower incisors, and tongue tip, tongue mid, and tongue root (as shown in Figure 5). All subjects engaged in conversation for approximately 5 minutes after sensors were attached to provide subjects the opportunity to familiarize themselves with the presence of the sensors in the oral cavity.

The collection experiment has been carried out in the quiet environment with a maximum background noise of 50 dB. Acoustic data was collected by a match condenser microphone EM9600, and articulatory data was collected in synchronization with the acoustic data.

4.1.4. Data Processing and Feature Extraction. The collected data were loaded into the VisArtico, a visualization tool for filtering using a low-pass filter (cut-off is 20 Hz). Meanwhile, the articulatory data were corrected for head movement using Cs5normpos tool, which is a kind of tool in the EMA control system of AG501.

The VisArtico program can visualize kinematic data while also allowing for calculation of tongue kinematic

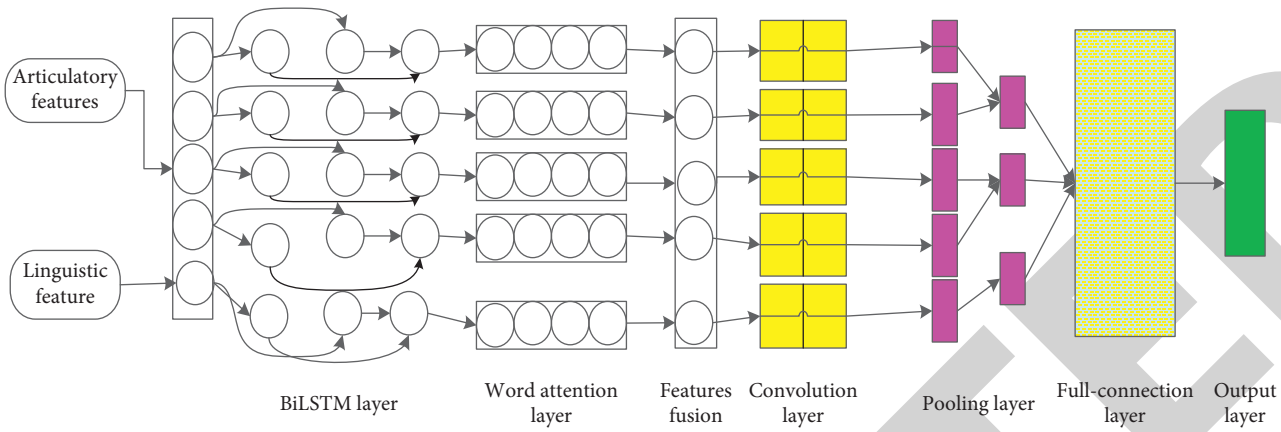

Figure 3: Integration model of BiLSTM-CNN.

Figure 4: Sensor placements on occlusal bite plane.

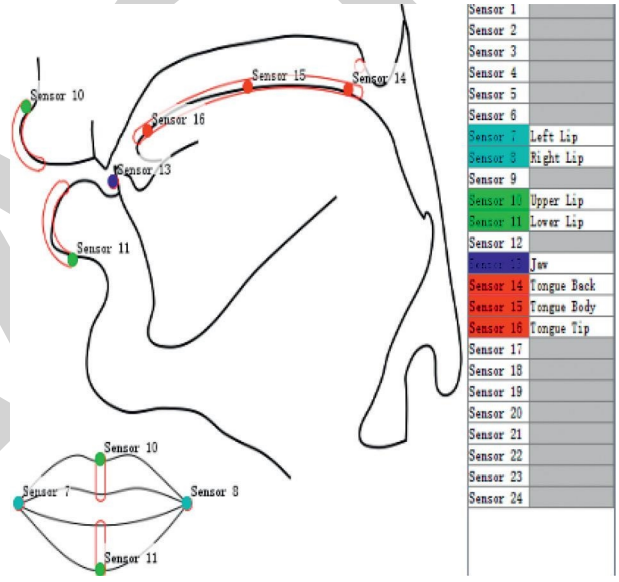

Figure 5: Sensor placements on the articulator.

parameters. In this paper, we extracted 8-dimensional articulatory features as shown in Table 1.

In this paper, we have chosen 299 samples of disyllables and sentences and then took 200 samples as the training data and 99 samples as the test data, respectively.

4.2. Model Comparison of EMA-to-F2 Conversion. In the EMA-to-F2 experiment, we compared the performances of the GMM-based, RNN-based, and BiLSTM-CNN-based methods. The Root Mean Square Error (RMSE) in Hz between the true and the predicted F2 was adopted as the evaluation measure parameter.

As a classical prediction model, GMM can approximate any function as long as the number of mixing elements is sufficient. In this study, we selected GMM with 500 Gaussian elements to accurately describe the joint probability density function of the articulatory features and acoustic features. According to the maximum likelihood criterion, the conditional probability of acoustic features has been obtained by approximate calculation of the joint probability density function of acoustic features and articulatory features, and the closed solution of the best acoustic features has been obtained. The result is shown in Figure 6 (the figure takes 80 frames of data as the example).

For the EMA-to-F2 conversion based on BiLSTM-RNN, the 21-frame input window (10 frames forward and 10 frames backward) has been used to train the network. We have trained 50 iterations for BiLSTM-RNN with 5 hidden layers and 100 hidden units per hidden layer. The training results are shown in Figure 7, which illustrated the RMSE and loss of training data. Along with increasing the iterations number, the RMSE between the true and predicted data and loss function values decreased. The optical model occurred at the 48th epoch, and the loss function value and RMSE reached their minimum, respectively.

The BiLSTM-CNN we proposed consisted of BiLSTM, word-attention layer, and the CNN (convolutional layer,

Table 1: Articulatory features.

Abbreviation for features	Description of features
Tongue root-X	The normalized back-forward velocity of the tongue root
Tongue root -Z	The normalized up-down velocity of the tongue root
Tongue body-X	The normalized back-forward velocity of the tongue mid
Tongue body-Z	The normalized up-down velocity of the tongue mid
Tongue tip-X	The normalized back-forward velocity of the tongue tip
Tongue tip-Z	The normalized up-down velocity of the tongue tip
Tongue constriction	The kinematics range of tongue tip at back-forward direction
Lip aperture	The open degree of lips at up-down direction

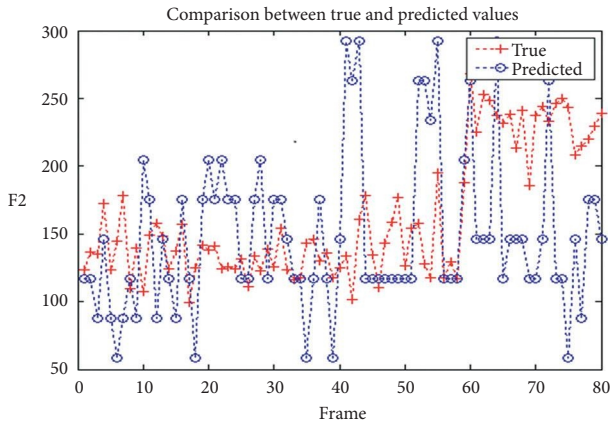

Figure 6: Comparison between true and conversed values of GMM-based F2.

pooling layer, full-connection layer, and regression layer). About the CNN part, we have chosen the convolutional layer with size of $169 * 169$, 4 full-connection layers, and the 1-dimensional regression layer. About the BiLSTM part, we took 5 hidden layers with 100 hidden units per hidden layer and adopted 21 frames (10 frames forward, 1 current frame, and 10 frames backward) as the input feature; meanwhile, the semantic feature needs input to the BiLSTM for feature fusion and training. In the training process, we initially set the learning rate to 0.005 and fixed the momentum at 0.8, with maximum epochs of 50. Then, we can find that BiLSTM-CNN is much better than the BiLSTM-RNN and GMM conversion model, and the comparisons of F2 between true value and the predicted values, using the GMM, BiLSTM-RNN, and BiLSTM-CNN based on word attention, all are shown in Figure 8.

From the figure, we can find that the predicted F2 using BiLSTM-CNN is most similar to the true value, and the predicted F2 using BiLSTM-RNN is less similar to BiLSTM-CNN. Furthermore, we used test data on GMM, BiLSTM-RNN, and BiLSTM-CNN based on word attention; the RMSE and correlation coefficient r of F2 can be obtained and shown in Table 2.

The correlation coefficient r has been used to analyze the correlation between the predicted features and the true features using Pearson product moment correlation method, which is a method to analyze the linear relationship between two variables. Here, it is supposed that there are two databases: articulatory feature input ($x = \{x_1, x_2, \ldots, x_n\}$) and acoustic features output ($y = \{y_1, y_2, \ldots, y_n\}$), and the size of the database is n. Thus, Pearson correlation coefficient can be defined as

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}, \quad (24)$$

where $\overline{x}$ and $\overline{y}$ represent the means of sample features x and y and x_i and y_i show ith values of $\overline{x}$ and $\overline{y}$, respectively. Correlation coefficient r can reflect the strength information of the linear relationship between variable sets x and y, ranging from -1 to 1. If x_i and y_i are multidimensional vector, the dimensionality of the vector should be reduced first, and then the correlation analysis should be carried out.

In the study, we can find that there are strong positive correlations between the predicted and true features on all three models, which is shown in Table 1. In detail, the correlation was, in turn, BiLSTM-CNN > BiLSTM-RNN > GMM.

4.3. Model Comparison of EMA-to-MFCC Conversion. In the EMA-to-MFCC experiment, we adopted MMCD as the parameter to evaluate the results of articulatory-to-MFCC conversion, which can be defined as the mean value of Euclidean distance between the predicted value and true value. Here, we used 12-dimensional MFCC as the acoustic feature and compared the performances of the GMM-based, RNN-based, and BiLSTM-CNN-based methods.

In the experiment, we selected GMM with 500 Gaussian elements to accurately describe the joint probability density function of the articulatory features and acoustic features. For the BiLSTM-CNN, we set the convolutional layer with size of $169 * 169$, 4 full-connection layers, the 1-dimensional regression layer, and 5 hidden layers with 100 hidden units per hidden layer and adopted 21 frames (10 frames forward, 1 current frame, and 10 frames backward) as the input feature.

In the training process, we initially set the learning rate to 0.005 and fixed the momentum at 0.9, with the maximum epochs of 60. Then, we can find that BiLSTM-CNN is much better than the BiLSTM-RNN and GMM conversion model, and the comparison results are shown in Table 3.

From the table, the MMCD of BiLSTM-CNN is the minimum among three models, and BiLSTM-RNN is better than GMM but not better than BiLSTM-CNN. Meanwhile, we can find that there are strong positive correlations

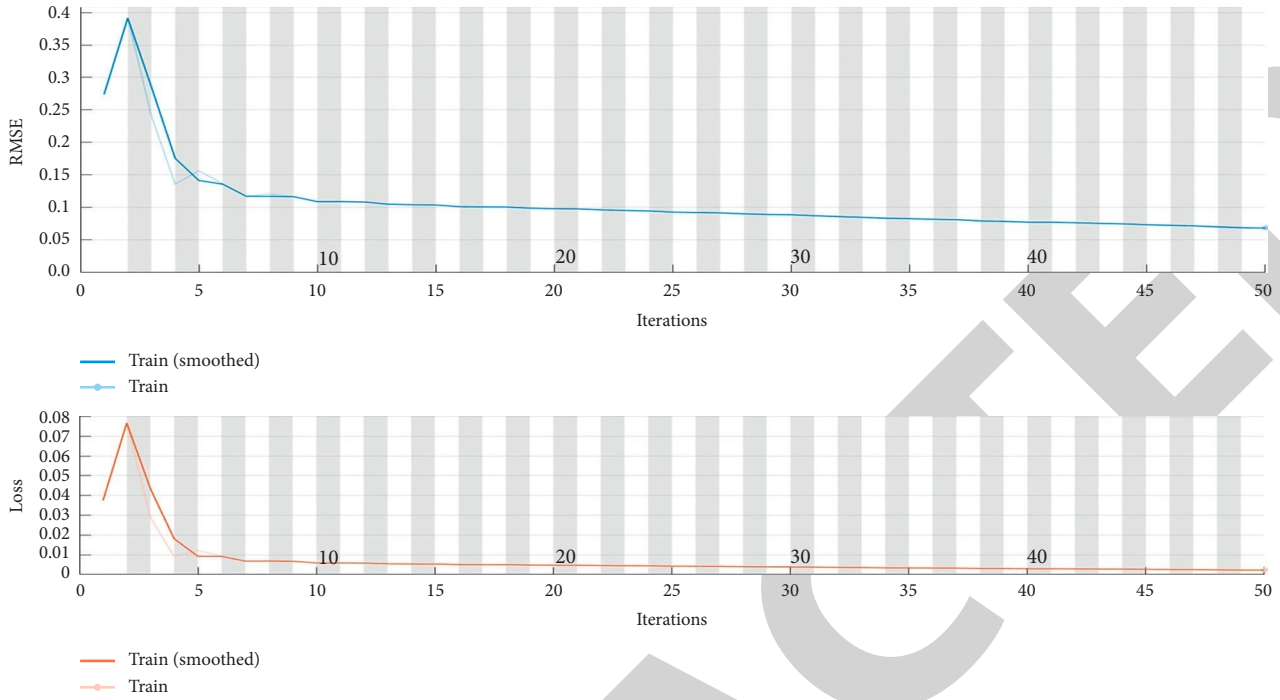

Figure 7: The RMSE and loss function values on the training database of BiLSTM-RNN.

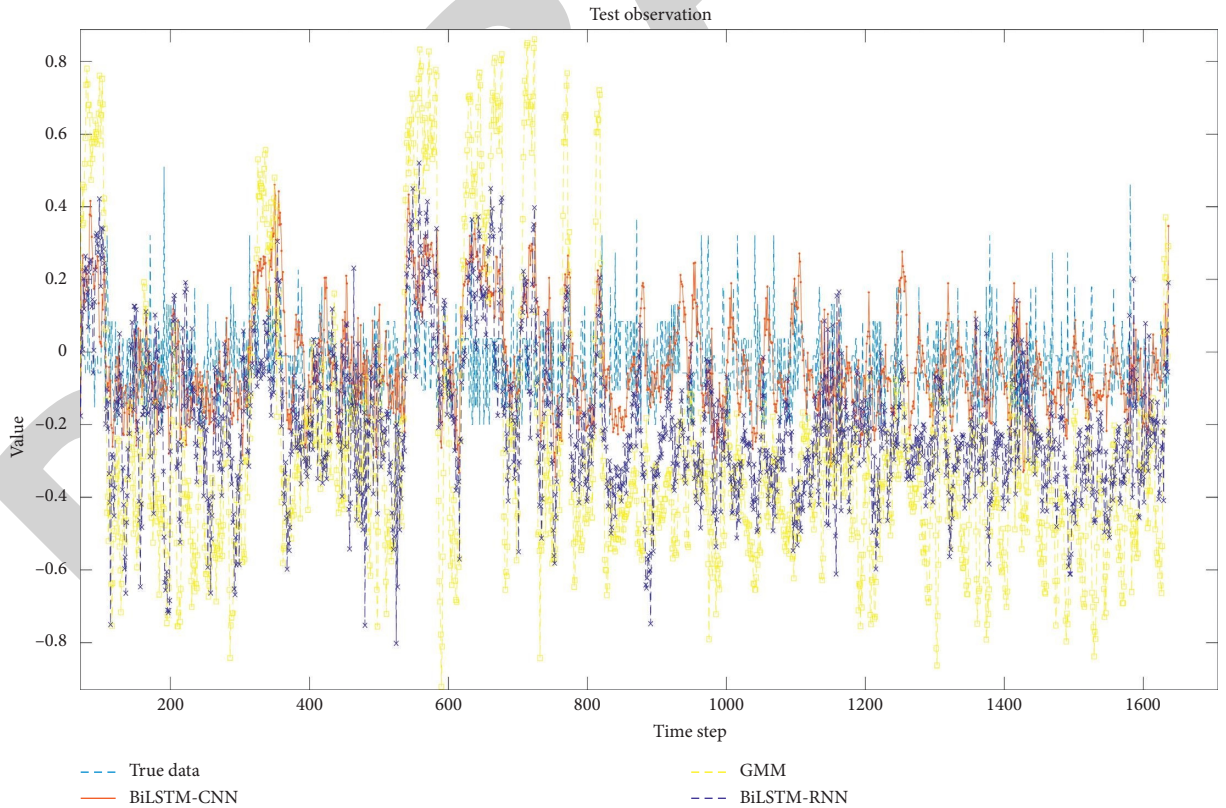

Figure 8: Kinematics comparison of F2 based on different network.

Table 2: The RMSE (Hz) and r on the test set when using GMM, BiLSTM-RNN, and BiLSTM-CNN for EMA-to-F2 conversion.

	GMM	BiLSTM-RNN	BiLSTM-CNN
RMSE	37.13	26.42	22.10
r	0.565	0.625	0.738

Table 3: The MMCD (dB) and r on the test set when using GMM, BiLSTM-RNN, and BiLSTM-CNN for EMA-to-MFCC conversion.

	GMM	BiLSTM-RNN	BiLSTM-CNN
MMCD	2.384	1.824	1.467
r	0.576	0.725	0.753

between the predicted and true features on all three models; in detail, the correlations are, in turn, BiLSTM-CNN > BiLSTM-RNN > GMM.

5. Discussion and Conclusion

This study provided a novel conversion method combining BiLSTM, CNN, and word-attention theory. In the current study, features of the tongue and the lip in 3D coordinates of AG501 have been extracted for the conversion and recognition research and acoustic features (i.e., F2 and MFCC).

From the conversion research, we found that the kinematics of tongue and lips can construct a simple graph, which are found from the application of CNN, because CNN has been used to graph signal processing widely. Meanwhile, because the database we used is Mandarin, as a kind of tone language, semantic feature plays an important role in the speech processing, especially in articulatory-to-acoustic conversion and speech recognition. So, we adopted word-attention theory in this study and achieved ideal effect, which proves that the semantic feature is helpful to the conversion study especially in Mandarin.

The current study broke the limitation of focusing on the vowel only and fused the semantic features and articulatory features. Due to the limitation of numbers of samples, we choose 299 disyllables only in this paper; the sample size was a little bit small, which will be considered in future efforts. The study in this paper should be the basement of the research of speech recognition and speech production. It can promote the fusion of artificial intelligence and Smart Campus in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Thanks are due to all the subjects in current experiment, to Guicheng Shao for technical assistance, to Jianmei Fu for modal design, and to Jianzheng and Dong Li for assistance in data collection. This work was supported by the Educational Reform Innovation Project of Shanxi Province of China (J2019174), Science and Technology Project of Xinzhou Teachers University (2018KY15), and Academic Leader Project of Xinzhou Teachers University.

References

[1] R. John, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, no. S1, p. 56, 1990.

[2] SAIL, "MRI-TIMIT: a multimodal real-time MRI articulatory corpus," 2014, https://sail.usc.edu/span/mri-timit/ed.

[3] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-based ultrasound-to-speech conversion for a silent speech interface," in *Proceedings of the 2017 INTERSPEECH*, Stockholm, Sweden, 2017.

[4] Y. Luo, *A Study on the Location and Coarticulation of Consonants Based on EPG—A Case Study of Zhuang and Miao Languages, Master, Experimental Phonetics*, Shang Normal University, Shanghai, China, 2017.

[5] K. Richmond, "Preliminary inversion mapping results with a new EMA corpus," in *Proceedings of the 2009 INTERSPEECH*, Brighton, UK, 2009.

[6] K. Richmond, Z. Ling, J. Yamagishi, and B. Ur, "On the evaluation of inversion mapping performance in the acoustic domain," in *proceeding of the 2013 INTERSPEECH*, Lyon, France, 2013.

[7] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[8] S. King and A. Wrench, "Dynamical system modeling of articulator movement," in *Proceedings of the 1999 ICPHS*, San Francisco, CA, USA, 1999.

[9] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proceedings of the 2000 Seminar on Speech Production*, Sydney, Aystralia, 2000.

[10] Z. L. Korin Richmond and J. Yamagishi, "Benigno UR and IA, on the evaluation of inversion mapping performance in the acoustic domain," in *Proceedings of the 2013 INTERSPEECH*, Lyon, France, 2013.

[11] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *Proceedings of the 2004 INTERSPEECH*, Jeju Island, Republic of Korea, 2004.

[12] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[13] J. Lin, W. Li, Y. Gao et al., "Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks," *Journal of Signal Processing Systems*, vol. 90, pp. 1077–1087, 2018.

[14] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An analysis of HMM-based prediction of articulatory movements," *Speech Communication*, vol. 52, no. 10, pp. 834–846, 2010.

[15] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "Integrating articulatory data in deep neural network-based acoustic modeling," *Computer Speech & Language*, vol. 36, pp. 173–195, 2016.

[16] C. C. Leonardo Badino, L. Fadiga, and G. Metta, "Deep-level acoustic-toarticulatory mapping for DBN-HMM based phone," in *Proceeedings of the 2012 IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, USA, 2012.
[17] J. Bai, F. Li, and H.-D. Ji, "Attention-based BiLSTM CNN Chinese microblog position detection model," *Computer Applications and Software*, vol. 3, no. 35, pp. 266–274, 2018.
[18] K. Sun, "Word attention-based BiLSTM and CNN ensemble for Chinese sentiment analysis," *Computer Science and Application*, vol. 10, no. 2, pp. 312–324, 2020.
[19] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks," in *Proceeedings of the 2016 INTERSPEECH*, San Francisco, CA, USA, 2016.
[20] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation," *Speech Communication*, vol. 99, pp. 161–172, 2018.
[21] Y.-M. Wang and K. Chen, "End-to- end audio-visual dual-mode speech recognition based on SDBN and BLSTM attention fusion," *Communication Science*, vol. 12, pp. 80–90, 2019.
[22] L.-Y. Wang, C.-H. Liu, D.-B. Cai et al., "Text emotion analysis based on CNN-BiLSTM network with attentional model," *Journal of Wuhan University of Technology*, vol. 4, no. 41, pp. 387–394, 2019.
[23] B. An Ji, *MSEE, Speaker Independent Acoustic-to-Articulatory Inversion, Doctor, Electrical and Computer Engineering*, Marquette University, Milwaukee, WI, USA, 2015.
[24] A. Ji, J. J. Berry, and M. T. Johnson, "Vowel production in Mandarin accented English and American English: kinematic and acoustic data from the Marquette University Mandarin accented English corpus," *Speech Communication*, vol. 19, Article ID 060221, 2013.
[25] R. Li, Z. Wu, Y. Huang, J. Jia, H. Meng, and L. Cai, "Emphatic speech generation with conditioned input layer and bidirection LSTMs for expressive speech synthesis," in *Proceedings of the 2018 ICASSP*, Calgary, Canada, 2018.
[26] J. C. Aviles and A. Kouki, "Position-aided mm-wave beam training under NLOS conditions," *IEEE Access*, vol. 4, pp. 8703–8714, 2016.
[27] L. Wu, F. Tian, L. Zhao, J. Lai, and T.-Y. Liu, "Word attention for sequence to sequence text understanding," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 1–8, New Orleans, LA, USA, 2018.
[28] G.-F. Ren, X.-Y. Zhang, D. Li, and etal, "Design and evaluation of Mandarin bi-modal emotion speech database," *Modern Electronic Technology*, vol. 41, no. 14, pp. 182–186, 2018.
[29] M. Stella, A. Stella, F. Sigona, P. Bernardini, M. Grimaldi, and B. G. Fivela, "Electromagnetic articulography with AG500 and AG501," in *Proceedings of the 2013 INTERSPEECH*, Lyon, France, 2013.