IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 23, 2019, accepted September 5, 2019, date of publication September 12, 2019,
date of current version September 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940700

Whisper to Normal Speech Conversion Using Sequence-to-Sequence Mapping Model With Auditory Attention

HAILUN LIAN[1], YUTING HU[1], WEIWEI YU[1], JIAN ZHOU[1], AND WENMING ZHENG[2], (Senior Member, IEEE)

[1]Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230601, China
[2]Key Laboratory of Child Development and Learning Science, Ministry of Education, Southeast University, Nanjing 210096, China

Corresponding author: Jian Zhou (jzhou@ahu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61301295, in part by the Anhui Provincial Natural Science Foundation under Grant 1708085MF151, and in part by the Natural Science Foundation for the Higher Education Institutions of Anhui Province under Grant KJ2018A0018.

ABSTRACT Whispering is a special pronunciation style in which the vocal cords do not vibrate. Compared with voiced speech, whispering is noise-like because of the lack of a fundamental frequency. The energy of whispered speech is approximately 20 dB lower than that of voiced speech. Converting whispering into normal speech is an effective way to improve speech quality and/or intelligibility. In this paper, we propose a whisper-to-normal speech conversion method based on a sequence-to-sequence framework combined with an auditory attention mechanism. The proposed method does not require time aligning before conversion training, which makes it more applicable to real scenarios. In addition, the fundamental frequency is estimated from the mel frequency cepstral coefficients estimated by the proposed sequence-to-sequence framework. The voiced speech converted by the proposed method has appropriate length, which is determined adaptively by the proposed sequence-to-sequence model according to the source whispered speech. Experimental results show that the proposed sequence-to-sequence whisper-to-normal speech conversion method outperforms conventional DTW-based methods.

INDEX TERMS Auditory attention mechanism, sequence-to-sequence, speech quality, whisper conversion.

I. INTRODUCTION

Whispered speech refers to low-energy pronunciation without vocal cord vibration. It is a special and essential style of speech communication [1]. For example, in places such as libraries and conference rooms where loud speech is prohibited, people generally use whispered speech for human-human communication or human-computer interaction. In addition, to protect the privacy of communication content, people prefer whispering for communication in public places. In the medical field, for patients with laryngectomy, whispering is the only means of communication. In recent years, whispering has become one of the most convenient silent interfaces in the field of human-computer interaction compared with the surface electromyogram (EMG) interface and the magnetic resonance imaging (MRI) interface.

Whispering is a low-energy signal compared with normal voiced speech because the vocal cords do not vibrate when people are whispering. The airflow exhaled from the lungs directly excites the sound cavity through the narrow half-opening glottis to generate unvoiced speech signals, so whispering does not contain the fundamental frequency (F0) [2]. Because of the absence of F0, the whisper has noise-like characteristics, and the energy of a whisper is approximately 20 dB lower than that of normal speech; thus, whispering is more susceptible to noise interference.

Due to wide applications of whispered speech in various communication scenarios, research on whispering has attracted much attention in recent years. For example, [3]–[5] studied whispered speech recognition. Reference [6] investigated the effectiveness of phase information for whispered speech emotion recognition. Production of synthetic whispers from neutral speech recordings has been studied in the feature spaces in, e.g., [7] to augment limited transcribed

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Luo.

whispered recordings for whispered speech recognition. Recently, reconstructing normal voiced speech from whispered speech has attracted the attention of many researchers [8]–[10]. The aim of whisper conversion is to convert whispered speech to normal speech to improve its intelligibility and/or perception quality.

Currently, there are two kinds of whisper conversion technologies. One is rule-based whisper conversion, which modifies parameters of the source-filter model such as mixed excitation linear prediction (MELP), code excited linear prediction (CELP), and linear prediction coding (LPC) according to transformation rules obtained by experimental observations or statistical analysis of the acoustic feature differences between whispered speech and its normal counterpart [11]–[14]. However, transformation rules are always generated by empirical observation or simple statistical modeling, so the naturalness and speech quality of rule-based whisper conversion need further improvement.

The other kind of whisper conversion approaches includes the Gaussian mixture model (GMM) method and the neural network method. Toda et al. first utilized GMM to model the joint spectral feature space of whispering and its normal counterpart. However, the speech spectral envelope estimated by the naive GMM model exhibits discontinuity and oversmoothing. To solve this problem, the dynamic spectral parameter is proposed to model dynamic acoustic space and achieved a performance improvement [15]. In addition, maximum likelihood parameter generation (MLPG) [16] has been used to generate smooth speech parameters, and global variance (GV) has been adopted to enhance the details of spectral parameters [17], [18], further improving the intelligibility and naturalness of the converted speech.

Unlike GMM, neural networks can fit complex nonlinear relationships. In the past few years, neural networks have been widely used for whisper-to-normal speech conversion. Li et al. adopted the restricted Boltzmann machine (RBM) to model the joint feature space composed of whispering and parallel normal speech [19] and obtained a preferable estimated target normal speech. Recently, deep neural networks (DNN) have been used to achieve better conversion performance without a predivided acoustic parameter space [20]. To reflect the interframe relationship, Nisha utilized a deep bidirectional long short time memory (DBLSTM) for speech conversion, and experimental results showed that the converted speech is more natural and is more similar to the target normal speech [21].

However, feature alignment is necessary for state-of-the-art statistical-based whisper conversion methods. Researchers frequently use the dynamic time warping (DTW) algorithm to align features by adding or removing speech frame features using a dynamic program algorithm where the speech acoustic and perception characteristics are not considered [22]. In the training phase, features aligned by DTW are used for model training, which may cause poor speech quality and/or speech intelligibility for the converted speech.

In addition, F0 estimation is another problem when converting whispering to normal speech. Owing to the absence of F0 in whispered speech, existing conversion methods always adopt a model to characterize the relationship between the whisper spectrum and the F0 of its normal counterpart. However, whispering is noise-like, and the spectrum of different phonemes has no significant difference. Thus, the relation between the whisper spectrum and the F0 is not obvious.

To solve these issues, we propose a sequence-to-sequence mapping framework to characterize the nonlinear relationship between original whispered speech features and target normal speech features. In the conversion phase, the converted normal speech is obtained by the trained model. Once the mel frequency cepstral coefficients (MFCC) of the estimated normal speech is obtained, it is used to train a DBLSTM model to characterize the relationship between the estimated MFCCs and F0 of the normal speech.

The remainder of this paper is organized as follows. Section II introduces the proposed sequence-to-sequence whisper-to-normal conversion model based on the auditory attention mechanism. Section III gives experimental results and discussion. The final section is devoted to the conclusion.

II. WHISPER-TO-NORMAL CONVERSION BASED ON SEQUENCE-TO-SEQUENCE MAPPING

The proposed SEQ2SEQ whisper-to-normal speech conversion framework shown in Fig. 1 consists of model training and speech conversion. In the training phase, speech transformation and representation using adaptive interpolation of the weighted spectrum (STRAIGHT) [23] is utilized to extract the spectral envelope, the aperiodic component and F0 of normal speech, and the spectral envelope of whispered speech. The spectral envelope is further transformed into MFCC. A sequence-to-sequence framework is trained for mapping the relationship between the MFCCs of a whisper and those of normal speech. Note that, once the framework is trained well, the estimated MFCCs of converted speech from the SEQ2SEQ model are used to train two DBLSTM models for estimating F0 and the aperiodic component, respectively. In the conversion stage, the MFCCs extracted from whispered speech are used to estimate MFCCs of the target normal speech, which are then used to estimate both F0 and the aperiodic component of the target speech.

The sequence-to-sequence mapping framework is known as a codec structure, which was originally proposed by Cho et al. for machine language translation [24]. The encoder maps the source word features to a high-dimensional eigenspace for decoding. The encoder of the sequence-to-sequence framework reads whispered speech features sequentially and summarizes them into a fixed-length context vector c, which characterizes the entire speech feature sequence information. Given the current hidden state s_t, the decoder takes the context vector c as the input and gradually generates normal speech features at each time step.

The sequence-to-sequence framework can encode a speech feature sequence into a context vector of fixed length and

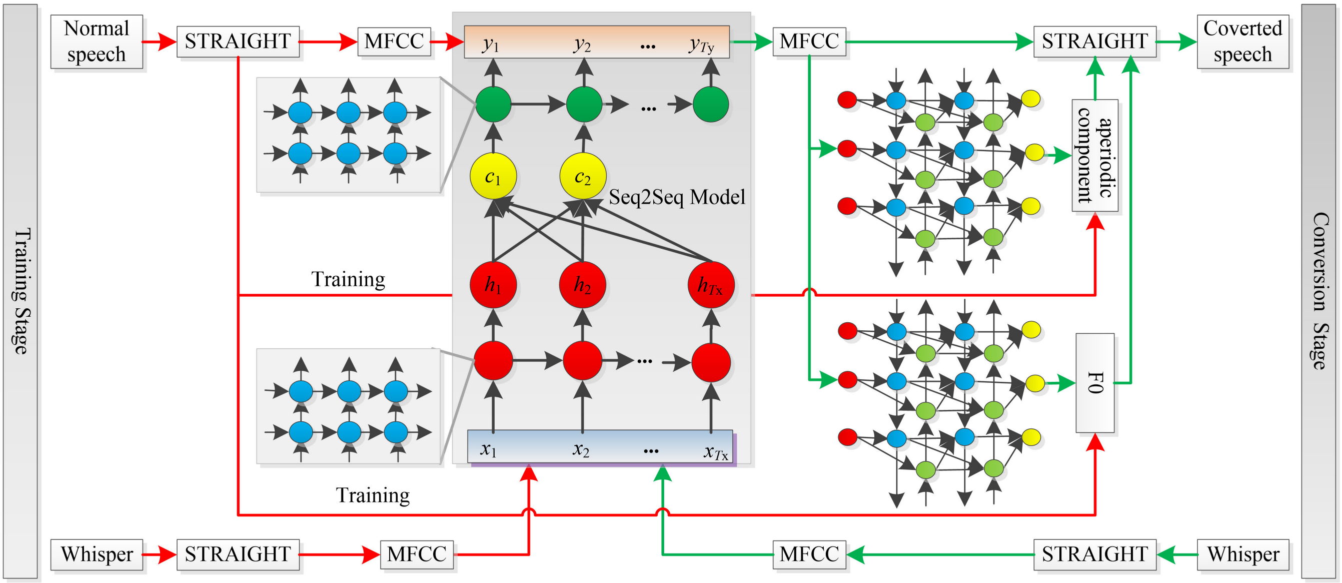


FIGURE 1. Scheme of proposed whisper-to-normal speech conversion model based on SEQ2SEQ framework.

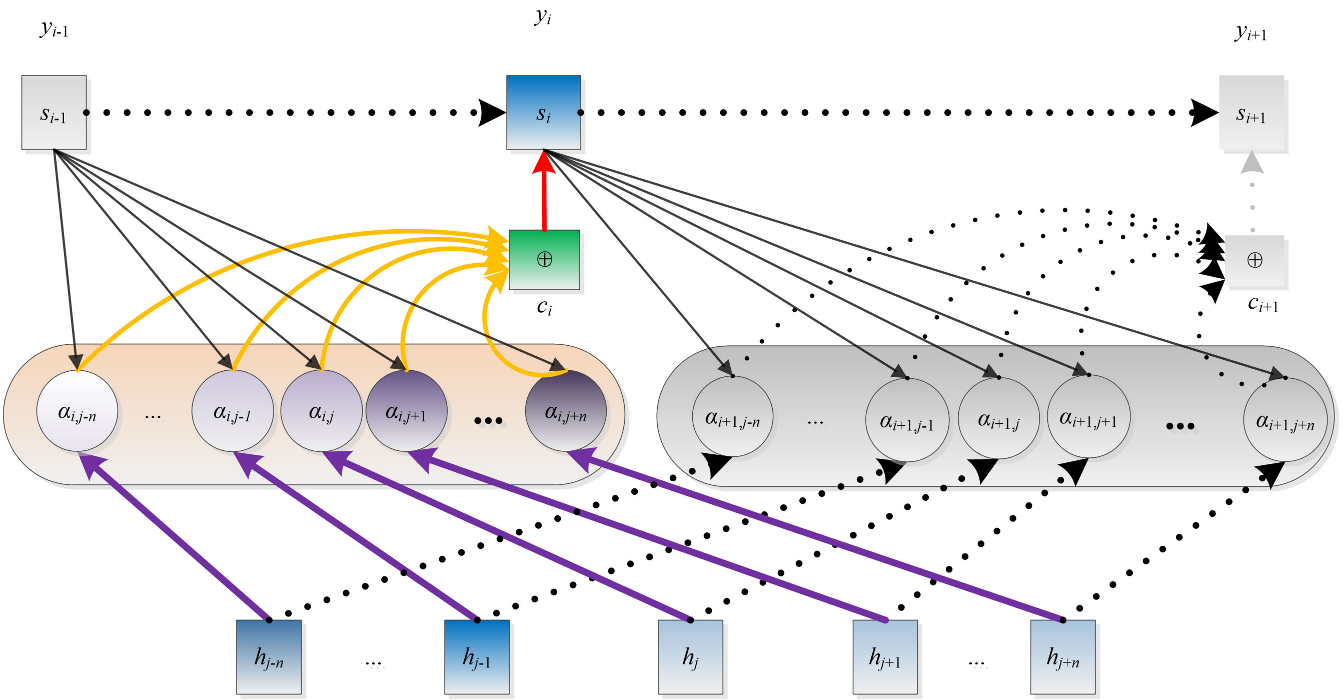


FIGURE 2. Context generative model for whisper-to-normal conversion based on SEQ2SEQ framework.

decode it back into another feature sequence with a different length from the input sequence. That is, the feature sequence length of the source whisper and the normal target speech are not required to be the same. Thus, in the proposed model, the parallel whisper and normal counterpart corpus do not require time-aligning in the training stage. Specifically, the source whisper feature sequence of length T_x is directly decoded into a target normal speech feature sequence of length T_y by the proposed sequence-to-sequence framework.

For speech signal, the speech of the current frame is strongly related to that of previous speech frames. To this end, a long short time memory (LSTM) network is utilized as the encoder of the sequence-to-sequence framework to characterize the implicit relationship between successive frames of whispered speech [25]. Similarly, for the decoder, we also adopt an LSTM to decode the speech contexts back into the feature sequence of target normal speech [26].

Suppose the source whispered speech feature sequence is represented as $\{x_1,\cdots,x_{T_x}\}$ and the target normal speech feature sequence as $\{y_1,\cdots,y_{T_y}\}$ where T_x and T_y represent the length of the source and target sequences, respectively.

In the proposed sequence-to-sequence framework, the LSTM network consists of an input gate, output gate, and forget gate as follows:

$$\begin{aligned} i_t &= \delta(W_i \cdot [h_{t-1}, x_t] + b_i), & (1)\\ f_t &= \delta(W_f \cdot [h_{t-1}, x_t] + b_f), & (2)\\ p_t &= f_t p_{t-1} + i_t * tanh(W_p[h_{t-1}, x_t] + b_p), & (3)\\ o_t &= \delta(W_o \cdot [h_{t-1}, x_t] + b_o), & (4)\\ h_t &= o_t tanh(p_t). & (5) \end{aligned}$$

where i, f, o, and p denote the input gate, forget gate, output gate, and cell states, respectively. δ is the sigmoid function. h_t is the hidden state at time step t.

A conventional sequence-to-sequence framework adopts p_t of the last hidden layer as its context vector once the whole sequence is completely encoded. However, for whisper-to-normal speech conversion, the length of the whisper feature sequence is always longer than that of the normal speech, and different phonemes of normal speech correspond to whisper frames of different length at different positions, so feature vectors of whispering used to obtain target feature vectors of normal speech are dynamically changed at different time steps. To model these various nonlinear relationships more effectively, we adopt an auditory attention mechanism to obtain a self-adaptive context vector, which is used to adaptively estimate the current hidden state and output of the decoder.

In the proposed sequence-to-sequence framework, we suppose that the current state of the decoder is related to all hidden states of the encoder, each of which has a different impact on the estimation of the current state of the decoder. To obtain a self-adaptive context, both past and future hidden states of the encoder are considered simultaneously to obtain the current context.

Suppose that $\{h_1, \cdots, h_{T_x}\}$ represents the hidden state of the encoder. The time-dependent context c_i for obtaining the *ith* feature vector of the converted target normal speech is computed as

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j, \tag{6}$$

with

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{j=1}^{T_x} exp(e_{ij})}, \tag{7}$$

$$e_{ij} = a(S_{i-1}, h_j). \tag{8}$$

where $a(S_{i-1}, h_j) = v^T tanh(WS_{i-1} + Uh_j)$ describes the similarity between the cell state of the decoder and the hidden state of the encoder.

The new state S_i of current frame is obtained as

$$(cell\_out_i, S_i) = lstm(y_{i-1}, S_{i-1}, c_i). \tag{9}$$

where S_{i-1} is the previous state of LSTM, c_i is the current context, $cell\_out_i$ represents the current LSTM output, and y_{i-1} is the feature of the previous target frame. y_i is obtained by

$$y_i = liner(cell\_out_i, c_{i+1}). \tag{10}$$

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATA PREPARATION AND EVALUATION METHODS

In this paper, 348 parallel whisper and normal speech utterances from the CSTR NAM TIMIT Plus corpus database[1] were used to evaluate the effectiveness of the proposed model for whisper-to-normal speech conversion. Each utterance was sampled at 8 kHz, with 16-bit PCM storage. Three hundred forty-eight utterances were randomly separated into a training set with 300 parallel utterances and a test set with 48 parallel utterances. The frame length was set to 40 ms with 5ms frameshifting. The 257-dimensional spectral envelopes were extracted by STRAIGHT. A 30-dimensional MFCC vector for each frame was obtained from the spectral envelope. The first-order difference $MFCC\_dynamic_k$ of the MFCC feature vector is derived as

$$\begin{aligned} MFCC\_dynamic_k = \frac{1}{3}(&-2 * MFCC_{k-2} - MFCC_{k-1} \\ &+ MFCC_{k+1} + 2 * MFCC_{k+2}). \end{aligned} \tag{11}$$

The mel cepstral distance (CD) [27], the short time objective intelligibility (STOI) [28], the perceptual evaluation of speech quality (PESQ) [29], the root mean squared error (RMSE), and the mean duration differences (MDD) were used to objectively evaluate the performance of the converted speech. The mean opinion score (MOS) and ABX preference test were chosen as subjective evaluation methods for converted speech. CD is a common measurement for spectrum conversion performance, which is computed as

$$CD = \frac{10}{log10}\sqrt{2\sum_{d=1}^{D}(C_d - C'_d)^2}. \tag{12}$$

where C_d and C'_d represent the *dth* element of the cepstral coefficients feature of reference normal speech and converted normal speech, respectively. D represents the dimension of the cepstral coefficient feature, which is set to 24. A higher CD value indicates a greater difference between converted normal speech and reference normal speech. The STOI score ranges from 0 to 1, and a larger STOI score indicates a higher intelligibility of converted normal speech. The PESQ is used to evaluate the overall speech quality of converted speech, and the PESQ value ranges from 0 to 5. A larger PESQ value indicates better quality of converted speech.

RMSE is a frequently used objective method to evaluate the similarity of the estimated F0 and the ground truth F0. A smaller RMSE value means more accurate estimation of F0. MDD is used to measure the time duration difference between the converted speech and the target speech.

[1] http://homepages.inf.ed.ac.uk/jyamagis/page3/page57/page57.html

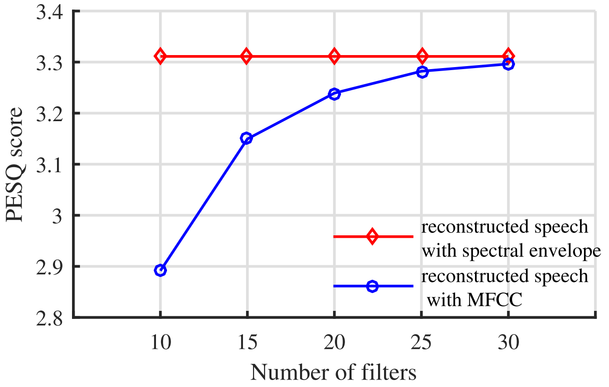


FIGURE 3. PESQ with different numbers of filters on reconstructed speech.

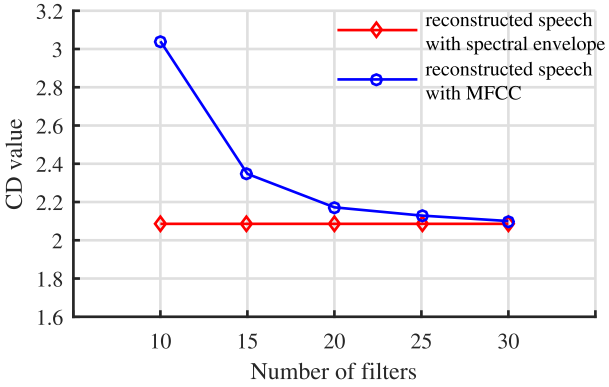


FIGURE 4. CD with different numbers of filters on reconstructed speech.

A smaller MDD value means the time duration of the converted speech is more similar to that of the target normal speech. MOS is a common subjective evaluation method for speech quality. The ABX preference test aims to determine which of the two converted speeches (denoted as A and B) by different conversion methods sounds like the target ground truth speech X. If it is ambiguous to distinguish between A and B, "no preference" is chosen.

To verify the effectiveness of the MFCC feature dimension on the normal speech reconstruction, different numbers of filters were used to generate the MFCC features, which were then used to reconstruct the converted speech. As seen in Fig. 3, Fig. 4, and Fig. 5 when the number of filters was set to 30, the smallest difference between the reconstructed speech and the reference normal speech was obtained. Thus, we used 30 filters for MFCC extraction in subsequent experiments.

For comparison, GMM- [18], DNN- [20], and DBLSTM- [21] based whisper-to-normal speech conversion were conducted. The parallel speech corpuses for training the GMM, DNN, and DBLSTM models were time-aligned by the DTW algorithm.

For the GMM-based whisper-to-normal speech conversion, three GMM models denoted as GMM_dynamic, GMM_f0, and GMM_ap were trained for MFCC, F0 and aperiodic component estimation, respectively. The dynamic and static features were combined as the GMM input. The joint density Gaussian mixture model (JDGMM) was adopted to model the probability distributions of the joint feature of parallel whispered and normal speech. The number of Gaussian components was set to 32 for GM_dynamic and GMM_f0 and 16 for GMM_ap.

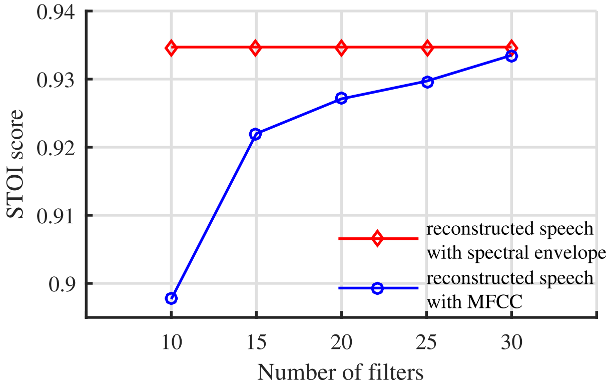


FIGURE 5. STOI with different numbers of filters on reconstructed speech.

For the DNN-based whisper-to-normal speech conversion, three DNN models were used to perform whisper-to-normal speech conversion: the spectral envelope mapping network (denoted as DNN_Dynamic), the F0 estimation network (denoted as DNN_f0), and the aperiodic component estimation network (denoted as DNN_ap). To verify the performance of dynamic features, we also used a DNN_static model where only static features were considered to estimate the spectral envelope. The network configurations of DNN_Dynamic, DNN_f0, DNN_ap, and DNN_static were 60-120-60-120-60, 60-120-60-30-1, 60-120-257-120-257, and 30-120-60-120-30, respectively. For all DNN models, the dropout was set to 0.8, the learning rate was 0.0001, the batch size was 100, and the training epoch was set to 1000.

The DBLSTM-based whisper-to-normal speech conversion model consisted of a spectral envelope estimation module (denoted as DBLSTM_dynamic), F0 estimation module (denoted as DBLSTM_f0), and aperiodic component estimation module (denoted as DBLSTM_ap). Furthermore, an additional spectral estimation model (denoted as DBLSTM_static) using only static features was constructed to evaluate the interframe relationship description ability of DBLSTM. The network configurations of DBLSTM_dynamic, DBLSTM_f0, DBLSTM_ap, and DBLSTM_static were 60-128-256-256-128-60, 60-128-128-128-128-1, 60-128-256-256-128-257, and 30-128-256-256-128-30, respectively. In the training stage, the dropout was set to 0.3, with a learning rate of 10-4, batch size of 10, time step of 100, and training epoch of 500 times. The backpropagation through time (BPTT) algorithm was adopted for DBLSTM model training.

For the proposed sequence-to-sequence whisper-to-normal speech conversion framework, a two-layer LSTM (256-256) was adopted as the encoder, and a two-layer LSTM (256-256) was used as the decoder. In the training stage, we used

TABLE 1. Objective evaluation results of converted speech based on different models.

MODEL	CD value	PESQ score	STOI score
GMM_dynamic	5.44	1.02	0.46
DNN_static	5.14	1.00	0.48
DNN_dynamic	5.06	1.10	0.51
DBLSTM_static	5.00	1.27	0.56
DBLSTM_dynamic	4.99	1.25	0.55
SEQ2SEQ	**2.77**	**1.69**	**0.75**

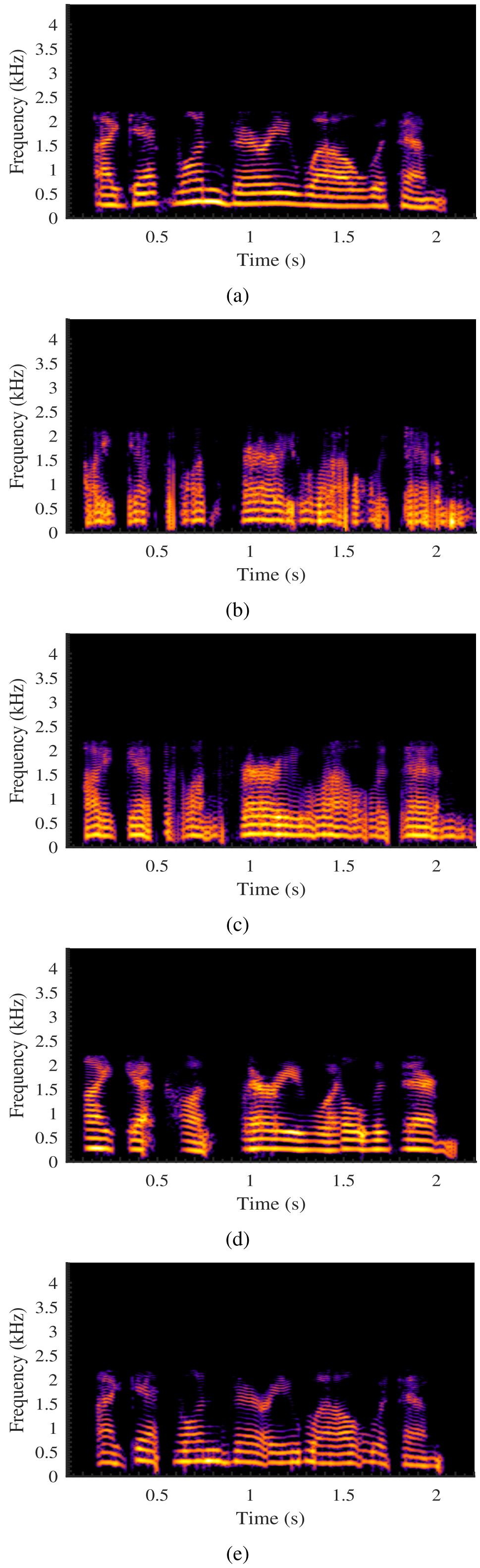
the target speech features of the previous frame to train the hidden state of the current frame, which was then used to obtain the converted speech feature of the current frame. In the speech conversion stage, we performed two types of conversion experiments. In the first experiment, we used the ground truth speech feature of the previous frame to obtain the speech feature of the current frame (denoted as SEQ2SEQ). In the second experiment, the estimated speech feature of the previous frame was used to the obtain speech feature of the current frame (denoted as SEQ2SEQ_pre). The estimated MFCC was used to compute the speech spectrum envelope of the converted normal speech.

The estimated MFCC from the sequence-to-sequence framework was also used to train two DBLSTM models i.e., DBLSTM_seqf0 and DBLSTM_seqap for estimating F0 and the aperiodic component of the normal speech, respectively. The configuration of DBLSTM_seqf0 was the same as that of DBLSTM_f0, and the configuration of DBLSTM_seqap was the same as that of DBLSTM_ap.

B. RESULTS

Table 1 shows the objective evaluation results of normal speech converted by different conversion models. It can be seen that the capability of described interframe characteristics of the conversion model have an essential effect on whisper-to-normal speech conversion performance. In fact, the GMM is a segment linear model that has a weak ability to describe nonlinear relationships, so the performance of the GMM-based whisper-to-normal speech conversion method is poorer than the other three methods. Although the DNN model has excellent nonlinear relation description ability, its interframe relation characterizing ability is simulated by a dynamic feature that is implemented by considering the feature difference between neighboring successive frames.

Compared with the DNN model, the DBLSTM model has a better interframe characterizing ability due to DBLSTM can take advantage of the relationship between long distance frames. To this end, there is no need for a dynamic feature in the DBLSTM method. As seen in Table 1, the DBLSTM-based whisper-to-normal speech conversion method obtained better conversion performance than the GMM and DNN methods. This implies that the static features of MFCC are adequate to characterize the difference between whispering and normal speech for the LSTM-based speech conversion model. Since LSTM was also adopted in the proposed SEQ2SEQ method, we adopt an MFCC static feature vector when modeling the relation between whispering and its normal speech counterpart in the SEQ2SEQ method.



(a)

(b)

(c)

(d)

(e)

FIGURE 6. Comparison of spectrograms. (a) target speech; (b) GMM; (c) DNN; (d) DBLSTM; (e) SEQ2SEQ.

From Table 1, one can see that an objective evaluation of the proposed SEQ2SEQ method outperforms the GMM, DNN and DBLSTM methods. Note that the SEQ2SEQ

TABLE 2. **MOS and MDD of converted speech using different conversion methods.**

Method	MOS	MDD(s)
GMM_dynamic	2.25	0.14
DNN_static	2.36	0.14
DNN_dynamic	2.43	0.14
DBLSTM_static	2.91	0.14
DBLSTM_dynamic	2.93	0.14
SEQ2SEQ	**3.55**	**0.06**

TABLE 3. **RMSE of estimated F0 and ground truth F0.**

Method	GMM	DNN	DBLSTM	SEQ2SEQ
RMSE(HZ)	115.70	89.78	80.44	57.60

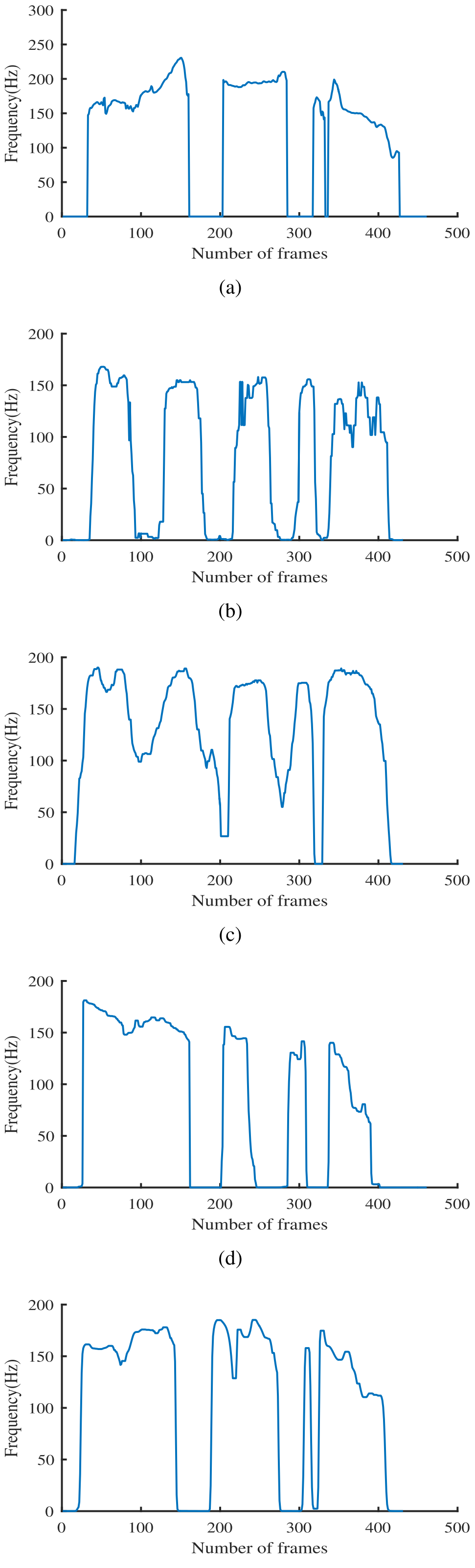
method also adopted LSTM as the encoder and decoder, so the proposed SEQ2SEQ method has similar interframe characterizing ability to DBLSTM. However, the SEQ2SEQ method does not require preprocessed speech time-aligning, which is required in the DBLSTM method. Specifically, the SEQ2SEQ method adopts the auditory attention principle to implement adaptive feature mapping between parallel whispered and normal speech corpuses. We attribute the improvement of performance to the adaptive feature mapping. In addition, the naive DTW-based feature aligning may reduce the speech quality and speech intelligibility of the converted normal speech.

Table 2 shows the MOS and MDD values of converted speeches using different conversion methods. It is obvious that the converted speech using SEQ2SEQ achieves the highest MOS and lowest MDD.

We also plotted spectrograms of normal speech, with the converted speeches using different conversion models for subjective evaluation. One can see in Fig. 6 (a)-(e) that the proposed SEQ2SEQ approach obtains the most similar spectrum of the converted speech to the target normal speech, compared with the GMM, DNN and DBLSTM methods.

Note that the LSTM in the SEQ2SEQ method was used only to map the relationship of the MFCC features of whispered and normal speech. For the STRAIGHT model, the F0 and aperiodic component are also necessary for reconstructing target normal speech. As aforementioned, we adopted two DBLSTMs for estimating F0 and the aperiodic component from MFCC features estimated by the SEQ2SEQ method. Once the MFCCs of the normal speech are obtained, the F0 and aperiodic component can be estimated by these two DBLSTM models, respectively.

The RMSE between the estimated F0 and ground truth F0 is shown in Table 3 One can find that the RMSE value of the proposed SEQ2SEQ method is lowest. Note that the only difference between the SEQ2SEQ method and the DBLSTM method is that the proposed SEQ2SEQ method uses the estimated MFCC of the converted speech, while the DBLSTM method uses the spectrum of the source whisper. This verifies the effectiveness of using the estimated MFCC to estimate F0 of the converted speech.



(a)

(b)

(c)

(d)

(e)

FIGURE 7. **Comparison of F0. (a) reference normal speech; (b) GMM; (c) DNN; (d) DBLSTM; (e) SEQ2SEQ.**

Table 4 shows the ABX preference test score of the voiced speech converted by different methods. The results in the first row indicate that the speech converted by DNN is more similar to the target normal speech than that by the GMM method.

TABLE 4. ABX test results of the converted speech obtained by using different conversion methods.

GMM	DNN	DBLSTM	SEQ2SEQ	No preference
33.62%	45.43%	-	-	20.95 %
-	30.24%	52.43%	-	17.33 %
-	-	15.17%	76.31%	8.52 %

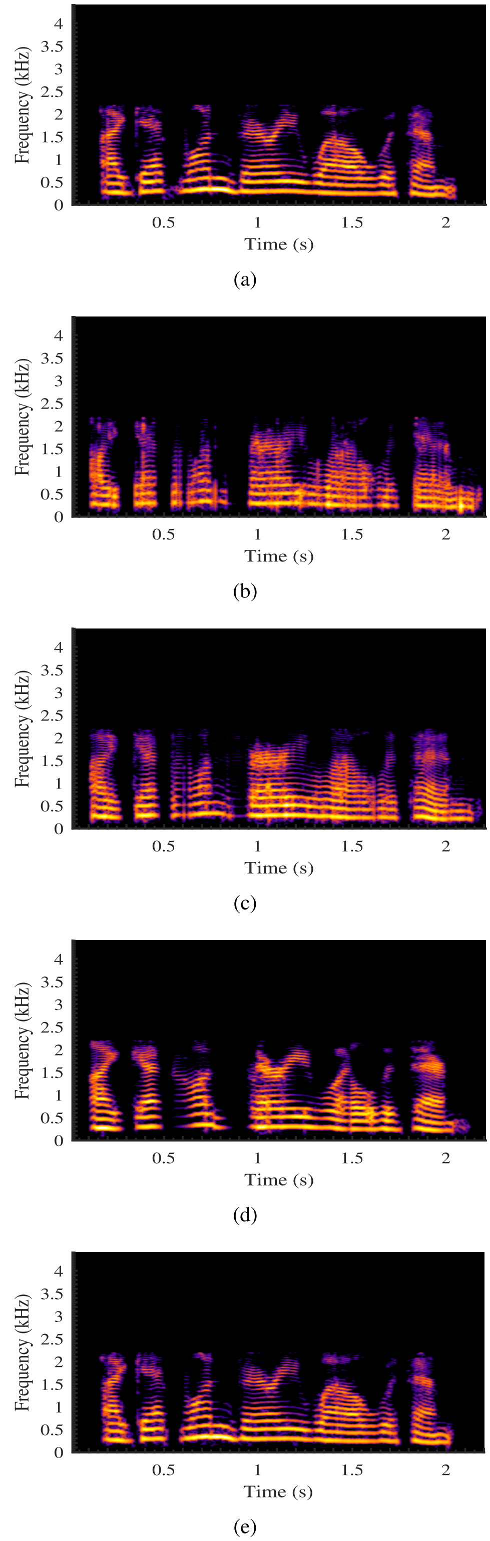
The results in Table 4 show that the converted speech using F0 estimated by the proposed SEQ2SEQ method is more similar to the target ground truth speech than that converted by the other three methods.

For subjective evaluation, Fig. 7 plots F0 of the reference normal speech and F0 estimated by different conversion models. In Fig. 7 (b)-(d), we can see that all methods can estimate F0 well through the spectral envelope. Specifically, F0 estimated by the SEQ2SEQ method is closest to the F0 of the reference target speech.

Although the F0 estimation method of the proposed sequence-to-sequence framework is similar to that of the DBLSTM method, the estimated F0 of the SEQ2SEQ method is more similar to the normal reference speech than that of the DBLSTM method. The reason may be that, for the DBLSTM method, the F0 estimation model considers the relationship of the whispered spectrum and F0 of the normal speech. However, in the proposed SEQ2SEQ method, the F0 estimation model considers the relationship of the normal MFCC features estimated by the SEQ2SEQ model and F0 of the reference normal speech. The F0 estimation results show that the MFCC features estimated by the proposed SEQ2SEQ model are more accurate and have a stronger correlation of F0 than the raw MFCC features of whispered speech.

To evaluate the spectrum envelope estimation performance of different conversion methods, the F0 estimation models in the GMM, DNN, DBLSTM, and SEQ2SEQ methods were replaced by the true F0 of reference normal speech. We plotted spectrograms of normal speech, the converted speech based on different conversion models without F0 estimation, in Fig. 8 for subjective evaluation.

We can see in Fig. 8 (a)-(e) that the spectrum estimated by the GMM method is over-smooth and details of the high-frequency components are unclear, resulting in blurred speech content, thus resulting in unsatisfactory speech naturalness. Although the spectrum estimated by the DNN method retains more high-frequency contents than that by the GMM method, the spectrum of converted speech is still not clear enough. Compared with the spectrum estimated by the DBLSTM method, the spectrum estimated by the proposed SEQ2SEQ method is closer to that of the target reference normal speech.



(a)

(b)

(c)

(d)

(e)

FIGURE 8. Speech spectrum estimated using different method with true F0. (a) reference normal speech; (b) GMM; (c) DNN; (d) DBLSTM; (e) SEQ2SEQ.

To further evaluate the performance of the proposed sequence-to-sequence whisper-to-normal speech conversion framework, the SEQ2SEQ method, the SEQ2seq_pre method, and the DBLSTM_dynamic method were evaluated on the TIMIT corpus database. The experimental results are shown in Table 5, where the SEQ2SEQ_pre method used the speech feature of the previous frame estimated by the sequence-to-sequence framework to estimate the speech feature of the current frame; i.e., the speech feature of each frame of the target speech was estimated by the sequence-to-

TABLE 5. Performance evaluations of different models with the TIMIT corpus database.

Model	CD	PESQ	STOI	MOS	RMSE
SEQ2SEQ	2.77	1.69	0.75	3.55	57.60
SEQ2SEQ_pre	6.55	1.09	0.21	1.00	81.68
DBLSTM_dynamic	4.99	1.25	0.55	2.93	80.44

TABLE 6. Performance evaluations of different models with our corpus database.

Model	CD	PESQ	STOI	MOS	RMSE
SEQ2SEQ	2.78	2.55	0.84	3.58	103.39
SEQ2SEQ_pre	6.52	1.02	0.56	2.92	125.79
DBLSTM_dynamic	6.74	0.82	0.48	2.84	127.29

sequence framework. However, the converted performance of SEQ2SEQ_pre decreased compared to DBLSTM_dynamic. We attribute this to the limited size of training data in the TIMIT corpus database, which comprises only 420 utterances of whispered speech, 348 utterances of which were selected for our model training.

To verify the effectiveness of the SEQ2SEQ_pre method, 1000 sentences from the TIMIT corpus database were selected and pronounced by a female to obtain 1000 utterances of normal speech and 1000 utterances of whispered speech. All 2000 sentences from the corpus can be accessed at ftp://210.45.212.96/ with username: download and password: download. The speech conversion experimental results are shown in Table 6. As seen in Table 6, SEQ2SEQ_pre achieved better conversion performance than the DBLSTM_dynamic method.

IV. CONCLUSION

We proposed a sequence-to-sequence whisper-to-normal speech conversion framework with an auditory attention mechanism. Unlike existing whisper-to-normal speech conversion models where whispered and normal speech corpuses need time-aligning before model training and speech conversion, the proposed sequence-to-sequence whisper-to-normal speech conversion framework does not require feature-aligning features. The proposed sequence-to-sequence framework adopts an additional full connection neural network (FCNN) to simulate the perception attention principle and generate context-adaptive encoding information for normal speech feature estimation. To characterize the interframe relationship of successive frames, the encoder and decoder of the proposed sequence-to-sequence speech conversion method adopts long-short term memory. In addition, better F0 estimation can be obtained from the precisely estimated normal speech spectrum than from the raw whispered speech spectrum. The proposed sequence-to-sequence framework has the ability of adaptive nonlinear mapping between whispered and normal speech. It can also be used for normal speech conversion.

REFERENCES

[1] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Commun.*, vol. 52, no. 4, pp. 301–313, Apr. 2010.

[2] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. Phys.*, vol. 24, nos. 7–8, pp. 515–520, Sep./Oct. 2002.

[3] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1408–1421, Jul. 2011.

[4] C. Xueqin, Z. Heming, and F. Xiaohe, "Performance analysis of mandarin whispered speech recognition based on normal speech training model," in *Proc. 6th Int. Conf. Inf. Sci. Technol. (ICIST)*, Dalian, China, 2016, pp. 548–551.

[5] T. Grozdić and S. T. Jovičić, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2313–2322, Dec. 2017.

[6] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.

[7] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1705–1720, Oct. 2016.

[8] I. V. Mcloughlin, J. Li, and Y. Song, "Reconstruction of continuous voiced speech from whispers," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 1022–1026.

[9] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.

[10] A. Ferreira, "Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information," in *Proc. Int. Symp. Signal, Image, Video Commun. (ISIVC)*, Tunis, Tunisia, 2016, pp. 159–166.

[11] F. Ahmadi, I. V. McLoughlin, and H. R. Sharifzadeh, "Analysis-by-synthesis method for whisper-speech reconstruction," in *Proc. IEEE Asia Pacific Conf. Circuits Syst. (APCCAS)*, Macao, China, Nov./Dec. 2008, pp. 1280–1283.

[12] M. Yang, F. Qiu, and F. Mo, "A linear prediction algorithm in low bit rate speech coding improved by multi-band excitation model," *Acta Acustica*, vol. 26, no. 4, pp. 329–334, Jul. 2001.

[13] H. R. Sharifzadeh, I. V. Mcloughlin, and F. Ahmadi, "Regeneration of speech in voice-loss patients," in *Proc. 13th Int. Conf. Biomed. Eng. (ICBME)*, Singapore, 2008, pp. 1065–1068.

[14] I. V. Mcloughlin, H. R. Sharifzadeh, L. T. Su, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Trans. Access. Comput.*, vol. 6, no. 4, p. 12, Jun. 2015.

[15] M. Ahangar, M. Ghorbandoost, S. Sharma, and M. J. T. Smith, "Voice conversion based on a mixture density network," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 329–333.

[16] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, Jun. 2000, pp. 1315–1318.

[17] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[18] T. Toda and K. Shikano, "NAM-to-speech conversion with Gaussian mixture models," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Lisbon, Portugal, 2005, pp. 1957–1960.

[19] J.-J. Li, I. V. McLoughlin, L.-R. Dai, and Z.-H. Ling, "Whisper-to-speech conversion using restricted Boltzmann machine arrays," *Electron. Lett.*, vol. 50, no. 24, pp. 1781–1782, Nov. 2014.

[20] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 2579–2583.

[21] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional LSTMs," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Hyderabad, India, 2018, pp. 491–495.

[22] M. V. Ramos, "Voice conversion with deep learning," M.S. thesis, Dept. Elect. Comput. Eng., Instituto Superior Técnico, Lisbon Univ., Lisbon, Portugal, 2016.

[23] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigné, "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 Commun.*, vol. 27, nos. 3–4, pp. 187–207, Apr. 1999.

[24] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," Jun. 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

[25] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[26] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw. (ICANN)*, Edinburgh, U.K., 1999, pp. 850–855.

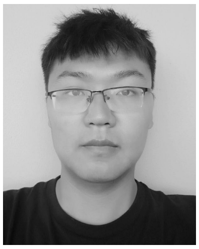
[27] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun, Comput., Signal Process. (PACRIM)*, Victoria, BC, Canada, May 1993, pp. 125–128.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4214–4217.

[29] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, document BS.1534-1, ITU-R, Geneva, Switzerland, 2001.

HAILUN LIAN received the B.S. degree in computer science and technology from Anhui Jianzhu University, China, in 2017. He is currently pursuing the M.S. degree in computer science and technology with Anhui University, China. His research interests include speech signal processing and deep learning.

YUTING HU received the B.S. degree in information and computational science from the Anhui University of Science and Technology, China, in 2017. She is currently pursuing the M.S. degree in computer science and technology with Anhui University. Her current research interests include speech signal processing and image processing.



WEIWEI YU received the B.S. degree in microelectronics from Anhui University, China, in 2016, where he is currently pursuing the M.S. degree in computer science and technology. His research interests include speech signal processing and deep learning.

JIAN ZHOU received the Ph.D. degree in information and communication engineering from Southeast University, China, in 2013. He has been an Associate Professor with the School of Computer Science and Technology, Anhui University, China, since 2014. His current research interests include speech and image processing, and pattern recognition.

WENMING ZHENG (M'08–SM'18) received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004, where he is currently a Professor with the Key Laboratory of Child Development and Learning Science, Ministry of Education. He has been with the Research Center for Learning Science, since 2004. His research interests include affective computing, pattern recognition, machine learning, and computer vision. He is currently an Associate Editor of the IEEE Transactions on Affective Computing and of *Neurocomputing*, and an Editorial Board Member of *Visual Computer*.