IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 19, 2019, accepted April 7, 2019, date of publication April 23, 2019, date of current version May 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912926

A BLSTM and WaveNet-Based Voice Conversion Method With Waveform Collapse Suppression by Post-Processing

XIAOKONG MIAO, XIONGWEI ZHANG, MENG SUN, CHANGYAN ZHENG, AND TIEYONG CAO
Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing 210007, China

Corresponding authors: Xiongwei Zhang (xwzhang9898@163.com) and Meng Sun (sunmengccjs@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471394, and in part by the Natural Science Foundation of Jiangsu Province for Excellent Young Scholars under Grant BK20180080.

ABSTRACT In recent years, neural network-based voice conversion methods have been rapidly developed, and many different models and neural networks have been applied in parallel voice conversion. However, the over-smoothing of parametric methods [e.g., bidirectional long short-term memory (BLSTM)] and the waveform collapse of neural vocoders (e.g., WaveNet) still have negative impacts on the quality of the converted voices. To overcome this problem, we propose a BLSTM and WaveNet-based voice conversion method cooperated with waveform collapse suppression by post-processing. This method firstly uses BLSTM to convert the acoustic features between parallel speakers, and then synthesizes pre-converted voice with WaveNet. Subsequently, several alternative iterations of BLSTM post-processing is performed, and the final converted voice is generated by WaveNet. The proposed method can directly generate converted audio waveforms and avoid the waveform-collapsed speech caused by a single WaveNet generation effectively. The experimental results indicate that acoustic features trained by using the BLSTM network could achieve better results than conventional baselines. From our experiments on VCC2018, the usage of WaveNet could alleviate the problem of over-smoothing, which contributes to improving the similarity and naturalness of the final results of voice conversion.

INDEX TERMS Voice conversion, speech synthesis, BLSTM, WaveNet.

I. INTRODUCTION

Voice conversion (VC) is a method for seeking to convert one speaker's voice into another voice while maintaining the content unchanged. The technology of voice conversion has been widely used in many fields, such as text-to-speech (TTS), speech enhancement, emotion conversion and other applications [1], [2]. In recent years, machine learning has contributed many solutions to solving the problems in voice conversion such as Gaussian mixture model (GMM) [3], [4], frequency warping [5]–[7], deep neural network (DNN) [8]–[10], and so on. These frameworks of voice conversion mainly consist of two phases: a training phase and a conversion phase. During the training phase, the relevant conversion function is extracted through the parallel corpus of source speaker and target speaker. During the conversion phase, the conversion function is applied on features extracted from new input voice. However, when waveform generation is conducted by a pre-trained parametric vocoder, over-smoothing would frequently appear, which leads to the missing of detailed information in the waveform of the converted voice and makes the converted voice sound buzzy.

In 2016, DeepMind developed a generative model for creating audio waveforms, called WaveNet. The model works by predicting the distribution for each audio sample conditioned on previous ones [11], which enables it to model audio waveforms accurately. Therefore, it is able to directly generate natural-sounding voice and alleviate the problem of over-smoothing. However, it is also reported that because the WaveNet uses causal convolution to predict next sound sample, the occasional instability of a generated sample will have impacts on samples generated subsequently, especially when inaccurate acoustic features are used as local conditional

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

parameters [12]. This phenomenon was also observed in our experiments. On the other hand, in Voice Conversion Challenge 2016, long short-term memory (LSTM) demonstrated its superiority on the mapping of spectral parameters. As an improved version of LSTM, BLSTM has been utilized in voice conversion given its strong ability on modeling contextual relations [8].

In this paper, we propose to combine BLSTM with WaveNet to achieve an unified cascaded model. The first step is to train a WaveNet vocoder, which can synthesize the target voice through the features of target voice. Then a feature conversion network of BLSTM (named by BLSTM1) is trained. The converted spectral parameters by BLSTM1 network are fed into the WaveNet as condition variables, then pre-converted voice is generated. The features of the pre-converted voice are extracted again and sent to a post-processing network modeled by another BLSTM (named by BLSTM2). Finally the converted voice is generated through the WaveNet vocoder. The proposed method effectively avoids waveform collapse, solves the problem of over-smoothing and improves the quality and effectiveness of voice conversion.

The rest of this paper is organized as follows. Section II mainly reviews some related work involved in this paper. Section III describes the proposed method of voice conversion. The training algorithm and its procedures are presented in section IV. The setup of experiments and analysis of results are given in section V. We conclude in section VI with a brief summary and present the future work.

II. RELATED WORK

BLSTM is an improvement of bidirectional recurrent neural network (RNN), which can model a certain amount of contextual information with cyclic connections and map the whole history of previous inputs to each output in principle.

However, in the back-propagation optimization of bidirectional RNN networks, it is found that the accumulation or attenuation of back-propagation gradients explode or vanish over time in long range contextual transmission. An effective way to overcome this problem is to introduce long-term and short-term memory architectures, which can store information in a linear storage unit for many temporal steps, and can learn the optimal amount of contextual information related to regression tasks [13]. Fig.1 shows a LSTM network with gated cells. It is a peephole LSTM unit with input, output and forget gates. BLSTM consists of many long short-term memory cells and the bidirectional RNN. BLSTM is found to be able to take into account both forward and backward sequential information, hence it is adopted in this paper to accurately depict acoustic features.

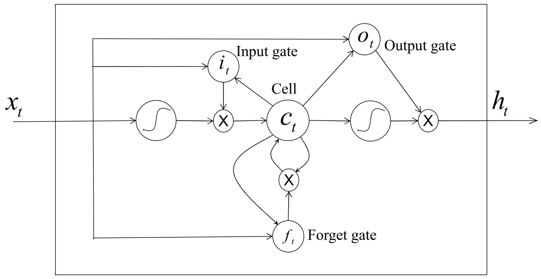


FIGURE 1. A long short-term memory cell [8].

A LSTM cell is defined by the following terms, [14]–[16]:

$$i_t = \sigma_i(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma_f(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \tag{2}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \tag{3}$$

$$o_t = \sigma_o(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \tag{4}$$

$$h_t = o_t \tanh(c_t), \tag{5}$$

where σ is sigmoid function, b_i, b_f, b_c, and b_o are bias vectors and i, f, o, c are the input gate, forget gate, output gate and cell vectors, respectively. All gates share the same size as the hidden vector h. The term x_t is the input to the memory cell at time step t. W_{mn} (e.g. W_{hi}, W_{ci}, etc.) are weight matrices. Please refer to [17] for details of BLSTM.

BLSTM has been used for voice conversion in [8]. The conversion process is shown in Fig.2. Fundamental frequency (F0), spectrum envelope and aperiodic frequency (AP) are extracted. These parameters are used for modeling voice conversion given the fact that high quality speech can be synthesized from these parameters. However, as mentioned earlier, there are still some problems to be solved, and the quality of converted voice can further be improved.

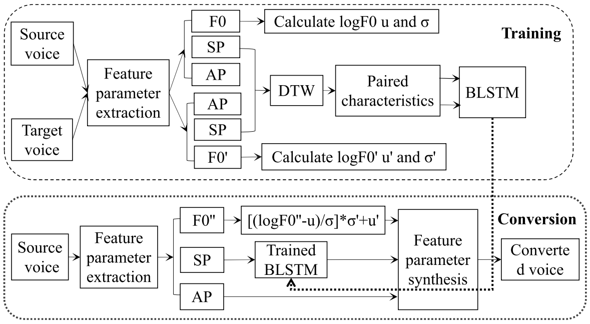
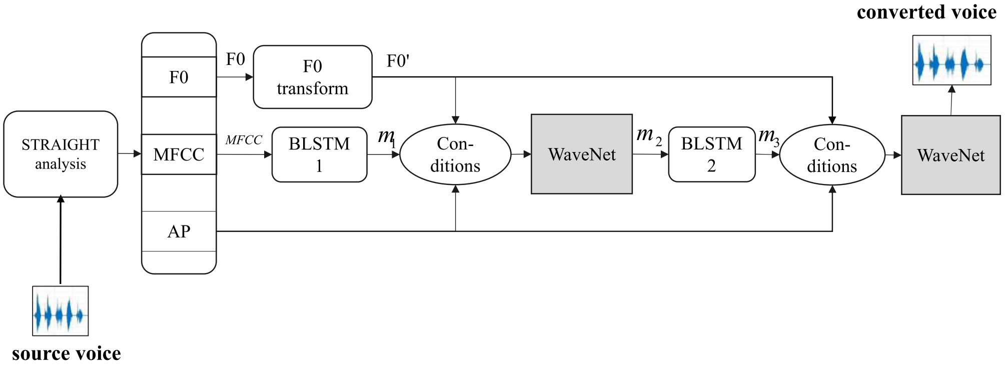


FIGURE 2. Voice conversion framework based on BLSTM, where AP, SP, F0, and DTW represent aperiodic frequency, spectrum envelope, pitch period, and dynamic time warping, respectively.

WaveNet is a deep auto-regressive and generative model for producing waveforms. WaveNet can model the conditional probability $P(X|\theta)$ given conditional inputs θ;

$$P(X|\theta) = \prod_{t=1}^{T} P(x_t|x_1, \cdots, x_{t-r}, \theta), \tag{6}$$

where t and r are the sample index and the size of the receptive field respectively. x_t is the current audio sample and T is the total number of audio sampling. θ represents the conditional feature vector.



wav：source voice waveform
F0：pitch AP:Aperiodic frequency
MFCC：Aligned source mel-frequency cepstral coefficient

F0'：converted pitch
m_1：BLSTM1 converted MFCC
Conditions：Upsampled all condition parameters
WaveNet：Trained WaveNet

m_2： The MFCC of pre-converted voice
m_3：BLSTM2 converted m_2
Conditions：Upsampled all condition parameters

FIGURE 3. The generation framework of voice conversion based on BLSTM and WaveNet.

The formulation of the gated activation units of WaveNet is defined as follows [18]:

$$z = \tanh\left(W_f * x + V_f * y\left(\theta\right)\right) \odot \sigma\left(W_g * x + V_g * y\left(\theta\right)\right), \tag{7}$$

where $*$ denotes a convolution operator and $\odot$ denotes an element-wise multiplication operator. σ is the sigmoid function. V_f is the convolution weight for the condition features. It is worth mentioning that $V_f * y(\theta)$ and $V_g * y(\theta)$ represent the convolutions of each layer. $y(\theta)$ is a modified version of the original condition features θ by adjusting its length to be consistent with that of x.

III. PROPOSED METHOD

A. OVERALL ARCHITECTURE

Fig.3 shows the diagram of our proposed method. The conversion process can be divided into three main stages: WaveNet for waveform generation, BLSTM1 for acoustic feature conversion, and BLSTM2 for post-processing. Firstly, we use STRAIGHT to extract three types of features of speech [19]: AP, F0 and mel frequency cepstral coefficients (MFCC). Then each type of features is processed and converted respectively. F0 is converted by log-linear transform. AP is processed by copying. MFCC is converted by BLSTM1. Then the converted parameters are used as WaveNet conditions to generate pre-converted voice. Then the MFCCs of pre-converted voices are extracted and processed by BLSTM2. Finally, the parameters after post-processing are used as the conditions of WaveNet to generate the final converted voice.

The number of waveform samples in time domain is normally larger than the number of frames. In order to align the frame-level conditional variables (i.e. MFCC, F0, AP) with the sample-points of time-domain, linear interpolation is utilized here. At the same time, minimum mean squared error (MSE) criterion is used as the loss function in BLSTMs and cross-entropy is used as the loss function in WaveNet training. Considering the influence of acoustic feature conversion and the waveform-collapsed speech caused by occasional error in the process of waveform generation, the post-processing training and a two-stage voice generation process are used in the whole process of network conversion. In this way, the quality of converted voice is expected to be further improved.

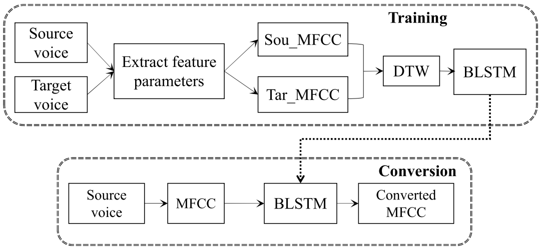


FIGURE 4. The framework of MFCC conversion based on BLSTM.

B. FEATURE CONVERSION

Feature conversion includes spectrum envelope conversion and F0 conversion. MFCC is used as the feature to be converted by BLSTM [20]. Fig.4 shows the framework for realizing the conversion of MFCC based on BLSTM.

For simplicity and effectiveness, log linear transform is applied to F0 conversion as follows,

$$p_t{}^{(Y)} = \frac{p_t{}^{(X)} - u^{(X)}}{\sigma^{(X)}} \times \sigma^{(Y)} + u^{(Y)}, \tag{8}$$

where $p_t{}^{(X)}$ and $p_t{}^{(Y)}$ are the original logF0 and the converted logF0, respectively. $u^{(X)}$ and $u^{(Y)}$ are the means, $\sigma^{(X)}$ and $\sigma^{(Y)}$ are the standard deviations of the training data for the

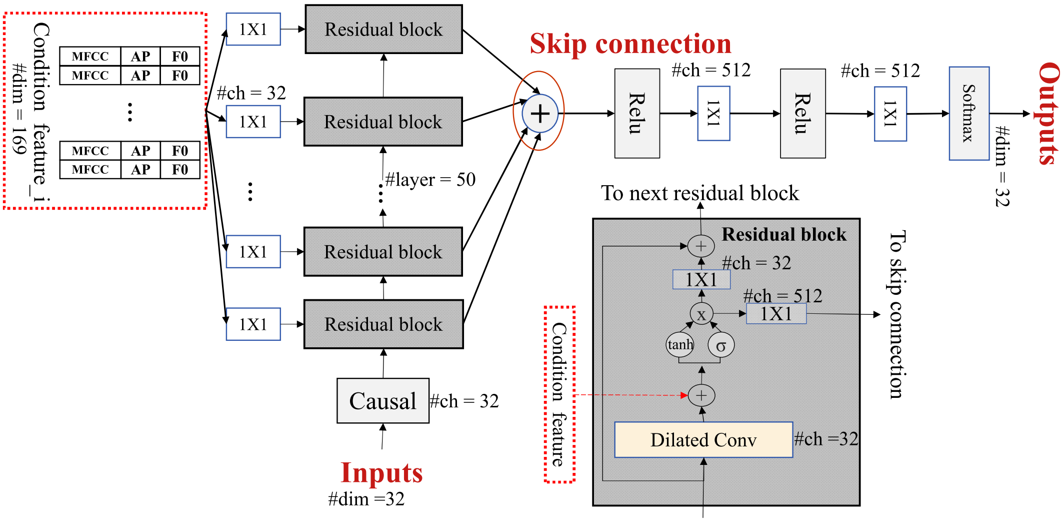


FIGURE 5. The structure of conditional waveNet vocoder. The vocoder consists of casual convolution, residual blocks, and skip connections. The residual blocks are denoted by dark grey boxes, whose conditional features are shown in dotted red boxes. The dark grey box in the bottom right corner shows the zoomed-in internal structure of a residual block, where *σ* and ***tanh*** represent sigmoid and tanh activation functions, respectively. ***Causal*** represents causal convolution and ***Relu*** represents rectified linear unit. ***Skip*** connection is to prevent the gradient dispersion and degradation caused by the increase of network layers. #***dim*** and #***ch*** represent dimensions and number of channels, respectively.

source and the target speakers, respectively [2]. These means and standard deviations are extracted from training data.

C. CONDITIONAL WAVENET VOCODER

As shown in Fig.5, a conditional WaveNet vocoder is used to generate speech samples. In our WaveNet model, MFCC, F0 and AP are used as condition variables. The WaveNet consists of 50 connected residual blocks. Each of the residual blocks includes a dilated causal convolution and a gate activated function. The input waveforms are quantified to 8 bits based on μ-law encoding and the generated waveforms are restored by the μ-law decoding [21].

In order to ensure the original condition features θ to be aligned to x, the method of linear interpolation is used to replace the copying method, which is to make each sample point of x correspond to correct and accurate conditional $y(\theta)$ as much as possible. Moreover, the linear interpolation also maintains the continuity between adjacent frames and the correlations between sample points. In practical operation, we only perform linear interpolation expansion on MFCC, but make copying on both F0 and AP to avoid using too many parameters. The method of linear interpolation upsampling can be seen in Fig.6, where n represents the number of sample points per frame and Δy represents incremental difference between adjacent sample points after interpolation.

D. POST-PROCESSING

WaveNet is based on causal convolution to predict the next voice sample. However, it sometimes generates very noisy speech called waveform collapse when only a limited amount of training data is available or significant acoustic mismatches exist between the training and testing data.

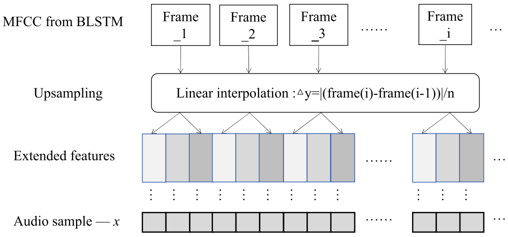


FIGURE 6. The computing of incremental difference between adjacent sample points [2].

Such limitations on the corpus and limited ability of the model are frequently observed in real-word scenarios, such as voice conversion and speech enhancement [12]. The term waveform-collapsed speech refers to the phenomenon that affected by the bad synthesizing effect of a previous less accurately predicted point, the subsequent synthesized speech is thus seriously distorted. In order to overcome the problem of speech collapse, a post-processing network is added after generating the pre-converted voice. Its motivation is to make the converted feature as close as possible to the target feature through the post-processing network, and to reduce the reconstruction error of features caused by a single sample-generation process.

IV. TRAINING ALGORITHM AND PROCESS

Fig.7 represents the training framework of the BLSTM voice conversion with conditional WaveNet. The details are given in Algorithm 1. Firstly, WaveNet vocoder is trained. The upsampled target features (MFCC, F0 and AP) and voices are fed into WaveNet. The WaveNet is subsequently trained to restore the target voices effectively given the conditional

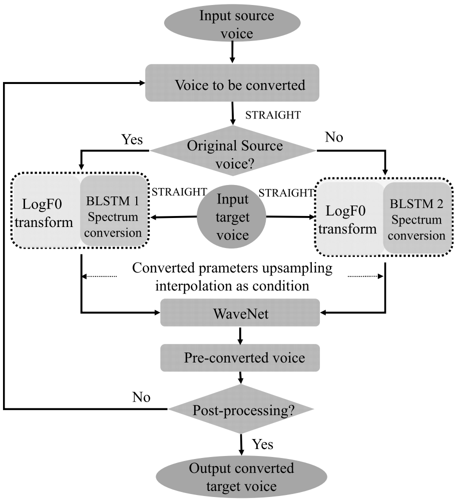


FIGURE 7. **The training framework of the BLSTM voice conversion with conditional waveNet, where STRAIGHT represents the step of the analysis and extraction of features, waveNet represents a pre-trained waveNet, and the dotted boxes contain two isolated operations, i.e. logF0 transform and spectrum conversion by using BLSTM.**

information. Subsequently, BLSTM1 is trained to perform feature conversion. When the input is the original source voice, the MFCCs of target voice and source voice aligned after pre-processing are fed into BLSTM1. BLSTM2 is then trained to perform post-processing. Pre-converted voice is generated by BLSTM1 and WaveNet. The MFCCs of pre-converted voice and target voice are sent to BLSTM2, and MSE is used again as loss function in BLSTM2. Finally, with the parameters (MFCC, F0 and AP) obtained from post-processing, converted speech waveform is synthesized by the trained WaveNet.

V. EXPERIMENTS AND RESULT ANALYSIS

A. EXPERIMENTAL SETUP

VCC2018 is the 2nd Voice Conversion Challenge to evaluate different voice conversion systems and approaches using the same voice data. The corpus includes two subsets, a parallel one (HUB) and a non-parallel one (SPOKE). We choose VCC2SF1, VCC2SF2, VCC2SM1 and VCC2SM2 as the source voice data. The source database contains these sentences uttered by two male and two female speakers. Each speaker has 81 sentences. The target database consists of two male and two female speakers. For each speaker, 66 sentences are used for training data, with the remaining 15 sentences are used for testing. The waveforms in the directory are in RIFF/WAVE format. The sampling rate is 22.05 kHz and is stored in 16-bit format.

To make comprehensive evaluation on datasets of different language, experiments on a mandarin dataset of *CASIA* is also conducted [22]. The folder of *same-text*-300 consists of two males and two females speakers. The female speaker *liuchang* and the male speaker *wangzhe* are used as source speakers, while the female speaker *zhaoquanyin* and the male speaker

Algorithm 1 The Whole Training Framework for Parallel Voice Conversion

```
Require:
$x$: source voice; $y$: target voice;
Initial WaveNet parameter $\theta_W$, BLSTM1 parameter $\theta_{B1}$ and BLSTM2 parameter $\theta_{B2}$;
Extract the acoustic characteristics of $x$: $m_x = MFCC_x, f_x = F0_x$ and $A_x = AP_x$;
Extract the acoustic characteristics of $y$: $m_y = MFCC_y, f_y = F0_y$ and $A_y = AP_y$;
DTW$(m_x, m_y)$: $m'_x = m_{xDTW}, m'_y = m_{yDTW}$;
Learning rate $\eta$, the number of WaveNet training iterations $n_1$, the total number of training iterations in the two previous steps $n_2$, the number of BLSTM1 training iterations $(n_2 - n_1)$, the number of total iterations $n_3$.
Begin step 1: Train the WaveNet as vocoder for waveform generation.
for epoch $= 1, \cdots, n_1$ do
    for training data in $(m'_y, f_y, A_y, y)$ do
        upsample $(m'_y, f_y, A_y)$ and generate $\hat{y}$ from the WaveNet:
        $\hat{y} = W(\, m'_y, f_y, A_y \,)$
        update $\theta_W$ with cross-entropy loss criterion:
        $\theta_W \leftarrow \theta_W - \eta_W \nabla_{\theta_W} L_{cross-entropy}\left(y, \hat{y}\right)$
    end for
end for
End
Begin step 2: Train BLSTM1 as the feature conversion network.
for epoch $= n_1, \cdots, n_2$ do
    for training data in $(\, m'_x, m'_y)$ do
        generate $m_1$ from the BLSTM1:
        $m_1 = B1(m'_x, m'_y)$
        update $B1$ $(m'_x, m'_y)$ with MSE criterion:
        $\theta_{B1} \leftarrow \theta_{B1} - \eta_{B1} \nabla_{\theta_{B1}} L_{MSE}(\, m'_y, m_1 \,)$
    end for
end for
End
Begin step 3: Train BLSTM2 as the post-processing network.
for epoch $= n_2, \cdots, n_3$ do
    for training data in $(\hat{y}, m_1, f_x, A_x)$ do
        $f'_x$: log-linear converted $f_x$.
        upsample $(m_1, f'_x, A_x)$ and generate $\hat{x}$ from the trained WaveNet:
        $\hat{x} = W\left(m_1, f'_x, A_x\right)$
        extract the acoustic characteristics:
        $\hat{x}$: $m_2 = MFCC_{\hat{x}}$; $\hat{y}$: $\hat{m}_y = MFCC_{\hat{y}}$
        generate $m_3$ from the WaveNet:
        $m_3 = B2\left(m_2, \hat{m}_y\right)$
        update $\theta_{B2}$ with MSE criterion: $\theta_{B2} \leftarrow \theta_{B2} - \eta_{B2} \nabla_{\theta_{B2}} \left(L_{MSE}\left(\hat{m}_y, m_3\right)\right)$
    end for
end for
upsample $(m_3, f'_x, A_x)$ and then generate the final voice:
$Y_{final} = W\left(m_3, f'_x, A_x\right)$
End
```

ZhaoZuoxiang are used as target speakers. The folders of *normal* in each speaker are used as training and testing. For each speaker, 66 sentences are randomly selected as training data, and 15 sentences are selected as testing data. The results on the mandarin dataset is shown in the section of experimental results. The feature extraction framework is based on STRAIGHT [23], which is used to extract a 39-dimensional MFCC, 129-dimensional AP, and 1-dimensional F0 for each frame with 25ms length and 5ms shift.

The learning rate of WaveNet was set as 0.0001. The optimization method was chosen as Adam. At the same time, conditional information and conditional filter also performed causal convolution with different layers. The total number of dilated causal convolution channels was 32, and the dilations of 50 layers were set to 5 sets of $[2^0, 2^1, 2^2, 2^3, \ldots, 2^9]$. The 1×1 convolutions in the residual block were set to 32 channels, and the number of 1×1 convolution channels between the skip-connection and the softmax layer was 512. The initial filter width was set to 32.

We have concluded through experiments that the WaveNet was trained with ground truth is better than WaveNet trained with estimated values, especially when the training data set is limited. In the BLSTM1 network, the number of hidden units was chosen as 50, and the initial learning rate is 0.0001. The optimization function was Adam. The epoch size was set to 1.0×10^5. And if the loss was less than 0.003, the training would be terminated. In the BLSTM2 post-processing network, the number of hidden units was chosen as 50, and the initial learning rate was 0.0002. The epoch size was set to 2.0×10^3. In addition, we had tried to replace the traditional linear conversion method of F0 by using the BLSTM neural network. However, it was found that the conversion effect was unsatisfactory, so linear transformation was still utilized to reduce the volume of the training model.

In order to prove the effectiveness of the proposed method, the following three methods were selected for comparison: GMM-VC, WaveNet-VC, BLSTM+WaveNet-VC, N12-VC.

Alternative BLSTM and Wavenet-VC(ABW-VC): It refers to the voice conversion method proposed in this paper.

GMM-VC(G-VC): A conventional GMM-based voice conversion system. We selected the GMM voice conversion method in VCC-2018 as the baseline of voice conversion.

WaveNet-VC(W-VC): Niwa *et al.* introduced the WaveNet-VC method, which directly models the target voice and source features [2]. A WaveNet-based model with 5 blocks (50 layers in total) was used. Specifically, dilations in the 10 layers were set to $[2^0, 2^1, 2^2, 2^3, \ldots, 2^9]$, and this was repeated five times to form a total of 50 dilated causal convolution layers. The number of channels for dilated causal convolutions and residual connections was 32.

BLSTM and WaveNet-VC(BW1-VC): A method used BLSTM network to transform the feature of MFCC, and then utilized the converted MFCC for WaveNet to synthesize the converted voice without post-processing. The BLSTM network had 50 hidden units, and the initial learning rate was 0.0001. The optimization method was Adam. The number of training epochs was set to 1.0×10^5.

N12-VC: This method used waveform filtering to generate waveforms. And it was submitted to VCC2018 [24]. The conversion results of this method were published and compared with the proposed method. This method demonstrated good performance in VCC2018 and was superior to most of the participating algorithms.

B. EXPERIMENTAL RESULTS

Objective and subjective tests were conducted on the HUB task corpus of Voice Conversion Challenge 2018 and a mandarin data set of CASIA, both of which have data from parallel speakers.

The objective measurement is the mel cepstral distortion (MCD), which is defined as:

$$\mathrm{MCD(con,tar)[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{I}({m_i}^{con} - {m_i}^{tar})^2}, \tag{9}$$

where ${m_i}^{con}$ and ${m_i}^{tar}$ are the mel cepstral coefficients of converted features and target features, respectively [21]. I refers to the total number of frames in a sentence.

The smaller the MCD value, the closer the converted voice is to the target voice. Fig.8 and Fig.9 show the MCD scores of different methods on two different databases (i.e. mandarin and English voice datasets), respectively. It should be noted that the *Source* in the figures represents the MCD values between the source voice and the target voice.

MCD is the objective evaluation metric of voice conversion. As is shown in the average bars of Fig.8 and Fig.9, it can be seen that the proposed algorithm (ABW-VC) performed better than two other algorithms based on Wavenet, i.e. W-VC and BW1-VC. However, it failed to reach a lower MCD value than G-VC with direct mel-cepstrum conversion by using Gaussian mixture models (GMM) which was demonstrated as a strong baseline in VCC2018. One possible reason would be that two different objective functions were utilized in their training, i.e. mel-cepstrum distortion for GMM and waveform reconstruction error for WaveNet. G-VC took directly minimizing MCD as its goal while WaveNet prefered to produce waveform sounds similar to the target voices. It is worth noting that all the values in Fig.8 and Fig.9 are the average values of all the test data.

To prove the excellent effect of WaveNet, Chen *et al.* focused on the high quality of WaveNet synthesized voice compared to speech synthesis straightly [20]. WaveNet was reported to be superior to traditional speech synthesizer in terms of naturalness or similarity. Fig.10 and Fig.12 show the mean opinion scores (MOS) of the conversion methods mentioned above for similarity. Fig.11 and Fig.13 show the mean opinion scores (MOS) for naturalness.

As shown in Fig.10, Fig.11, Fig.12 and Fig.13, the proposed algorithm reached the highest average scores of MOS for naturalness and similarity among different databases, which demonstrated the effectiveness of the method for

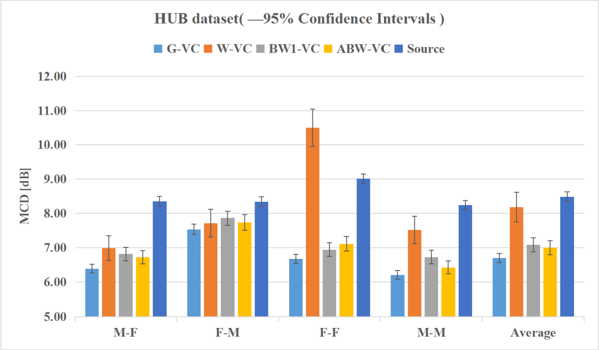


FIGURE 8. MCD scores of conversion on the HUB dataset, (F: Female, M: Male, F-M: Female to male conversion).

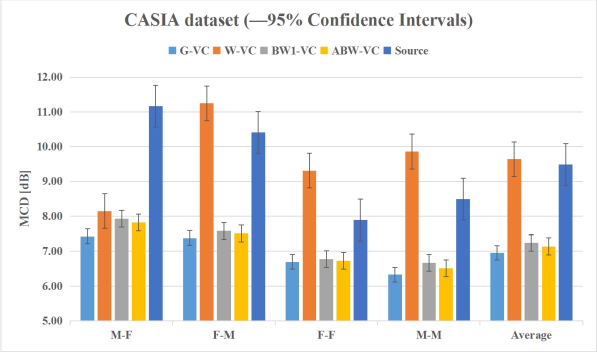


FIGURE 9. MCD scores of conversion on the CASIA dataset, (F: Female, M: Male, F-M: Female to male conversion).

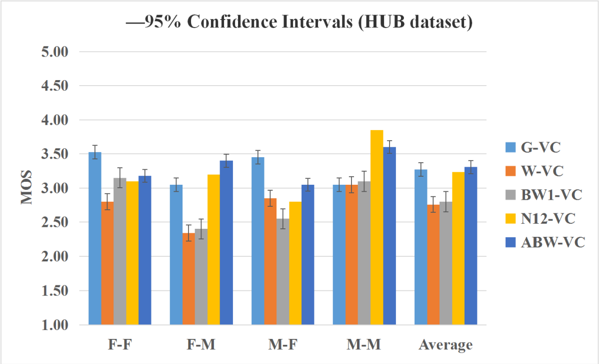


FIGURE 10. MOS of the converted voice from HUB dataset with the 95% confidence intervals for similarity (F: Female, M: Male).

both English and Mandarin. Although the conversion of F-F (female to female) was slightly poor, other metrics of the proposed VC system were obviously better than those from the baselines. Especially for the case of F-M (female to male), the performance was greatly improved. It can be seen that our method achieved an above-average accuracy for similarity in cross-gender and M-M (male to male). The improvement of F-F conversion is trivial, which may bring difficulties to distinguish the similarity of the converted voice and the target voice and the similarity of the target voice and the original voice.

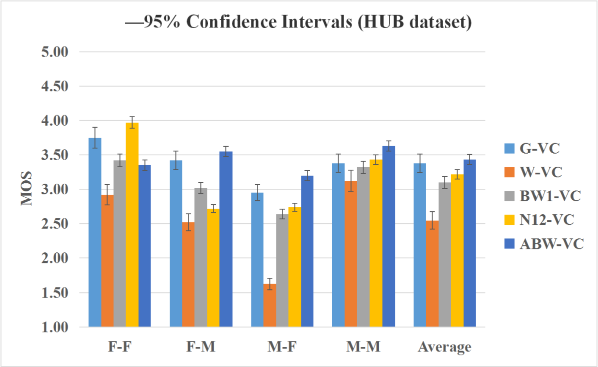


FIGURE 11. MOS of the converted voice from HUB dataset with the 95% confidence intervals for naturalness (F: Female, M: Male).

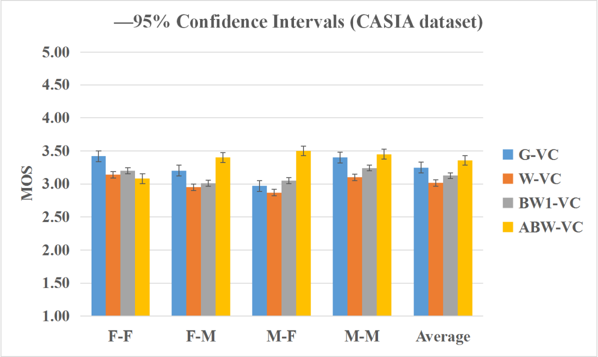


FIGURE 12. MOS of the converted voice from CASIA dataset with the 95% confidence intervals for similarity (F: Female, M: Male).

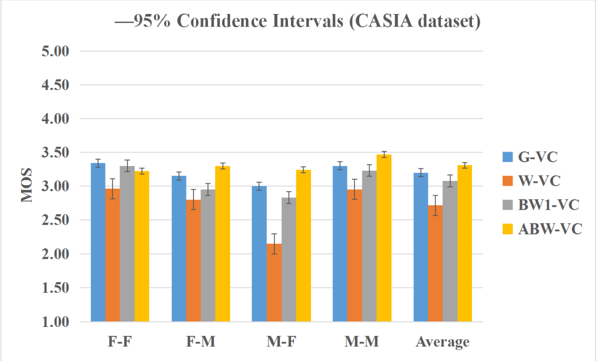


FIGURE 13. MOS of the converted voice from CASIA dataset with the 95% confidence intervals for naturalness(F: Female, M: Male).

At the same time, it is worth noting that the performance of ABW-VC was better than that of BW1-VC, which showed that the additional post-processing part improved the overall quality of the converted voice effectively.

VI. CONCLUSION

In this paper, a BLSTM and WaveNet based voice conversion method with waveform collapse suppression by post-processing was presented. We studied how to train the model of features conversion by BLSTM and how to realize the model of WaveNet with local conditional parameters. The WaveNet-converted voice was optimized through subsequent iterations to prevent waveform collapse. By comparing with several different methods of voice conversion on the Mandarin and English datasets, it was found that although the improvement of the proposed method in objective measures was trivial, it was able to improve the effect of subjective auditory greatly. Meanwhile, it made the converted voice sound more natural and fluent, especially in cross-gender voice conversions. At the same time, the proposed method of voice conversion took WaveNet with post-processing to solve the problem of over-smoothing caused by other speech synthesizers and to reduce the occurrence of waveform-collapsed speech effectively.

In a word, our method achieved a relatively high degree of naturalness and similarity comparing to other baseline methods on average. As future work, we will study the reason why the effect of conversion between female speakers was trivial, and keep optimizing the method of voice conversion to improve its performance.

REFERENCES

[1] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2001, pp. 301–304.

[2] J. Niwa, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Statistical voice conversion based on WaveNet," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5289–5293.

[3] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2005, pp. 9–12.

[4] Y. Stylianou, O. Cappê, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[5] D. Erro, A. Alonso, L. Serrano, E. Navas, and I. Hernáez, "Towards physically interpretable parametric voice conversion functions," in *Proc. 6th Adv. Nonlinear Speech Process. Int. Conf.*, 2013, pp. 75–82.

[6] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, E. S. Chng, and M. Dong, "Sparse representation for frequency warping based voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4235–4239.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de 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, 1999.

[8] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4869–4873.

[9] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong, and E. S. Chng, "High quality voice conversion using prosodic and high-resolution spectral features," *Multimedia Tools Appl.*, vol. 75, no. 9, pp. 5265–5285, 2016.

[10] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 954–964, Jul. 2010.

[11] A. van den Oord *et al.* (2016). "WAVENET: A generative model for raw audio." [Online]. Available: https://arxiv.org/abs/1609.03499

[12] Y.-C. Wu, K. Kobayashi, T. Hayashi, P. L. Tobing, and T. Toda. (2018). "Collapsed speech segment detection and suppression for WaveNet vocoder." [Online]. Available: https://arxiv.org/abs/1804.11055v2

[13] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of BLSTM Neural Networks for enhanced emotion classification in dyadic spoken interactions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4157–4160.

[14] F. A. Gers and J. Schmidhuber, "LSTM recurrent networks learn simple context-free and context-sensitive languages," *IEEE Trans. Neural Netw.*, vol. 12, no. 6, pp. 1333–1340, Nov. 2001.

[15] Z. Huang, W. Xu, and K. Yu. (2015). "Bidirectional LSTM-CRF models for sequence tagging." [Online]. Available: https://arxiv.org/abs/1508.01991

[16] M. Coto-Jiménez and J. Goddard-Close, "LSTM deep neural networks postfiltering for improving the quality of synthetic voices," in *Proc. Mexican Conf. Pattern Recognit.*, 2016, pp. 280–289.

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

[18] Y. Zhao, S. Takaki, H.-T. Luong, J. Yamagishi, D. Saito, and N. Minematsu, "Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a WaveNet vocoder," *IEEE Access*, vol. 1, pp. 60478–60488, 2018.

[19] C. Liu and D. Kewley-Port, "STRAIGHT: A new speech synthesizer for vowel formant discrimination," *Acoust. Res. Lett. Online*, vol. 5, no. 2, pp. 31–36, 2004.

[20] K. Chen, B. Chen, J. Lai, and K, Yu, "High-quality Voice Conversion Using Spectrogram-Based WaveNet Vocoder," in *Proc. Interspeech*, 2018, pp. 1993–1997.

[21] Y. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU Non-Parallel Voice Conversion System for the Voice Conversion Challenge 2018," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018.

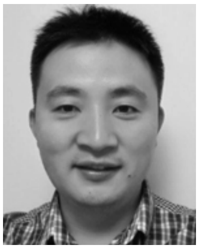
[22] W. J. Han, H. F. Li, H. B. Ruan, and L. Ma, "Review on speech emotion recognition," *J. Softw.*, vol. 25, no. 1, pp. 37–50, 2014.

[23] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, 2018, pp. 1983–1987.

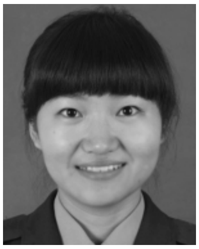
[24] J. Lorenzo-Trueba *et al.* (2018). "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods." [Online]. Available: https://arxiv.org/abs/1804.04262

XIAOKONG MIAO received the B.S. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, in 2014, and the M.S. degree from the Department of Electronics and Optical Engineering, Ordnance Engineering College, Shijiazhuang, China, in 2017. He is currently pursuing the Ph.D. degree with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing, China. His research interests include signal processing, voice conversion, and machine learning.

XIONGWEI ZHANG received the Ph.D. degree in signal and information processing from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1992. He is currently a Professor with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing. His research interests include speech signal processing, machine learning, and pattern recognition.



MENG SUN received the Ph.D. degree from the Department of Electrical Engineering, Katholieke University Leuven, in 2012. He is currently an Associate Professor with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing, China. His research interests include speech processing, unsupervised/semi-supervised machine learning, and sequential pattern recognition.



CHANGYAN ZHENG received the B.S. and M.S. degrees from the Department of Electronics and Optical Engineering, Ordnance Engineering College, Shijiazhuang, China, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing, China. Her research interests include signal processing, speech enhancement, and machine learning.

TIEYONG CAO received the Ph.D. degree from the Institute of Communications Engineering, PLA University of Science and Technology, Nanjing, China, in 2004. He is currently a Professor with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing. His research interests include signal processing, machine learning, and image processing.