Voice Conversion by Dual-Domain Bidirectional Long Short-Term Memory Networks with Temporal Attention

Xiaokong MIAO Meng SUN and Xiongwei ZHANG
Army Engineering University, Nanjing, China
E-mail: miao_xk@163.com, sunmengccjs@gmail.com, xwzhang9898@163.com

Abstract— **Voice conversion(VC) is a method for seeking to convert one speaker's voice into another person's voice while maintaining the content unchanged. One of the key steps is to construct a mapping of features from the source speaker to the target speaker. Given the strong ability to model contextual information, bidirectional long short-term memory network (BLSTM) is usually taken as the mapping tool. In this paper, an improved version of BLSTM, dual-domain BLSTM, is considered as a baseline where dynamic time warping (DTW) is conventionally taken as a tool to align speech frames. In order to alleviate the negative impacts of temporal misalignment in DTW, a casual attention mechanism is introduced to improve the dual-domain BLSTM. Experiments demonstrated the effectiveness of the proposed approach by yielding lower mel-ceptral distances and higer MOS scores than the baselines.**

I. Introduction

Voice conversion (VC) aims to modify the speech of one speaker to make it sound like another specific speaker while keeping its linguistic information unchanged [1]. 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 [2]. In recent years, the research of machine learning has contributed many solutions to solving the problems in voice conversion such as deep neural networks (DNN) [3-4], Gaussian mixture models (GMM) [5-6], long short-term memory networks (LSTM) and its bidirectional version BLSTM, and so on. These models have performed as effective non-linear mapping approaches for voice conversion [7-8]. Given the sequential nature of speech, recurrent models are usually adopted to model the contextual information of spectral features, especially for LSTM and BLSTM. In this paper, BLSTM is taken as the basic model given the fact that it is good at modeling long-range sequential information from both directions with the help of its memory blocks and peephole connections [9]. BLSTM has been reported achieving superior performance to competitive baselines mentioned above [10].

As traditional supervised learning, the voice conversion recipes consist of two stages: a training stage and a conversion stage. During the training stage, parallel VC training firstly requires to make the alignment of speech frames between the source and target speakers, usually by using dynamic time warping (DTW). Subsequently, the mapping from source features to target features is modeled by BLSTM. During the conversion stage, the mapping function of BLSTM is applied on features extracted from the new input voice of the source speaker to yield converted voices [11].

In this paper, two aspects of improvements of parallel voice conversion will be studied. The first one is to introduce a dual-domain BLSTM, which is able to realize the conversion between source and target speakers mutually, by training only once. The second one is to alleviate the alignment errors from DTW by introducing temporal attention to the dual-domain voice conversion.

The rest of this paper is organized as follows. Section 2 introduces the baseline model BLSTM and its dual-domain version in details. Section 3 presents the proposed approach. The setup of experiments and analysis of results are given in Section 4. Finally, the conclusion is given in Section 5.

II. Dual-domain BLSTM and our motivation

A typical method of voice conversion by using BLSTM is given in Fig.1. In this method, BLSTM has been used for voice conversion as a feature-mapping model [12], whose main idea is to introduce an adaptive gating mechanism to decide to which degree the LSTM units keeping the previous state and memorizing the extracted features of the current data frame [13].

During the training stage, fundamental frequency (F0), spectral envelope and aperiodic frequency (AP) are extracted from the utterances of both source and target speakers. These parameters are used for modeling voice conversion given the fact that high quality speech can be synthesized from these parameters. Among these features, Mel-cepstral coefficients (Mcep) are extracted to represent spectral envelope. Subsequently, parallel Mcep sequences of the source and target speech for training are aligned through DTW. Then, the aligned source and target mceps are used as the input and output features to train the BLSTM model. During the conversion stage, given the Mcep features from a new utterance of the source speaker, its corresponding converted Mcep features are predicted by the learnt BLSTM. Finally, the converted voice is synthesized according to the predicted Mcep features, the transformed log F0 and the copied aperiodic frequency.

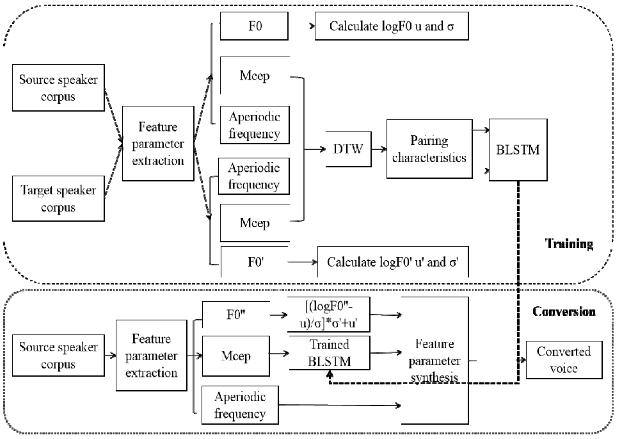

Fig.1 The architecture of a typical BLSTM-based voice conversion method, where AP, Mcep, F0 and DTW represent aperiodic frequency, mel-cepstrum, pitch period and dynamic time warping, respectively [11].

As an extension of BLSTM presented above, the basic idea of dual-domain BLSTM is given in Fig.2, where the data flows have two directions: upward and downward. The dual-domain BLSTM is designed to jointly learn temporal relationship between two data streams, by utilizing the power of BLSTMs in learning both short-term and long-term dependencies. It uses two BLSTM networks (a top one and a bottom one) to achieve this, where the shared units in the middle part bridge the gap of the two BLSTMS as shown in Fig.1.

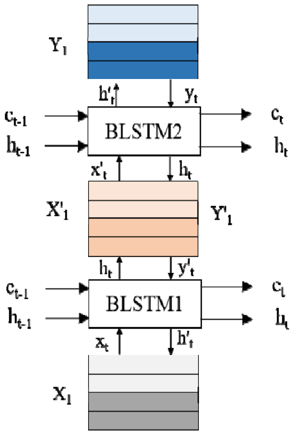

Fig.2 A dual-domain BLSTM trained jointly on dataset

In this paper, dual-domain BLSTM is chosen as our basic model. Dual-domain BLSTM is able to jointly learn temporal dependencies from two domains: both independent temporal dependencies per domain, as well as common cross-domain temporal dependencies. This model has been successfully applied to sequence prediction and classification tasks [14-16].

III. Dual-domain attention-based BLSTM for voice conversion

The dual-domain BLSTM network can realize the conversion between speakers for each other. However, there are some decreases in the quality of converted voice compared with its single-domain counterpart, i.e. conventional BLSTM. In order to improve the performance of dual-domain BLSTM, an attention-based dual-domain BLSTM for voice conversion is proposed.

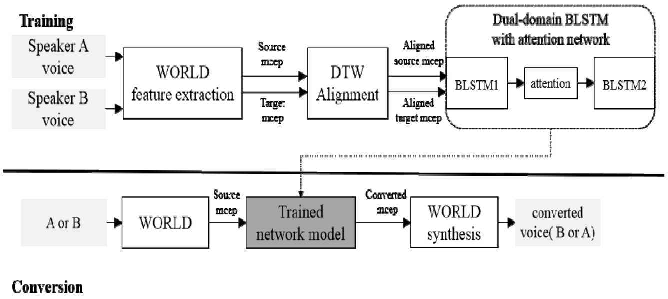

Fig.3 Schematic diagram of VC for dual-domain BLSTM with attention. It is to alleviate the misalignment error caused by DTW by introducing an attention layer. Meanwhile, the retaining of DTW can accelerate the convergence of the training iterations of the attention layer. WORLD [17] is chosen as the vocoder in this diagram.

A. The general recipe

As illustrated in Fig. 3, the proposed approach is divided into two stages: training stage and conversion stage. During the training stage, speaker A and speaker B are used to denote the involved speakers, instead of using source speaker and target speaker, because in the dual-domain BLSTM model, the speakers are mutually transformed and any speaker can be the source one or the target one.

In the training stage, both A and B are used as both source training and target training to train the network. In the conversion stage, any person selected as the voice to be converted, either for A or for B, can generate the corresponding converted voice. BLSTM1 and BLSTM2 are two BLSTM networks with the same structure. BLSTM1 is the first network from speaker A to speaker B, while BLSTM2 vice versa. The two BLSTM networks are connected and temporally adjusted by an attention layer, as will be presented below.

B. Dual-domain BLSTM with attention mechanism

Attentive neural networks have recently demonstrated success in a wide range of tasks ranging from question answering, machine translations, speech recognition, to image captioning [18-20].

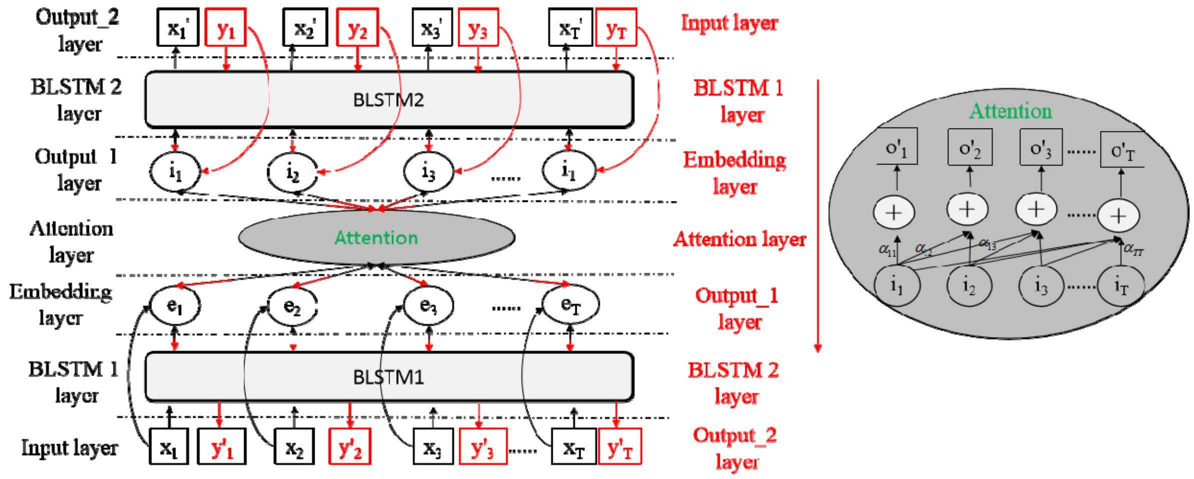

Fig.4 Dual-domain BLSTM model with attention. In attention layer, considering the mapping of sequence to sequence, the idea of causal weighting is adopted. Right ellipse indicates the internal structure of the attention layer, where α_{ij} represents the weight value of the *i*th frame input sequence to the *j*th frame output sequence.

In this section, we implement the attention mechanism to the sequence-to-sequence task of voice conversion. Although DTW provides a relatively accurate alignment of speech frames of any two parallel utterances, there can still have errors of misalignments. Therefore, an attention mechanism is introduced to further adjust the temporal mapping between parallel utterances to reduce its negative impacts. Let H be a matrix consisting of output vectors $[h_1, h_2, \ldots, h_T]$ that the BLSTM layer produced, where T is the number of frames in this sequence. The symbol r is the weighted sum of the output vectors in H and α is the local weights of the frames [13]:

$$M = \tanh(H) \tag{1}$$

$$\alpha = softmax\left(w^T M\right) \tag{2}$$

$$r = H\alpha^T \tag{3}$$

where $H \in \Re^{d^w \times T}$, d^w is the dimension of the mcep, w is a trained parameter vector and w^T is its transpose. The dimension of w, α, r *is* d^w, T, d^w separately [13].

As shown in Fig.4, the dual-domain BLSTM with attention network contains three sub-networks: BLSTM1 BLSTM2 and the attention layer. Suppose the input vector is X= $[x_1,x_2,x_3,\cdots,x_T]$, and E=$[e_1,e_2,e_3,\cdots,e_T]$ is the data passes through BLSTM1, I=$[i_1,i_2,i_3,\cdots,i_T]$ is the data passes through the attention layer. X= $[x'_1,x'_2,x'_3,\cdots,x'_T]$ is the final output from BLSTM2. If the input vector is Y= $[y_1,y_2,y_3,\cdots,y_T]$, the training process is shown by the red arrow in Fig.4. The data passes through BLSTM2, the attention layer and BLSTM1 sequentially. The training of outputs and other parameters remains unchanged.

IV. Experiments

A. Dataset and Experimental Setup

In our experiments, we use CMU ARCTIC corpus [21]. Given a fact that a dual-domain voice conversion is studied in this paper, we focused on comparing the results of inter-gender conversion. Speech signals were sampled at 16kHz and the analysis window was 25ms with 5ms frame shift. WORLD was used to extract spectral envelope, aperiodic component (AP) and LogF0. 25-dim mceps extracted from the spectral envelope were converted by the proposed voice conversion method. LogF0 was linearly converted and then the AP of source voice was copied to synthesize the converted voice[22]. For the training of Dual-domain attention-based BLSTM for voice conversion, 956 and 107 parallel sentences were selected as the training and validation data. 55 parallel sentences were used as testing data during conversion.

The proposed voice conversion model was composed of two BLSTMs and one attention layer. The architectures of the two BLSTMs were the same. There were two hidden layers in the network where each hidden layer was a bidirectional LSTM [23]. In each layer, the number of units was set to [25, 128, 128, 25]. The proposed voice conversion model was trained by back-propagation through time algorithm and optimized by stochastic gradient descent. The learning rate was adjusted automatically with the model optimization, and the range of change was from 1.0×10^{-3} to 1.0×10^{-5}.

In order to prove the effectiveness of the proposed method, the following two methods were selected for comparison: GMM-VC, Dual-domain BLSTM-VC.

GMM-VC (**G-VC**): A conventional GMM-based voice conversion system.

Dual-domain BLSTM-VC (**DB-VC**): A dual-domain BLSTM without attention layer.

Dual-domain with attention BLSTM-VC (**DAB-VC**): i.e. the voice conversion method proposed in this paper.

B. Experimental Results

Objective and subjective tests were conducted on the CMU ARCTIC corpus. The objective measurement is the mel cepstral distortion(MCD), which is defined as,

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{I}\left(m_i^{con} - m_i^{tar}\right)^2} \qquad (4)$$

where m_i^{con} and m_i^{tar} are the mel cepstral coefficients of converted features and target features, respectively [24-25]. I refers to the total number of frames in a sentence.

MCD is the objective evaluation metric of voice conversion. The smaller the MCD value, the closer the converted voice is to the target voice.

The MCD scores of the above three systems for inter-gender voice conversion are summarized in Table 1.

Table 1: The MCD of the aforementioned three different systems. (F: female, M: male. bdl and slt represent two different speakers, respectively.)

Conversion	F-M	M-F
Speaker-pair	slt-bdl	bdl-slt
Source-Target	8.45	8.45
G-VC	6.32	6.21
DB-VC	6.89	6.73
DAB-VC	6.18	6.04

To evaluate the quality and similarity of the converted speech from these three systems, we conduct a Mean Opinion Score (MOS) test for voice conversion results.

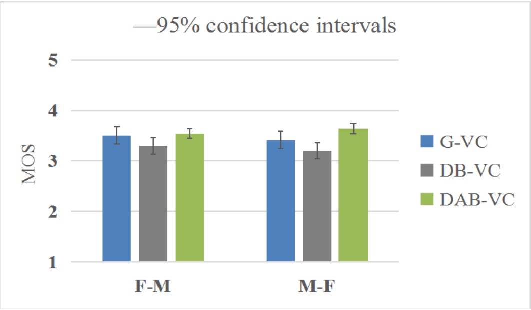

Fig. 5 MOS test results with the 95% confidence intervals. (5-point scale: 5: excellent, 4: good, 3: fair, 2:poor, 1: bad.)

Results show that our proposed approach can achieve superior performance than some baseline methods in terms of quality and similarity of the converted speech.

V. Conclusions

In this paper, a dual-domain BLSTM networks with attention based voice conversion method was presented. Comparing with traditional based GMM and basic BLSTM voice conversion methods, the proposed model realized the mutual conversion between any two speakers, and also improved the quality of the converted speech.

In future, neural vocoder such as WaveNet or LPCNet will be considered to further improve the subjective quality of the converted voice, together with the approach proposed in this paper.

The source codes regarding the approach will come soon. For the reader's reference, some audio samples for the subjective listening tests are accessible via this link:

https://github.com/miaoxk/dual_domain-attention-VC.

Acknowledgment

The work is supported by the Natural Science Foundation of Jiangsu Province (BK20180080) and the Natural Science Foundation of China (61471394).

References

[1] Tanaka, Kou , et al. "AttS2S-VC: Sequence-to-Sequence Voice Conversion with Attention and Context Preservation Mechanisms". 2018.

[2] J. Niwa, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Statistical voice conversion based on wavenet", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5289–5293, 2018.

[3] 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 and Applications, vol. 75, no. 9, pp. 5265–5285, 2016.

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

[5] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9–12, 2005

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[7] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in Proc. INTERSPEECH. ISCA, 2013, pp. 369–372.

[8] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," Trans. Audio, Speech & Language Processing, vol. 22, no. 12, pp.1859–1872, 2014.

[9] S. Hochreiter and J. "Schmidhuber, Long short-term memory". Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] D. Huang, L. Xie, Y. S. W. Lee, et al., "An automatic voice conversion evaluation strategy based on perceptual background noise distortion and speaker similarity," in Proc. SSW. ISCA, 2016, pp. 46–53.

[11] X. K. Miao, X. W. Zhang, M. SUN, et al..“A BLSTM and WaveNet-Based Voice Conversion Method With Waveform Collapse Suppression by Post-Processing” online:https://ieeexplore.ieee.org/document/8695725?source=authoralert

[12] L. Sun, S. Kang, K. Li, and H. Meng. “Voice Conversion Using Deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks”, IEEE International Conference on Acoustics, Speech and Signl Processing (ICASSP), pp. 4869–4873, 2015.

[13] Zhou P , Shi W , Tian J , et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation, Classification[C]// Meeting of the Association for Computational Linguistics. 2016.

[14] T. Perrett, D Damen. “DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition”. arXiv:1904.08634,2019

[15] Qin Y, Song D, Chen H, et al. “A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction”. arXiv:1704.02971v4. 2017.

[16] Y Liu, C .Y Gong, L Yang, et al. “DSTP-RNN: a dual-stage two-phase attention-based recurrent neural networks for long-term and multivariate time series prediction”. arXiv:1904.07464. 2019.

[17] M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 1877–1884, 2016.

[18] Bahdanau D , Cho K , Bengio Y . “Neural Machine Translation by Jointly Learning to Align and Translate”. Computer Science, 2014.

[19] Hermann, Karl Moritz , et al. “Teaching Machines to Read and Comprehend.” 2015.

[20] Chorowski, Jan , et al. "Attention-Based Models for Speech Recognition." Computer Science 10.4(2015):429-439

[21] J. Kominek and A. W. Black, “The cmu arctic speech databases,” in Proc. SSW. ISCA, 2004, pp. 223–224.

[22] Sun L , Li K , Wang H , et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training[C]// IEEE International Conference on Multimedia & Expo. IEEE, 2016.

[23] Jie Wu, Dongyan Huang, Lei Xie and Haizhou Li, "Denoising Recurrent Neural Network for Deep Bidirectional LSTM based Voice Conversion", Interspeech 2017, Stockholm, Sweden, August 20-24

[24] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in Communications, Computers and Signal Processing, vol. 1. IEEE, 1993, pp. 125–128.

[25] Y. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, “The NU NonParallel Voice Conversion System for the Voice Conversion Challenge 2018”, Odyssey 2018 The Speaker and Language Recognition Workshop, 2018.