A COMPACT FRAMEWORK FOR VOICE CONVERSION USING WAVENET CONDITIONED ON PHONETIC POSTERIORGRAMS

Hui Lu[1,2], *Zhiyong Wu*[1,2,3,*], *Runnan Li*[1,2], *Shiyin Kang*[4], *Jia Jia*[1,2], *Helen Meng*[1,3]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[2]Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[4]Tencent AI Lab, Tencent, Shenzhen, China
{lu-h17, lirn15}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk
shiyinkang@tencent.com, jjia@tsinghua.edu.cn

ABSTRACT

Voice conversion can benefit from WaveNet vocoder with improvement in converted speech's naturalness and quality. However, nowadays approaches segregate the training of conversion module and WaveNet vocoder towards different optimization objectives, which might lead to the difficulty in model tuning and coordination. In this paper, we propose a compact framework to unify the conversion and the vocoder parts. Multi-head self-attention structure and bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) are employed to encode speaker independent phonetic posteriorgrams (PPGs) into an intermediate representation which is used as the condition input of WaveNet to generate target speaker's waveform. In this way, we unify the conversion and vocoder parts into a compact system in which all parameters can be tuned simultaneously for global optimization. We compared the proposed method with the baseline system that consists of separately trained conversion module and WaveNet vocoder. Subjective evaluations show that the proposed method can achieve better results in both naturalness and speaker similarity.

Index Terms— Voice conversion, WaveNet, phonetic posteriorgrams(PPGs), self-attention, BLSTM

1. INTRODUCTION

Voice conversion (VC) is a technique to modify the speech from source speaker to make it sound like being uttered by target speaker while keeping the linguistic content unchanged.

Traditional VC methods mainly consist of two key components: conversion function and vocoder [1]. The conversion function converts features extracted from the source speaker's speech into acoustic features of the target speaker. Then the vocoder uses these converted features to synthesize speech waveform of the target speaker. However, traditional conversion functions rely heavily on parallel corpus [2-4] and time-alignment. Conventional vocoders [5,6] are of low quality and not robust. These restrictions greatly hindered the performance of traditional VC methods. Many approaches have been proposed to overcome these limitations [7-9]. Two of the most remarkable works are the introducing of PPGs to facilitate non-parallel VC [10] and the utilization of WaveNet vocoder [11] to improve the speech quality [12,13].

Motivated by the above progress, the N10 system [14] in the Voice Conversion Challenge 2018 (VCC 2018) [15] has proposed to use both PPGs-based non-parallel data conversion function and WaveNet vocoder for VC and achieved both high naturalness and high speaker similarity of the converted speech. However, in N10 system, the conversion part and the WaveNet vocoder are separately trained and tuned, which may hinder the global optimization of the whole system. Moreover, the method utilizes STRAIGHT [6] spectral features as the intermediate features bridging the two parts. Predicting the waveform from such human defined features may not achieve the best performance. Furthermore, training and tuning of two systems are usually more time-consuming and laborious than a single system.

In this paper, we propose a method to unify the conversion function and the vocoder. We first extract speaker independent linguistic features from the source speaker's speech. Then instead of converting the source speaker's linguistic features into the target speaker's acoustic features, we employ a combination of multi-head self-attention structure [16] and BLSTM [17] structure to encode the

* Corresponding author

linguistic features and F0. The encoded features are up-sampled into the desired resolution and fed into WaveNet to generate the target speaker's waveform. Consequently, our system doesn't need intermediate acoustic features. Since the training data can be extracted only from the target speaker's corpus, we don't need a parallel corpus and to do time-alignment for the data. We train only one system in which all the parameters are tuned simultaneously to optimize the generation of the target speaker's waveform. Experiments show that the proposed method can achieve better performance than the baseline system in both speech naturalness and speaker similarity.

The rest of this paper is organized as follows. Section 2 describes a state-of-the-art nonparallel VC system as our baseline system. Section 3 describes our proposed method. Experimental setup and results are presented in section 4. Section 5 concludes this paper.

2. BASELINE SYSTEM

Our baseline system is mainly based on the VC system using PPGs [10]. However, instead of using STRAIGHT vocoder as in [10], we train a WaveNet vocoder conditioned on acoustic features to generate the target speaker's speech waveform. The baseline system's framework is similar to that of the N10 system in the VCC 2018, however, we didn't use the adaptation method to train the WaveNet vocoder as they did [14]. The main architecture of the baseline system is depicted in Fig.1.

2.1. Phonetic posteriorgrams (PPGs)

PPGs are generally considered as speaker independent features containing linguistic information. A PPG is a time sequence representing the posterior probability of each senone for each time frame of the utterance. The number of frames each senone lasts reveals the duration information of the senone. Though we use a speaker-independent automatic speech recognition (SI-ASR) system to extract PPGs from an utterance, we don't convert PPGs into the corresponding phoneme or word sequences for two reasons: 1) PPGs can capture the composition of different senones in a certain sound and its subtle temporal changes, which can be utilized to obtain more accurate estimates of acoustic parameters. 2) The duration information may be damaged due to mapping different posterior probability distributions into one phoneme.

2.2. Framework overview

The training of the baseline system involves three stages. In training stage 1, the SI-ASR system is trained on a separate ASR corpus to serve as a PPGs extractor. In training stage 2, a BLSTM conversion function is trained to convert PPGs into Mel-cepstral coefficients (MCEPs) frame by frame. In training stage 3, the WaveNet vocoder conditioned on MCEPs and logarithm F0 is trained. The SI-ASR system trained in stage 1 is used in stage 2 to obtain the PPGs representation of the speech. The up-sample layer before the WaveNet vocoder up-sample the condition input to match the speech waveform's time resolution. While the WaveNet vocoder in Fig.1 is trained on the ground truth MCEPs, it can also be trained on the MCEPs predicted by the trained BLSTM conversion function; however, in the latter case, the trained WaveNet vocoder may generate speech with more serious quality degradation according to our experiments.

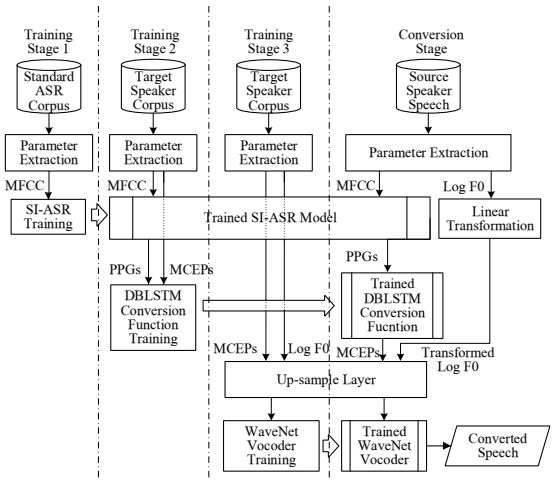

Fig.1: The framework of the baseline system.

During conversion, the source speaker's speech is first transformed into its PPGs representation via the SI-ASR model. The trained BLSTM conversion function converts the PPGs into the MCEPs which is concatenated with the logarithm F0 to form the local condition input to WaveNet vocoder. The WaveNet vocoder finally generates the converted speech waveform. The logarithm F0 of the source speaker is transformed into that of the target speaker simply by a linear transformation.

2.3. Limitations

Despite the good performance of the baseline system, it still has the following limitations: 1) The BLSTM conversion function and the WaveNet vocoder are trained separately. Hence parameters of the two parts are not tuned jointly to optimize the generation of the waveform. Besides, errors from each part may compound. 2) The utilization of MCEPs as intermediate features may not be consistent with the whole system's optimization objective. 3) It is time-consuming to train and tune two separate neural network architectures.

3. PROPOSED METHOD

To overcome the above limitations, we propose a modified conditional WaveNet based on self-attention and BLSTM to unify the conversion part and the vocoder part for VC.

3.1. Framework overview

As shown in Fig.2, the training process of the proposed method involves two stages. The training stage 1 is the same as in the baseline system. In training stage 2, the network being trained is a unification of a conversion system and a vocoder system. There is no intermediate process of mapping

PPGs into acoustic features. We directly train a WaveNet synthesizer conditioned on PPGs and logarithm F0. We call it a WaveNet synthesizer because the WaveNet architecture is conditioned mainly on linguistic features (i.e. PPGs).

The conversion process is also simpler compared to that in the baseline system. For an arbitrary source speaker's speech, we extract its PPGs representation through the trained SI-ASR model. The PPGs and the linear-transformed logarithm F0 are fed into the condition network, and then the WaveNet synthesizer generates the target speaker's speech waveform according to the condition input.

3.2. Condition network

In speech synthesis systems, the contextual information of linguistic features is important for producing natural and high-quality speech. The main reason is that the same phoneme or syllable may sound differently in different contexts due to the co-articulation effects. Traditional statistical parametric speech synthesis system would use a series of human-defined rules to model the contextual information of linguistic features [18]. For example, in the phoneme level, the preceding and succeeding two phonemes are added to the contexts; while in the phrase level, the position of the current phrase in major phrases is taken into consideration. Recent end-to-end Text-to-Speech (TTS) system [19] employed a bidirectional RNN architecture to model the linguistic features' contextual information. While PPGs can represent fine details of speech sounds and their subtle changes, they are bad at capturing coarse-grained information such as phrase level contexts. We need to explore an approach to take advantage of the contextual information of PPGs to improve their representation ability.

In the proposed method, we modify the up-sample layer in the baseline system to a more sophisticated network as shown in Fig.3. We call it condition network since it is a preprocessing of the WaveNet synthesizer's condition input. The condition network consists of two layer blocks, each includes a self-attention layer and a BLSTM layer. Self-attention and BLSTM are both state-of-the-art approaches for sequence modeling. However, self-attention and BLSTM process a sequence in quite different ways. Consider the procedure of computing output at a certain time step from a sequence, self-attention would directly attend to all elements in the sequence to compute a set of attention weights and obtain the output as a weighted sum of all these input elements [16]. While BLSTM would take in only the current input element and context information vectors from both directions to compute the output. We believe that the self-attention structure can easily capture the global context information far from the current linguistic frame while the BLSTM structure is more capable of acquiring the local context information near the current linguistic frame. The combination of the two structures can encode the PPGs into effective representations with sufficient context information to synthesize speech. We add a residual connection between the self-attention layer's input and the output as well as a layer normalization [20] operation after the self-attention layer to accelerate the training process. Experiments show that the modified conditional WaveNet can improve the converted speech's naturalness and target speaker similarity.

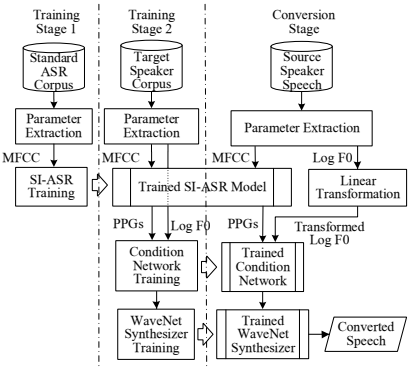

Fig.2: The framework of the proposed method.

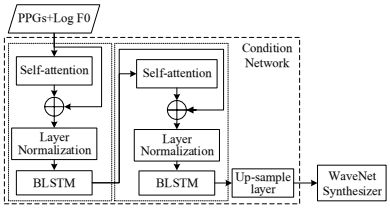

Fig.3: The condition network in the proposed method.

4. EXPERIMENTS

4.1. Experimental setup

To evaluate the proposed method and the baseline system, we used datasets of four American English speakers from the CMU ARCTIC [21]: two males (RMS, BDL) and two females (SLT, CLB). Among them, RMS and SLT are used as target speakers, BDL and CLB are used as source speakers, to test different VC system's performance.

All the acoustic features of speech are extracted with 25-ms window length and 5-ms window shift. The SI-ASR system for PPGs extraction is implemented using the Kaldi speech recognition toolkit [22] and trained on TIMIT corpus [23]. The PPG is extracted as a 128-dimensional data sequence representing probabilities of each 128 senones on all time frames of an utterance.

For the baseline system, the BLSTM conversion function consists of two layers, each layer is composed of a forward LSTM-RNN and a backward LSTM-RNN with their outputs concatenated. The numbers of hidden units of these LSTM-RNNs are all 64. The WaveNet architecture has two dilated blocks, each consists of 10 layers with the dilation rate from 1 to 512. The numbers of residual channels and skip channels are 128 and 256, respectively. The waveform is μ-law encoded into 256-dimension. The WaveNet architecture is the same for both the baseline system and the proposed method. The up-sample layer in both systems is just an operation in which different frames of the input are repeated multiple times to match the resolution of the target waveform.

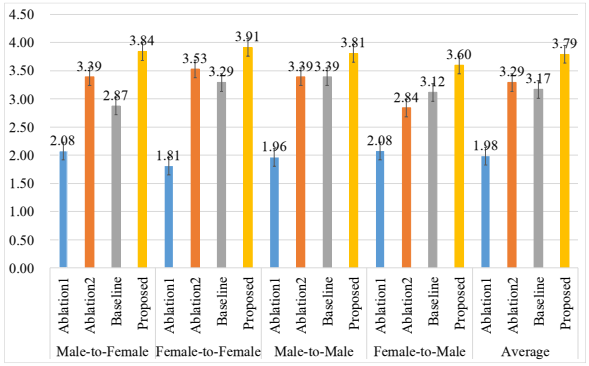

Fig.4: Naturalness results for 4 systems and 4 types of conversion tasks.

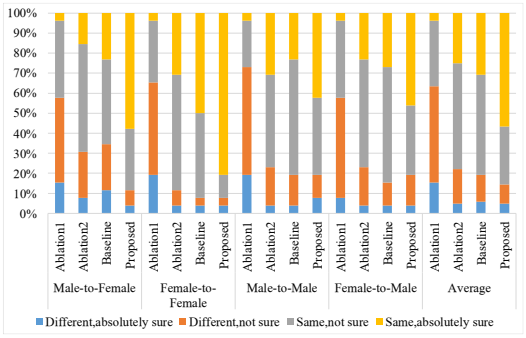

Fig.5: Similarity results of target speaker for 4 systems and 4 types of conversion tasks.

For the condition network in the proposed method, the numbers of hidden units in the two self-attention structures are 130 and 128, respectively, and numbers of heads are 13 and 8, respectively. The numbers of hidden units of the LSTM-RNNs in two BLSTM layers are both 128.

Four systems were evaluated in the experiments. In addition to the baseline system (***Baseline***) and the proposed system (***Proposed***), we also conducted two ablation studies on the proposed method to evaluate the effectiveness of its network structure. Ablation test system 1 (***Ablation 1***): which comes from the proposed system but without condition network, which means the WaveNet synthesizer is directly conditioned on the raw PPGs and the logarithm F0. Ablation test system 2 (***Ablation 2***): which comes from the proposed system but without self-attention architecture, i.e. the condition network only consists of a two-layer BLSTM.

4.2. Subjective evaluation

We build VC systems for two target speakers: a male speaker (RMS) and a female speaker (SLT). Subjective listening tests for four types of conversion are conducted: male-to-female (BDL to SLT), female-to-female (CLB to SLT), male-to-male (BDL to RMS) and female-to-male (CLB to RMS). The naturalness and speaker similarity are evaluated perceptually on the four types of conversion pairs.

Followed voice conversion challenge 2016 [24], we conducted Mean Opinion Score (MOS) listening test for naturalness and use Same/Different paradigm to measure the speaker similarity. In naturalness MOS listening tests, subjects were asked to evaluate the converted speech samples on a scale from 1 (completely unnatural) to 5 (completely natural). In similarity evaluation tests, the scale for judging is: "Same, absolutely sure", "Same, not sure", "Different, not sure" and "Different, absolutely Sure".

Four sentences were converted by each of the 4 systems in 4 types of conversion case. Hence, 64 utterances in total were evaluated[1]. 12 native Chinese speakers without listening impairment participated in the evaluation tests. Each utterance is presented to at least 6 subjects.

[1] Samples: https://light1726.github.io/voice_conversion_demo/

4.3. Experimental results

The results of the naturalness MOS and the similarity evaluation tests are depicted on Fig.4 and Fig.5 respectively. As can be seen, the proposed method outperforms the baseline system in both speech naturalness and speaker similarity, indicating the effectiveness of the proposed compact framework. Without the condition network, the ablation test system 1 performs badly in both naturalness and speaker similarity evaluations. We owe the great performance improvement from the ablation test system 1 to the proposed method to the condition network that effectively models the contextual information of PPGs. The ablation test system 2 also surpass the ablation test system 1 due to the utilization of BLSTM to model the context information. Although ablation test system 2 achieves comparable performance in both naturalness and speaker similarity when compared with the baseline system, its performance is still worse than that of the proposed method, which indicates that the using of multi-head self-attention can make up for the BLSTM in modeling the global context. Thanks to the condition network, all phonemes can be synthesized accurately.

5. CONCLUSION

In this paper, we propose a compact framework for voice conversion based on WaveNet conditioned on PPGs. The proposed method unifies the feature conversion function and the WaveNet vocoder. We propose a condition network based on self-attention and BLSTM to encode the PPGs into an effective condition input to the WaveNet. Subjective evaluations show that the proposed method outperforms traditional two-step VC methods and the condition network is effective for our VC methods.

Acknowledgements: This work is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N CUHK404/15), National Natural Science Foundation of China (61433018, 61375027). We would also like to thank Tencent AI Lab Rhino-Bird Joint Research Program (No.JR201803) and Tsinghua University - Tencent Joint Laboratory for the support.

6. REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] 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.

[3] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplarbased sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[4] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," [in] *Proc. ICASSP, IEEE*, pp. 4869–4873, 2015.

[5] 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.

[6] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[7] D. Erro, A. Moreno, and A. Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[8] F. Fang, J. Yamagishi, I. Echizen, and J. LorenzoTrueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," [in] *Proc. ICASSP*, IEEE, pp. 5279–5283, 2018.

[9] C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, and H. M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," [in] *Proc. APSIPA*, IEEE, pp. 1–6, 2016.

[10] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," [in] *Proc. ICME*, IEEE, pp. 1–6, 2016.

[11] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. w. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," [in] *Proc. SSW*, p. 125, 2016.

[12] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," [in] *Proc. Interspeech*, vol. 2017, pp. 1138–1142, 2017.

[13] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based WaveNet vocoder," [in] *Proc. Interspeech*, pp. 1993–1997, 2018.

[14] 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*, pp. 1983–1987, 2018.

[15] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," [in] *Proc. Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

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

[18] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[19] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," [in] *Proc. ICML*, 2017.

[20] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[21] J. Kominek and A. W. Black, "The cmu arctic speech databases," [in] *Fifth ISCA workshop on speech synthesis*, 2004.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," [in] *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, number EPFLCONF-192584, 2011.

[23] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[24] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," [in] *Proc. Interspeech*, pp. 1632–1636, 2016.