INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Jointly Trained Conversion Model and WaveNet Vocoder for Non-parallel Voice Conversion using Mel-spectrograms and Phonetic Posteriorgrams

Songxiang Liu[1], *Yuewen Cao*[1], *Xixin Wu*[1], *Lifa Sun*[2], *Xunying Liu*[1], *Helen Meng*[1]

[1]Human-Computer Communications Laboratory,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[2]SpeechX Limited, Shenzhen, China

{sxliu, ywcao, wuxx, lfsun, xyliu, hmmeng}@se.cuhk.edu.hk

Abstract

The N10 system in the Voice Conversion Challenge 2018 (VCC 2018) has achieved high voice conversion (VC) performance in terms of speech naturalness and speaker similarity. We believe that further improvements can be gained from joint optimization (instead of separate optimization) of the conversion model and WaveNet vocoder, as well as leveraging information from the acoustic representation of the speech waveform, e.g. from Mel-spectrograms. In this paper, we propose a VC architecture to jointly train a conversion model that maps phonetic posteriorgrams (PPGs) to Mel-spectrograms and a WaveNet vocoder. The conversion model has a bottle-neck layer, whose outputs are concatenated with PPGs before being fed into the WaveNet vocoder as local conditioning. A weighted sum of a Mel-spectrogram prediction loss and a WaveNet loss is used as the objective function to jointly optimize parameters of the conversion model and the WaveNet vocoder. Objective and subjective evaluation results show that the proposed approach is capable of achieving significantly improved quality in voice conversion in terms of speech naturalness and speaker similarity of the converted speech for both cross-gender and intra-gender conversions.

Index Terms: voice conversion, WaveNet, mel-spectrograms, non-parallel data, joint training, phonetic posteriorgrams

1. Introduction

Voice conversion (VC) aims to modify a speech utterance spoken by a source speaker to another as if it were uttered by a target speaker, while keeping the linguistic content unchanged. Most conventional VC systems contain a conversion model and a vocoder. The conversion model is trained to learn the mapping function between time-aligned source and target acoustic features, which can be Gaussian mixture models (GMMs) [1, 2], artificial neural networks (ANNs) [3, 4, 5, 6], etc. The vocoder can be a source-filter vocoder, e.g., STRAIGHT [7], WORLD [8], or a WaveNet vocoder [9].

Since parallel speech data between the source and target speakers is expensive to collect, VC techniques using non-parallel data (i.e., non-parallel VC) have been studied. The most successful non-parallel VC approach is the N10 system [10] in the Voice Conversion Challenge 2018 (VCC 2018) [11], which combines a phonetic posteriorgram (PPG)-based conversion model [12] and a WaveNet vocoder. The PPG-based conversion model learns the mapping function from PPGs to acoustic features, which are Mel-cepstrums (MCEPs), F0, voiced/unvoiced flag (VUV) and aperiodicities. The WaveNet vocoder is trained to generate speech waveform conditioned on acoustic features. The N10 system can achieve high conversion performance in terms of speech naturalness and speaker similarity of the converted speech according to the VCC 2018 evaluation results.

However, there are two ways to further improve the conversion performance of the N10 system. The N10 system uses MCEPs as spectral features, which are shown to be worse than Mel-spectrograms for VC in terms of speech naturalness and speaker similarity of the converted speech and F0 conversion [13]. The Mel-spectrogram is a low-level acoustic representation of speech waveform, which is commonly used for local conditioning of a WaveNet vocoder in current state-of-the-art text-to-speech (TTS) architectures [14]. Therefore, we believe that one way to improve the N10 system is to use Mel-spectrograms as acoustic features. The conversion model and the WaveNet vocoder in the N10 system are separately trained, which may not be globally optimal for the whole conversion pipeline. Hence, we believe that a second way to improve the N10 system is to jointly train the conversion model and the WaveNet vocoder. In [15], a compact framework of a condition network and a WaveNet vocoder is proposed to solve the problem of separate training in the N10 system. However, the proposed model in [15] bypasses the conversion model and does not take advantage of the acoustic features.

In this paper, we propose a VC architecture to jointly train a conversion model that maps phonetic posteriorgrams (PPGs) to Mel-spectrograms and a WaveNet vocoder. The conversion model includes several multi-head self-attention [15, 16] and bidirectional LSTM (BLSTM) [17] blocks and one bottle-neck (BN) layer. The output of the BN layer is fed directly into the WaveNet vocoder as local conditioning, which has a similar structure to the text-to-wave model in [18]. Since PPGs represent articulation of speech sounds in a speaker-normalized space and correspond to speech content [12], PPGs are also fed into the WaveNet vocoder through a residual-like connection and we believe that this enables the WaveNet vocoder to reduce pronunciation errors. A weighted sum of a Mel-spectrogram prediction loss and a WaveNet loss is used as the objective function to jointly optimize parameters of the conversion model and the WaveNet vocoder. Objective and subjective evaluation results show that the proposed approach is capable of achieving improved quality in voice conversion in terms of speech naturalness and speaker similarity of the converted speech for both cross-gender and intra-gender conversions.

The contributions of this paper are two folds:

- The proposed approach leverages low-level acoustic representation (i.e., Mel-spectrograms) of speech waveform for non-parallel VC.
- The proposed approach jointly trains the conversion model and the WaveNet vocoder using a multi-task learning mechanism.

http://dx.doi.org/10.21437/Interspeech.2019-1316

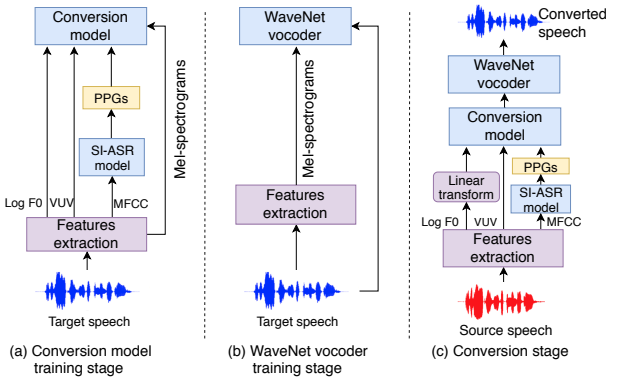

Figure 1: *Training and conversion stages of the Baseline 1.*

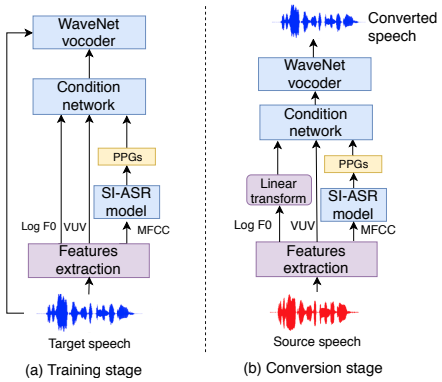

Figure 2: *Training and conversion stages of the Baseline 2.*

The rest of the paper is organized as follows: Section 2 introduces the baseline approaches. Section 3 describes the proposed approach. Experimental setups and evaluation results are presented in Section 4. Section 5 concludes this paper.

2. Baseline Approaches

The work most related to this paper is [10] and [15]. To validate the advantage of jointly training the conversion model and the WaveNet vocoder, we setup a model similar to that in [10] as **Baseline 1**. To validate the advantage of the PPG-to-Mel-spectrogram conversion model, we setup a model similar to that in [15] as **Baseline 2**.

2.1. Baseline 1

The Baseline 1 model is similar to the N10 system in VCC 2018, whose training and conversion stages are shown in Figure 1. Unlike [10], we used Mel-spectrograms as spectral features and did not use adaptation techniques for the WaveNet vocoder training.

During the training stage, as shown in Figure 1(a) and (b), a conversion model and a WaveNet vocoder are trained separately. During the conversion model training stage, logarithmic F0 (Log F0), voiced/unvoiced flag (VUV), Mel-cepstral coefficients (MFCCs) and Mel-spectrograms are extracted from the target speech. MFCCs are then used to compute PPGs using a speaker-independent automatic speech recognition (SI-ASR) model. A BLSTM-based conversion model is then trained to learn the mapping function from PPGs, Log F0 and VUV to Mel-spectrograms. Pitch information has been shown to help learn the mapping relationship between PPGs and spectrograms in [19, 20]. Hence, the conversion model also takes in Log F0 and VUV as inputs. During the WaveNet vocoder training stage, Mel-spectrograms are first extracted from the target speech. Then the μ-law quantized target speech [9] and the Mel-spectrograms are used to train the WaveNet vocoder.

During the conversion stage, as shown in Figure 1(c), Log F0, VUV and MFCCs are first extracted from the source speech. PPGs are then computed from the MFCCs by the SI-ASR model. Log F0 is linear transformed using the pitch statistics of the source and target speakers [21]. The conversion model takes in PPGs, VUV and the tranformed Log F0 as inputs and outputs predicted Mel-spectrograms, which are used by the WaveNet vocoder to generate converted speech.

2.2. Baseline 2

The Baseline 2 model is similar to that in [15]. The training and conversion stages are shown in Figure 2.

During the training stage, a condition network and a WaveNet vocoder are jointly trained. The condition network employs multi-head self-attention and BLSTM structure to encode PPGs into an intermediate representation, which is used as local conditioning by the WaveNet vocoder. Log F0, VUV and MFCCs are first extracted from the target speech. PPGs are then computed by the SI-ASR model from the MFCCs. Log F0, VUV and PPGs drive the condition network to compute local conditioning of the WaveNet vocoder. The WaveNet loss, which is a cross-entropy loss, is adopted as the objective function to optimize the parameters of the condition network and the WaveNet vocoder jointly.

During the conversion stage, the acoustic features from the source speech are processed in the same way as in Section 2.1. The PPGs, VUV and transformed Log F0 drive the condition network and the WaveNet vocoder to generate the converted speech.

3. Proposed Approach

3.1. Model Architecture and Joint Training Mechanism

The architecture of jointly training the conversion model and the WaveNet vocoder is illustrated in Figure 3(a). The combination of the multi-head self-attention and BLSTM structures has been shown to facilitate the learning of both global and local contextual information from PPGs [15], which inspires the design of the conversion model architecture of the proposed approach. The conversion model (gray-shaded box in Figure 3) contains an encoder module and a fully-connected (FC) layer. The encoder module consists of one prenet, N blocks of multi-head self-attention and BLSTM layers and one bottleneck (BN) layer, as shown in Figure 3(c). The BN features from the conversion model are fed into the WaveNet as local conditioning. PPGs obtained from an SI-ASR can represent articulation of speech sounds in a speaker-normalized space and correspond to speech content [12]. We believe that PPGs can help the WaveNet vocoder reduce pronunciation errors. Hence, we also feed PPGs into the WaveNet. One can think of this as a residual-like connection, enabling the model to learn from the BN features useful information that is lacking in PPGs for waveform generation. Parameters of the conversion model and the WaveNet vocoder are optimized using a multi-task learning

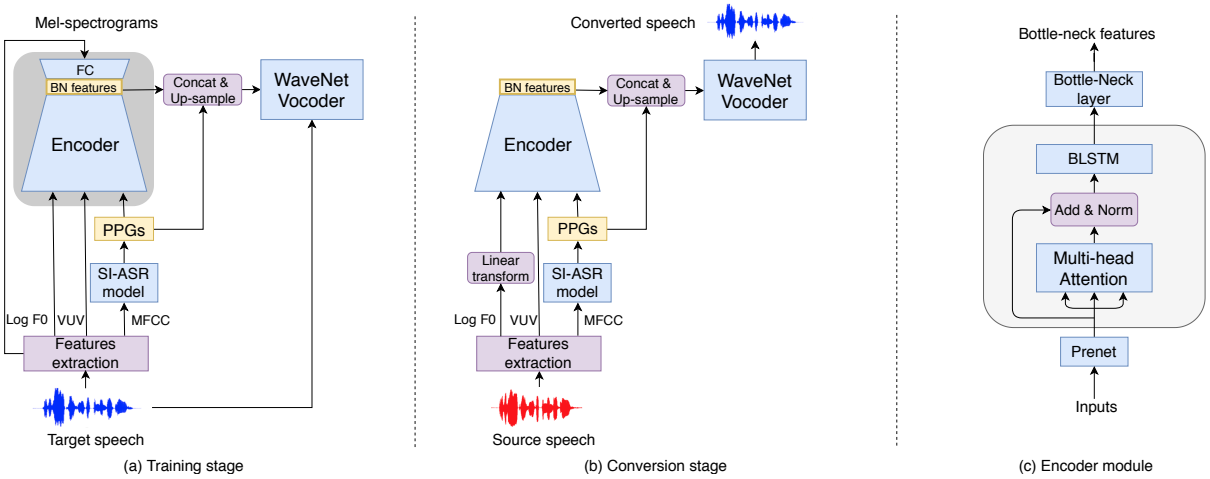

Figure 3: *Training stage (a), conversion stage (b) and the encoder module (c) of the proposed approach. The model in the gray-shaded box in (a) is the conversion model. 'BN features' in (a) and (b) means bottle-neck features.*

scheme. The overall loss function is a linear combination of a WaveNet loss and a Mel-spectrogram L1-loss:

$$loss_{joint} = loss_{WaveNet} + \alpha \times loss_{Mel-spectrogram}, \quad (1)$$

where α is a hyper-parameter, specifying the weight of the Mel-spectrogram L1-loss.

Since the top FC layer of the conversion model conducts shallow linear transformation of the BN features, we can also regard the BN features as latent acoustic features, while the PPGs are speaker-independent linguistic features.

3.2. Training and Conversion

The training and conversion stages of the proposed approach are shown in Figures 3(a) and (b), respectively.

During the training stage, MFCC, Log F0 and VUV features are first extracted from the target speech signals. PPGs are then computed by the SI-ASR model from the MFCC features. The conversion model takes Log F0, VUV and PPG features as inputs and outputs the predicted Mel-spectrograms and also the BN feaures. The BN features and PPGs are then concatenated and up-sampled by repeating to match the time resolution of the speech waveform, before being fed into the WaveNet vocoder as local conditioning. Parameters of the conversion model and the WaveNet vocoder are optimized jointly using the joint loss (i.e., $loss_{joint}$) in Equation (1).

During the conversion stage, we remove the top FC layer from the conversion model since we only need the BN features. PPGs, VUV features and the transformed Log F0 are obtained in the same way as in Baseline 1 and 2, which are used to compute the BN features from the conversion model. Finally, the WaveNet vocoder takes the concatenated and up-sampled BN features and PPGs as local conditioning to generate the converted speech.

4. Experiments

4.1. Experimental Setups

The CMU ARCTIC speech dataset [22] is used to conduct VC experiments. We use "rms" and "slt" as source speakers, and use "bdl" and "clb" as target speakers. We randomly choose 1000, 50 and 50 utterances from the dataset for training, validation and testing, respectively. The TIMIT corpus [23] is used to train the SI-ASR model for PPG extraction. The SI-ASR model has the same setting as in [20]. The sampling frequency is 16 kHz. All the acoustic features are computed using a 25-ms window and 5-ms frame shift. 13-dimension MFCCs are used and PPGs are set to having 131 senones in the SI-ASR model. We use 80-band Mel-spectrograms.

The WaveNet architecture in this paper includes 2 causal convolution blocks, with 10 dilated layers for each. The dilation rate in each block starts from 1 to 512. The filter size of the causal dilated convolution is 3. The number of residual channels and skip channels are 128 and 256, respectively. The speech waveform is μ-law quantized into 256-way categorical distribution.

The BLSTM-based conversion model in Baseline 1 consists of one FC layer, two BLSTM layers with 256 hidden units in each direction and one linear output layer. The condition network in Baseline 2 has two self-attention and BLSTM blocks, which share the same network structure. The number of hidden units is 128 and number of self-attention heads is 8.

The encoder module of the proposed conversion model as shown in Figure 3(c) contains one FC prenet with 128 units, 2 multi-head self-attention and BLSTM blocks and one BN layer with 64 units. The multi-head self-attention and BLSTM blocks have the same network structure as in Baseline 2. A dropout rate of 0.1 is applied to all layers in the conversion model. The weight of the Mel-spectrogram L1-loss α in Equation (1) is tuned using the validation set and set to be 0.001 finally.

Two ablation studies are setup. To validate the efficacy of using the Mel-spectrogram prediction loss, we remove the top FC layer of the conversion model (**ablation 1**). To validate the impact of PPG residual connection, we remove the PPG residual connection and only feed the BN features to the WaveNet vocoder (**ablation 2**).

4.2. Objective Evaluation

We use WORLD [8] to extract MCEPs and F0 features from the converted utterances and target utterances in the test set. Mel-cepstral distortion (MCD) and root mean square error of F0 (F0 RMSE) are computed, as shown in Table 1. We can see that the

Table 1: *MCD and F0 RMSE results of the proposed approach and the reference approaches.*

Setting	Cross-gender		Intra-gender	
	MCD (dB)	F0 RMSE (Hz)	MCD (dB)	F0 RMSE (Hz)
Baseline 1	8.166	50.003	8.134	49.778
Baseline 2	8.153	48.016	8.229	48.118
Ablation 1	8.319	52.138	8.288	49.600
Ablation 2	8.348	49.287	8.283	49.017
Proposed	**7.991**	**44.712**	**7.95**	**47.642**

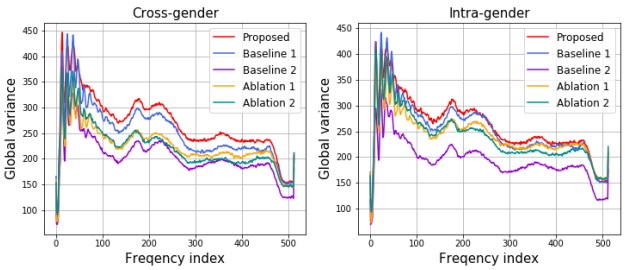

Figure 4: *Visualization of Global variances (GVs) of the proposed approach and the reference approaches.*

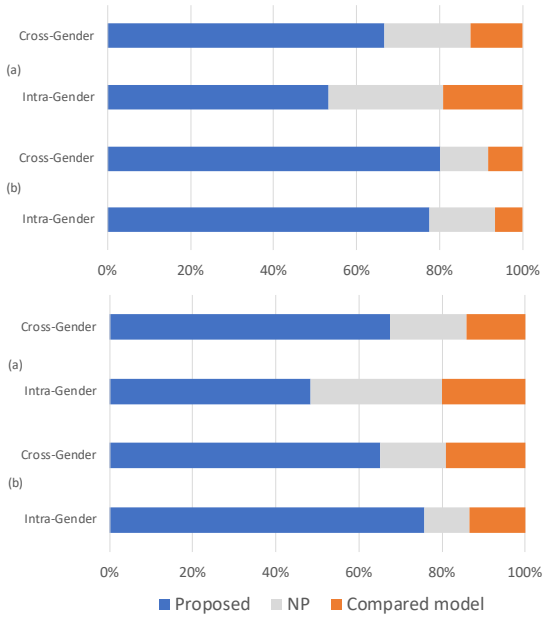

Figure 5: *Speech naturalness AB test (above) and speaker similarity XAB test (below) results. (a) The proposed approach with Baseline 1. (b) The proposed approach with Baseline 2. 'NP' means no preference. The p-values of student t-test on the evaluation results from top to bottom are* 9.52×10^{-5}, 7.10×10^{-3}, 1.05×10^{-4}, 1.20×10^{-5}, 1.20×10^{-3}, 3.91×10^{-2}, 3.01×10^{-3} *and* 5.73×10^{-4}, *respectively.*

proposed approach achieves the lowest MCD among the compared approaches in both the cross-gender and intra-gender conversion cases. According to the F0 RMSE results in Table 1, the proposed method achieves the best F0 conversion performance, which has great impact on the speech naturalness and speaker similarity of the converted speech. Note that in the proposed method, pitch information is encoded into the BN features and then fed into the WaveNet vocoder. MCD and F0 RMSE results of the ablations are also shown in Table 1, which are higher than those of the proposed approach. This validates the efficacy of using the Mel-sepctrogram prediction loss and the advantage of PPG residual connection.

Most conventional VC techniques suffer from the problem of over-smoothing in the converted speech. Higher spectrum global variance (GV) [2] indicates sharpness of the converted speech. Figure 4 plots the average GVs of the proposed approach, the baselines and the ablation approaches. We can see that the proposed method offers the best GV for both cross-gender and intra-gender conversions.

4.3. Subjective Evaluation

Two subjective evaluations are conducted: speech naturalness AB test and speaker similarity XAB test. The proposed approach is subjectively compared with Baselines 1 and 2. In the AB test, paired speech samples (A and B) from the proposed approach and the baseline approaches are presented to listeners, who are asked to indicate the samples with better speech naturalness or show no preference. In the XAB test, X indicates the target reference sample. Paired speech samples (A and B) with the same text content as the reference are presented and the listeners are asked to determine which one has closer timbre to the reference, or if they are equally close. Each conversion (cross-gender and intra-gender) has 20 samples for evaluation. 10 Chinese speakers who are proficient in English have participated in the evaluations and they are allowed to replay each sample pair as many times as necessary in both evaluations[1].

The subjective evaluation results are illustrated in Figure 5. We can see that the proposed approach significantly outperforms the baseline approaches in terms of speech naturalness and speaker similarity of the converted speech for both cross-gender and intra-gender conversions.

5. Conclusions

In this paper, we have investigated the joint training of a PPG-to-Mel-spectrogram conversion model and a WaveNet vocoder for non-parallel voice conversion. Bottle-neck features from the conversion model and PPGs are fed into the WaveNet vocoder as local conditioning. A weighted sum of a Mel-spectrogram prediction loss and a WaveNet loss is used to optimize the conversion model and the WaveNet model jointly. Objective and subjective evaluation results show that the proposed approach is capable of achieving better conversion performance in terms of speech naturalness and speaker similarity than the baseline approaches. Since the WaveNet vocoder requires a large amount of speech data for training, applying the proposed approach in this paper to a multi-speaker corpus is our future work.

6. Acknowledgements

This project is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14208817).

[1] Some audio samples can be found in "https://ooshaunoo.github.io/JntTrn-PPGMelsp-VC-samples/"

7. References

[1] 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.

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

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

[4] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 19–23.

[5] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using rnn pre-trained by recurrent temporal restricted boltzmann machines," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 3, pp. 580–587, 2015.

[6] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4869–4873.

[7] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[8] 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.

[9] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[10] 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.

[11] 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.

[12] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

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

[14] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[15] H. Lu, Z. Wu, R. Li, S. Kang, J. Jia, and H. Meng, "A compact framework for voice conversion using wavenet conditioned on phonetic posteriorgrams," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

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

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

[18] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[19] S. Liu, L. Sun, X. Wu, X. Liu, and H. Meng, "The hccl-cuhk system for the voice conversion challenge 2018," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 248–254.

[20] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance," *Proc. Interspeech 2018*, pp. 496–500, 2018.

[21] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4. IEEE, 2007, pp. 410–414.

[22] J. Kominek and A. Black, "The cmu arctic speech databases for speech synthesis research," Tech. Rep. CMU-LTI-03-177 http://festvox. org/cmu arctic/, Language , Tech. Rep., 2003.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus, 1993," *Linguistic Data Consortium, Philadelphia*.