A Cyclical Post-filtering Approach to Mismatch Refinement of Neural Vocoder for Text-to-speech Systems

Yi-Chiao Wu¹, Patrick Lumban Tobing¹, Kazuki Yasuhara¹, Noriyuki Matsunaga², Yamato Ohtani², and Tomoki Toda³

¹Graduate School of Informatics, Nagoya University, Japan
²AI, Inc., Japan
³Information Technology Center, Nagoya University, Japan

yichiao.wu8g.sp.m.is.nagoya-u.ac.jp, ohtani@ai-j.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

Recently, the effectiveness of text-to-speech (TTS) systems combined with neural vocoders to generate high-fidelity speech has been shown. However, collecting the required training data and building these advanced systems from scratch are time and resource consuming. An economical approach is to develop a neural vocoder to enhance the speech generated by existing or low-cost TTS systems. Nonetheless, this approach usually suffers from two issues: 1) temporal mismatches between TTS and natural waveforms and 2) acoustic mismatches between training and testing data. To address these issues, we adopt a cyclic voice conversion (VC) model to generate temporally matched pseudo-VC data for training and acoustically matched enhanced data for testing the neural vocoders. Because of the generality, this framework can be applied to arbitrary TTS systems and neural vocoders. In this paper, we apply the proposed method with a state-of-the-art WaveNet vocoder for two different basic TTS systems, and both objective and subjective experimental results confirm the effectiveness of the proposed framework.

Index Terms: temporal mismatch, acoustic mismatch, cycle-consistent, voice conversion, post-filter for text-to-speech

1. Introduction

Text-to-speech (TTS) is a technique to generate speech according to the given text. Benefitting from the thriving development of neural network (NN), the advanced TTS systems with NN-based waveform generation models [1, 2] achieve very impressive speech fidelity. However, the high quality and quantity requirements of training data, the burden of data pre-processing, and the time and resource consuming training process make it difficult to build an advanced TTS system from scratch.

A flexible and economical approach to developing a high-quality TTS system is to enhance the speech generated by low-cost or existing TTS systems using an NN-based generation model such as the WaveNet (WN) [3–6] vocoder. However, there are two challenges for training and testing the NN-based vocoder. First, if the NN-based vocoder is trained with natural acoustic features and waveforms, it will suffer from the acoustic mismatch problem in the testing stage. The acoustic mismatch between the synthetic testing acoustic features, which are extracted from the TTS-generated waveforms, and the natural training acoustic features causes a significant speech quality degradation. Secondly, even if training the vocoder with the synthetic acoustic features and the natural waveforms, the temporal structure mismatch between TTS-generated and natural waveforms still degrades the performance of the vocoder.

To tackle these problems, a cycle-voice conversion (Cycle-VC) [7] model is adopted to respectively generate temporally matched pseudo converted acoustic features for training the NN-based vocoder and acoustically matched enhanced acoustic features in the testing stage. Specifically, the Cycle-VC model includes two conversion paths. The first path converts the synthetic acoustic features to the natural ones, and the second path is composed of a natural to synthetic conversion model following the synthetic to natural conversion model of the first path. The enhanced and the pseudo converted acoustic features can be respectively attained from the first and second paths. Because both the enhanced and the pseudo converted acoustic features are converted by the Cycle-VC model, their acoustic mismatches are less than that of the synthetic and natural acoustic features. Since the pseudo converted acoustic features are converted from the natural acoustic features, their temporal structures are matched to the natural waveforms.

Both objective and subjective evaluations are conducted. The experimental results show the speech quality degradations caused by the acoustic and temporal mismatches and the effectiveness of the proposed framework. To sum up, the contributions of this paper are three folds:

• This paper argues that TTS-generated speech with manually determined phoneme alignment still has very different temporal structures from the related natural speech, and these temporal mismatches cause significant speech quality degradations.

• A WN vocoder trained and tested with the proposed framework does enhance the TTS-generated speech.

• The proposed framework can be generalized for arbitrary TTS systems and neural generation models.

2. Related work

For TTS systems with an NN-based vocoder, Tacotron2 [1] has shown an early success by independently training an autoregressive (AR) mel-spectral predictor and then training a WN vocoder with the output of the well-trained mel-spectral predictor. ClariNet [2] improved it with a non-AR parallel WN-like [8] vocoder and a jointly training manner. The authors of [9] also proposed a generative adversarial network [10] (GAN)-based framework to jointly optimize its mel-spectral predictor and vocoder. However, these methods are exclusive for specific mel-spectral predictors and difficult to be combined with arbitrary existing TTS systems.

Furthermore, GAN-based [11] and WN-based [12, 13] denoising models also have been proposed to directly operate the speech enhancement in the waveform domain. However, because the noisy and clean training data are usually paired, which
3. Cycle-spectral conversion

As shown in Fig. 1, the Cycle-VC system is composed of a target-to-source (TtoS) model and a source-to-target (StoT) model. The conventional VC system usually consists of only a StoT model, but the Cycle-VC system adopts an additional TtoS model to advance the speech modeling capability of the StoT model with the cycle-consistency. Moreover, the self-converted target features are suitable for training or fine-tuning the NN-based vocoders. That is, these self-converted target features are alignment-free to the target waveforms, and their acoustic characteristics are similar to the converted features.

4. Proposed post-filter for TTS

4.1. Acoustic and temporal mismatches

Because of the data-driven nature, NN-based vocoders are vulnerable to unseen testing data [24–26]. Specifically, NN-based vocoders are usually trained with a pair of natural acoustic features and waveforms, but the input acoustic features in the testing stage are predicted from other models as shown in Fig. 2 (a). The acoustic mismatch between training and testing data causes significant speech quality degradation.

In this paper, we argue that even if NN-based vocoders are directly trained with a pair of synthetic acoustic features and natural waveforms as Fig. 2 (b), the temporal mismatch problem still causes severe quality degradation. Specifically, although directly training the vocoder with synthetic acoustic features and natural waveforms can alleviate the acoustic mismatch problem in the testing stage, the temporal mismatch between them still markedly degrade the vocoder. Even if the synthetic acoustic features are extracted from manually tuned TTS-generated speech, which has synchronized phoneme durations, short pauses, and silence segments to the natural target speech, there are still some different temporal structures between TTS-generated and natural waveforms. These temporal mismatches in the vocoder training stage usually cause severe quality degradations such as mispronunciation.

4.2. Post-filter with cyclical mismatch refinement

The proposed method is composed of a Cycle-VC training, a vocoder training, and a post-filter testing stages. As shown in Fig. 3, synthetic acoustic features, which are extracted from TTS-generated speech, are taken as the source, and natural acoustic features are taken as the target of the Cycle-VC model. The StoT model is trained with the paired synthetic and natural acoustic features, and the TtoS model is trained with the cycle-consistency.

As the proposed framework shown in Fig. 4, an NN-based vocoder is trained with a pair of temporally matched natural
5. Experiments

5.1. Corpus and TTS system

An internal Japanese corpus, which included a female and a male speakers, with sampling rate 48 kHz was adopted for developing single-speaker TTS systems. Each speaker had 800 training and 100 testing utterances, and the average length of utterances was around 4 seconds. WORLD-based acoustic features [27], which included 60-dimensional mel-cepstral feature (\textit{mcep}), one-dimensional log-scaled fundamental frequency (\textit{F}_0), five-dimensional aperiodicity (\textit{ap}), and their delta and delta-delta terms, were adopted for the TTS systems. The minimum description length was set to 1.0.

Two basic and low-cost TTS systems, Hidden Markov Model (HMM)-based and deep neural network (DNN)-based systems, were adopted, and both of them were trained in a speaker-dependent (SD) fashion using the very limited training data. Specifically, the HMM-based systems were trained with the Hidden Semi-Markov Model (HSMM) training script of HTS (ver. 2.3.2) [28], and the manual phoneme segmentations were adopted to initialize the phoneme HSMMs. The DNN-based systems were composed of four independent feedforward DNNs for respectively predicting the \textit{F}_0, \textit{ap}, \textit{mcep}, and durations. Both systems adopted a maximum likelihood parameter generation (MLPG) [29]. The manually refined phoneme segmentations were utilized to generate synthetic speech, which had the same phoneme durations as the natural speech, of the entire training and testing sets. Moreover, a traditional spectral post-filter [28–30], which was included in the HTS demo, with an enhanced coefficient 1.4 (\beta = 0.4) was only applied to the DNN-based TTS system.

5.2. Cycle-VC model and WN vocoder

Both natural and TTS utterances were downsampled to 24 kHz for SD Cycle-VC models and SD WN vocoders. WORLD-based acoustic features, which included 45-dimensional \textit{mcep}, one-dimensional log-scaled \textit{F}_0 and unvoiced/voiced (\textit{U/V}) binary code, and three-dimensional coded \textit{ap}, were adopted for the Cycle-VC model input and the auxiliary features for the WN vocoders. Note that the outputs of the Cycle-VC models were only the \textit{mcep} features. The settings of the Cycle-VC model followed the previous work [7], and the training epoch was set to 15. Furthermore, several SD WN vocoders were involved in the evaluations to show the effectiveness of the proposed framework for different speakers and systems. The architecture and training processing of the WN vocoders followed our previous work [31] with 200,000 iterations.

5.3. Objective evaluations

Figures 5 shows the relationships among natural (N), synthetic (S), pseudo converted (P), and enhanced (E) \textit{mceps} of the proposed neural post-filters with DNN-based TTS systems on a mel-cepstral distortion (MCD) plane. The distance between any two points represents the average MCD of them. Specifically, although the TTS and natural speech waveforms have the
same contexts and durations, the distance between the natural and synthetic mceps (MCD(S, N)) is still the longest distance on the plane, and this result implies the temporal mismatches of the TTS and natural speech waveforms and the severe mismatches of their mceps. The smaller value of MCD(E, P) than the value of MCD(E, N) indicates that the proposed framework does alleviate the acoustic mismatches between the training and testing mceps of NN-based vocoders. The smaller value of MCD(E, N) than the value of MCD(S, N) also shows the effectiveness of the Cycle-VC model to enhance the synthetic mcep. Furthermore, Fig. 6 has a similar tendency as Fig. 5, which shows the generality of the proposed framework even with different TTS systems.

5.4. Subjective evaluations

As shown in Table 1, seven training and testing combinations of WN vocoders were included in the subjective evaluations. Specifically, the WN vocoder trained with natural acoustic features was tested by natural and DNN/HMM-based synthetic acoustic features, which were the natural and acoustic mismatch (AM) scenarios, respectively. The WN vocoders trained and tested with the DNN/HMM-based synthetic acoustic features were the temporal mismatch (TM) scenarios. The WN vocoders respectively trained and tested with DNN/HMM-based pseudo converted and enhanced acoustic features were the neural post-filter (NPF) scenarios. The subjective evaluations included the systems with these seven scenarios and the DNN-based and HMM-based TTS systems. Note that all systems were trained in an SD manner, so the total number of systems in the subjective tests was 18.

For each testing system and scenario, we randomly selected 50 testing utterances to form the subjective set and the total number of the utterances was 900. Ten subjects were involved in a preference test and a mean opinion score (MOS) test, and most of them were native speakers. Each subject evaluated a part of the subjective set, and each utterance in the subjective set was at least evaluated by one subject. The final results were the average scores of the testing speakers.

As shown in Fig. 7, the enhanced utterances, which were generated by the WN vocoders with the proposed post-filtering framework, were respectively compared with the utterances suffering the AM and TM problems. The results show the effectiveness of the proposed framework to alleviate the AM and TM problems and imply the generality of it for different TTS systems. The results also confirm our assumption that although the TTS-generated speech has the same phoneme durations as the natural speech, the different temporal structures still cause a severe TM problem. Moreover, the AM problem also causes significant speech quality degradations according to the results, so the proposed framework is essential for the post-filter application with an NN-based vocoder.

Figure 8 shows the results of the MOS test, where each subject was asked to give a score (1–5) to evaluate the speech quality of the given utterance. The higher the score, the higher the speech quality. The results show that the proposed post-filter markedly enhanced the TTS-generated speech, even if a traditional spectral post-filter was already applied to the DNN-based TTS. Although there is still a room to improve the speech quality to attain the same quality as the upper bound, which is the WN vocoder trained and tested with the natural acoustic features, the significant improvements (> 1) of MOSs for both TTS systems still show the effectiveness of the proposed framework as a post-filter for arbitrary TTS systems.

6. Conclusions

In this paper, the harmful effects of the acoustic and temporal mismatches for the TTS post-filter with an NN-based vocoder are explored. The proposed framework adopts the Cycle-VC framework to get the temporally matched pseudo converted acoustic features for the training of the NN-based vocoder and the acoustically matched enhanced acoustic features for the testing of the neural post-filter. Both objective and subjective tests of different TTS systems and speakers show the generality and effectiveness of the proposed framework. For future works, we intend to explore the proposed framework with more different NN-based vocoders.

7. Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 17H06101 and JST, CREST Grant Number JP-MJCR19A3.
8. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Shervery-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.

[2] W. Ping, K. Peng, and J. Chen, “ClariNet: Parallel wave generation in end-to-end text-to-speech,” in Proc. ICLR, May 2019.

[3] 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,” in Proc. ICLR, 2016, p. 125.

[4] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” in Proc. Interspeech, Aug. 2017, pp. 1118–1122.

[5] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, “An investigation of multi-speaker training for WaveNet vocoder,” in Proc. ASRU, Dec. 2017, pp. 712–716.

[6] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation,” in Proc. ICASSP, Apr. 2018, pp. 5664–5668.

[7] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, “Voice conversion with cyclic recurrent neural network and fine-tuned wavenet vocoder,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6815–6819.

[8] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L.-C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, “Parallel Wavenet: Fast high-fidelity speech synthesis,” in Proc. ICLR, July 2018, pp. 3915–3923.

[9] Y. Zhao, S. Takaki, H.-T. Luong, J. Yamagishi, D. Saito, and N. Memenatsu, “Wasserstein gan and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a wavenet vocoder,” IEEE Access, vol. 6, pp. 60 478–60 488, 2018.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” in Proc. NIPS, Dec. 2014, pp. 2672–2680.

[11] S. Paszke, A. Bonafonte, and J. Serra, “Segan: Speech enhancement generative adversarial network,” in Proc. Interspeech, 2017.

[12] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hashgawajohnson, “Speech enhancement using bayesian wavenet,” in Proc. Interspeech, 2017, pp. 2013–2017.

[13] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5069–5073.

[14] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.-H. Ling, “Dnn-based stochastic postfilter for hmm-based speech synthesis,” in Proc. Interspeech, 2014, pp. 1954–1958.

[15] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi, “A deep generative architecture for postfiltering in statistical parametric speech synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 2003–2014, 2015.

[16] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “The effect of neural networks in statistical parametric speech synthesis,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 4455–4459.

[17] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, “Generative adversarial network-based postfilter for statistical parametric speech synthesis,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 4910–4914.