EFFECTIVE WAVENET ADAPTATION FOR VOICE CONVERSION WITH LIMITED DATA

Hongqiang Du[1,2], Xiaohai Tian[2], Lei Xie[1], Haizhou Li[2,3]*

[1]Audio, Speech and Langauge Processing Laboratory, School of Computer Science,
Northwestern Polytechnical University, China
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Machine Listening Lab, University of Bremen, Germany

ABSTRACT

WaveNet has shown its great potential as a direct conversion model in voice conversion. However, due to the model complexity, WaveNet always requires a large amount of training data, which has limited its applications in voice conversion, where training data is scarce. In this paper, we propose a WaveNet adaptation method that effectively reduces the need of adaptation data. We first train a speaker independent WaveNet conversion model with multi-speaker dataset. Adaptation is then applied with limited target speaker's data. Specifically, singular value decomposition (SVD) is applied to dilated convolution layers of WaveNet to reduce the number of parameters, which makes adaptation more effective with limited data. Experiments conducted on CMU-ARCTIC and CSTR-VCTK corpus show that the proposed method outperforms baseline methods in terms of both quality and similarity.

Index Terms— Voice Conversion (VC), WaveNet adaptation, Singular Value Decomposition (SVD)

1. INTRODUCTION

Voice conversion (VC) aims to modify speech signal of a source speaker to sound like that of a target speaker without changing the linguistic content. A traditional voice conversion framework consists of feature extraction, feature conversion and speech generation. Various feature conversion approaches have been proposed to transform spectral feature from source speaker to target speaker, e.g. Gaussian mixture model (GMM)) [1, 2], frequency warping [3, 4], exemplar based methods [5] and deep neural network [6, 7]. At the same time, vocoders are widely studied for speech generation, e.g. parametric vocoders (WORLD [8]) and neural vocoders (WaveNet [9]). WaveNet vocoder is one of the most successful implementations, which is proposed for direct waveform modeling and generation in a data-driven manner. Its effectiveness for high quality speech generation has been demonstrated in several VC studies [10, 11]. Recently, a speaker dependent (SD) WaveNet is proposed to jointly optimize the feature conversion and speech generation [12].

Despite the recent progress, the quality of generated speech depends very much on the amount of training data. In order to improve the speech quality with a limited amount of training data, average modeling approaches are proposed. An average model is first trained with the data from multiple speakers. Adaptation is then applied on the average model towards the target speaker with a small amount of data. In practice, different techniques are employed for such model adaptation. For example, maximum a posterior (MAP) approach [13], eigenvoice based approach [14] and i-vector based approach [15]. In [16], a phonetic posteriorgram (PPG) based average modeling approach (AMA) is proposed, where both feature- and model-based AMA are investigated. To further improve the naturalness of the generated speech, WaveNet adaptation is also investigated [17, 18].

In this paper, we propose an effective WaveNet adaptation method for voice conversion with limited data. Inspired by [12], the proposed method makes use of WaveNet to map the PPG to the waveform samples directly. Instead of training an individual model for each target speaker [12], we first train a speaker independent (SI) WaveNet model with multiple speaker's data. Adaption is then applied to the SI WaveNet towards the target speaker with a small amount of target speech. Specifically, the singular value decomposition (SVD) technique is employed to reduce the number of WaveNet parameters, which effectively improve the performance of WaveNet adaptation with limited data. Moreover, the WaveNet is trained between PPGs and the corresponding time-domain speech signals of the same speaker. Hence, only target speech is required for adaptation. The experimental results show that our proposed method can improve the VC results on both quality and similarity with limited data.

*Lei Xie is the corresponding author, lxie@nwpu.edu.cn.

This research work is supported by the National Key Research and Development Program of China (No.2017YFB1002102). This research is also supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2018-006) and by the Programmatic Grant No. A18A2b0046 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain, Project Title: Human Robot Collaborative AI for AME).

2. VOICE CONVERSION WITH WAVENET

2.1. WaveNet vocoder framework

WaveNet vocoder [9] is an auto-regressive model that aims to directly generate raw audio signals conditioned on the acoustic features. In order to model time-domain audio signals effectively, WaveNet vocoder employs an architecture based on a stack of dilated convolution layers and a gated activation unit to learn the joint probability over a sequence of waveform samples. Residual and skip connections are utilized to speed up the convergence and deep model training.

2.2. Voice Conversion with WaveNet

WaveNet has been successfully used as a vocoder in voice conversion, where a feature conversion model is trained for spectral feature transformation, that is followed by a speaker dependent (SD) WaveNet vocoder [10] or adapted WaveNet vocoder [17, 18] for converted speech generation.

In [17], an average model adaptation approach on both feature conversion model and WaveNet vocoder is proposed. This system consists of three steps: average model training, average model adaptation and run-time conversion. During training stage, two average models are first trained with a multi-speaker corpus. Specifically, a feature conversion model is trained to model the relationship between PPG features and acoustic features, while a speaker independent (SI) WaveNet vocoder is trained to reconstruct speech signals from acoustic features. During adaptation, both averaged feature conversion model and SI WaveNet vocoder are fine-tuned with the the same small amount of adaptation data from the target speaker. At run-time, the PPG features extracted from source speech are first converted by averaged feature conversion model. The converted features are then taken by the adapted WaveNet vocoder to generate the converted speech signals.

3. EFFECTIVE WAVENET ADAPTATION FOR VOICE CONVERSION

3.1. WaveNet model for voice conversion

An SI model trained with multi-speaker corpus captures the common characteristics of multiple speakers. It generally requires less data to adapt an SI model towards the target speaker than to train a speaker dependent model from scratch.

Inspired by [12], we introduce a novel average modeling approach using WaveNet to jointly optimize the feature conversion and speech generation process. The proposed framework departs from previous adaptation-based approaches [17, 18] where the feature conversion model and speech generation model are trained separately. It utilizes the phonetic posteriorgrams (PPG) as a local condition to generate time domain speech signal directly.

The proposed framework consists of three steps: speaker independent (SI) WaveNet conversion model training, SI WaveNet model adaptation and run-time conversion.

Fig 1 shows the training process of the SI WaveNet conversion model. We first extract PPG from multiple speaker's corpus to represent the linguistic content of the speech. The first dimension of mel-cepstral coefficients is used to control the energy contour of the generated speech. $F0$ in logarithmic domain ($\log f0$) and voiced/unvoiced (V/UV) flag features are also extracted to represent the prosody. As BAPs are shown useful for waveform reconstruction [19], BAPs and the above features are concatenated to form a vector as the input of SI WaveNet.

For a specific target speaker, we adapt the SI WaveNet conversion model with a small amount of target speaker's data. At run-time conversion, we first extract PPG, energy, $\log f0$, V/UV flag and BAPs features from a source speech. A linear transformation is performed to obtain the converted $\log f0$ as follows.

$$\log f0_y = (\log f0_x - \mu_x) * (\frac{\sigma_y}{\sigma_x}) + \mu_y, \qquad (1)$$

where μ_x and σ_x are the mean and variance of the source speaker's $\log f0$, respectively. μ_y and σ_y are the mean and variance of the target speaker's $\log f0$. $\log f0_x$ and $\log f0_y$ are the source and converted $f0$ in logarithmic domain, respectively. The converted $\log f0_y$ together with other extracted features are are then taken by the adapted WaveNet conversion model as the input to generate converted speech.

3.2. WaveNet factorization with SVD for adaptation

The model complexity is another factor which affects the data requirement for adaptation. To reduce the model complexity, in [20], we proposed a data-efficient SD WaveNet vocoder. We applied the SVD technique to factorize WaveNet vocoder, which reduced the number of parameters significantly. Compared with original SD WaveNet vocoder, the factorized SD WaveNet vocoder maintained similar performance, in terms of both quality and similarity, while using much less training data. We propose to apply the SVD technique in SI WaveNet to reduce the model complexity.

As shown in Fig. 1, SVD technique is applied to factorize the WaveNet. Specifically, a 1×1 convolution layer is added after each dilated convolution layer of WaveNet to reduce the number of model parameters. Consequently, we reduce the required amount of target speech for adaptation.

4. EXPERIMENTAL SETUP

4.1. Database and feature extraction

Both CMU-ARCTIC [21] and CSTR-VCTK [22] databases were used in the experiments. CSTR-VCTK database, consisting of 44 hours speech from 109 speakers, was used for

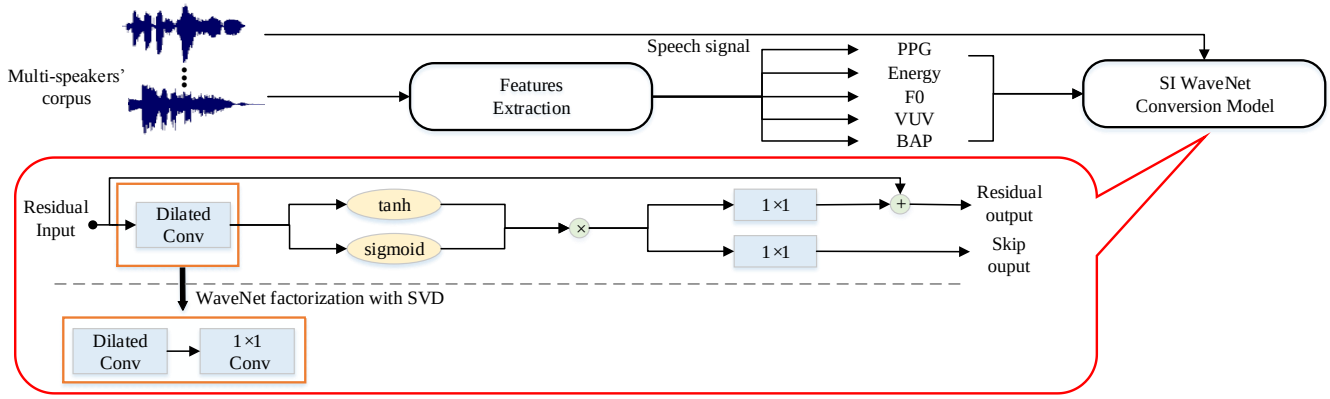

Fig. 1. Diagram of the training process of proposed effective SI WaveNet adaptation for voice conversion. In the red box, WaveNet is factorized by inserting a 1×1 convolution layer after each dilated convolution layer, which reduces the parameters of WaveNet. By reducing the number of parameters, we use less data for effective adaptation of WaveNet.

all the SI model training. Voice conversion experiments were conducted on CMU-ARCTIC database. Four speakers were selected, including two male speakers (bdl and rms) and two female speakers (clb and slt). Voice conversion was carried out between following pairs: clb to bdl, clb to slt , rms to bdl and rms to slt. For each system, we selected 20 and 50 sentences for adaptation, while another 20 non-overlap utterances were used for evaluation. All audio files were downsampled to 16 kHz.

WORLD vocoder [8] was employed to extract 513 dimensional spectrum, 1 dimensional aperiodicity coefficient and 1 dimensional $f0$ with 5ms frame shift. Then 40 dimensional mel-cepstral coefficients (MCCs) were calculated from spectrum using speech signal processing toolkit (SPTK) [1]. 42-dimensional phonetic posteriorgram (PPG) features were extracted by the PPG extractor trained on the Wall Street Journal corpus (WSJ) [23].

4.2. Systems and setup

- **AMA-WORLD**: This is the average modeling approach for voice conversion with WORLD vocoder. The AMA [16] consists of one feed-forward layer and two long short-term memory layers. Each hidden layer consists of 1024 units. The network input is PPG features (42-dim); While the output is acoustic feature (127-dim), consisting of the V/UV flag (1-dim) combining with MCC (40-dim), $\log f0$ (1-dim) and aperiodicity (1-dim) with their dynamic and accelerate features.
- **AMA-WaveNet**: We use the same setting as AMA-WORLD except that the adapted WaveNet vocoder is used for speech generation [17].
- **WaveNet-adp**: The proposed WaveNet adaptation voice conversion system. The PPG (42-dim), energy (1-dim), 3 dimensional $\log f0$ (static, delta and delta delta) together with the aperiodicity (1-dim) and the voiced/unvoiced/ (V/UV) flag (1-dim) are used as the local condition input of the WaveNet.
- **WaveNet-SVD-adp**: We use the same setting as WaveNet-adp except that SVD is applied to WaveNet to reduce the model parameters.

All the WaveNet model consists of 30 dilated convolution layers, which are divided into 3 blocks, with 10 dilated convolution layers in each block. The hidden units of residual connection and skip connection are set to 256. The networks are trained using the Adam optimization method with a constant learning rate of 0.0001 for 600,000 steps. The mini-batch size is 14,000 samples. The speech is encoded by 8 bits μ-law.

5. EVALUATIONS

5.1. Objective evaluation

We use root mean squared error (RMSE) to evaluate distortion between the target and converted speech. A lower RMSE indicates a smaller distortion. Given a speech frame, the RMSE is calculated as $RMSE = \sqrt{\frac{1}{F}\sum_{f=1}^{F}\left(20\log_{10}\frac{|Y(f)|}{|X(f)|}\right)^2}$, where, $|X(f)|$ and $|Y(f)|$ represent the magnitude of synthesized speech and natural speech at f-th frequency bin, respectively. F is the number of frequency bins.

Table 1 shows the average RMSE results for different systems. With 20 adaptation utterances, we observe that the proposed WaveNet-SVD-adp outperforms both AMA-WaveNet and WaveNet-adp baselines. When the number of adaptive data increases to 50, the RMSE of WaveNet-adp decreases from 15.62 dB to 14.15 dB, and achieves similar performance of WaveNet-SVD-adp (14.21 dB). This suggests that WaveNet factorized with SVD is more effective for

[1] https://sourceforge.net/projects/sp-tk/

Table 1. Comparison of average RMSE (dB) for different systems.

System	Utterances	Vocoder	RMSE
AMA-WORLD	20	WORLD	13.50
AMA-WaveNet	20	WaveNet	14.31
WaveNet-adp	20		15.62
WaveNet-adp (50)	50		14.15
WaveNet-SVD-adp	20		**14.21**

adaptation with limited data.

We also notice that AMA-WORLD achieves the lowest RMSE. RMSE is an indirect measurement. It has been observed that conventional vocoders usually give a lower RMSE than those with WaveNet vocoder [11].

5.2. Subjective evaluation

For subjective evaluation, we first conducted AB and XAB preference tests to assess speech quality and speaker similarity. Then multiple stimuli with hidden reference and anchor (MUSHRA) [24] is utilized to evaluate our models. For each system, 20 samples were randomly selected from the 80 converted samples for listening tests. 10 listeners participated in all listening tests.

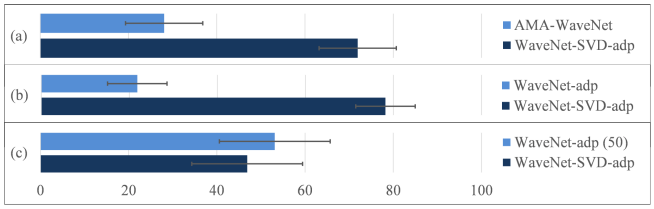

Fig. 2. Quality preference tests of converted speech samples with 95% confidence intervals of different systems for (a) WaveNet-SVD-adp vs AMA-WaveNet, (b) WaveNet-SVD-adp vs WaveNet-adp, (c) WaveNet-SVD-adp vs WaveNet-adp (50).

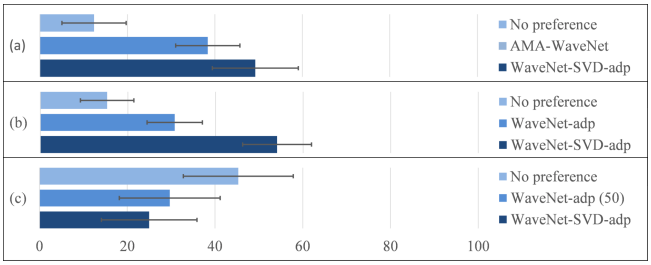

Fig. 3. Similarity preference tests of converted speech with 95% confidence intervals for (a) WaveNet-SVD-adp vs AMA-WaveNet, (b) WaveNet-SVD-adp vs WaveNet-adp, (c) WaveNet-SVD-adp vs WaveNet-adp (50).

Fig. 2 shows the quality preference results of AB tests. With 20 adaptation utterances, as shown in Fig. 2 (a) and (b), WaveNet-SVD-adp outperforms AMA-WaveNet and WaveNet-adp. When the number of adaptation utterances increases to 50, WaveNet-SVD-adp is close to WaveNet-adp (50), they are not significantly different as observed in Fig. 2 (c).

Fig. 3 shows the similarity preference results of XAB tests. The similarity results are consistent with that of AB test. We observe that WaveNet-SVD-adp outperforms AMA-WaveNet and WaveNet-adp. WaveNet-SVD-adp and WaveNet-adp (50) are not significantly different in terms of speaker identity.

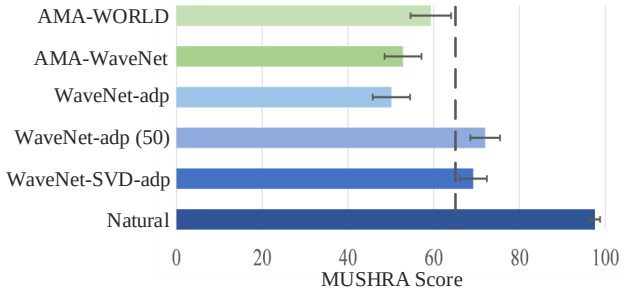

Fig. 4. MUSHRA scores with 95% confidence intervals on speech quality of waveforms reconstructed with different systems.

The results of MUSHRA listening test are shown in Fig. 4. Given 20 utterances for adaptation, listeners favor AMA-WaveNet more than WaveNet-adp. After applying SVD on WaveNet, WaveNet-SVD-adp outperforms clearly other systems with 20 utterances for adaptation. MUSHRA scores of WaveNet-SVD-adp and WaveNet-adp (50) fall into each other's confidence intervals, which means that they are not significantly different.

Therefore, we conclude that the SI WaveNet conversion model factorized with SVD is more effective with limited data for adaptation. The synthesized samples with different systems can be found from the website [2].

6. CONCLUSIONS

In this study, we propose an effective WaveNet adaptation method for voice conversion. The proposed method utilizes a speaker independent WaveNet to map the PPG to waveform directly, Singular value decomposition is employed to reduce the model complexity, which further reduce the data requirement of target speaker. The experimental results demonstrate that our proposed method outperforms the baseline methods with limited data in terms of quality and similarity.

[2] https://dhqadg.github.io/WaveNet-VC/

7. REFERENCES

[1] Tomoki Toda, Alan W Black, and Keiichi 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.

[2] Alexander Kain and Michael W Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. IEEE, 1998, vol. 1, pp. 285–288.

[3] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[4] Elizabeth Godoy, Olivier Rosec, and Thierry Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2011.

[5] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 313–317.

[6] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad, "Voice conversion using artificial neural networks," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3893–3896.

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

[8] Masanori Morise, Fumiya Yokomori, and Kenji 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] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[10] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori, and Tomoki Toda, "Statistical voice conversion with wavenet-based waveform generation.," in *Interspeech*, 2017, pp. 1138–1142.

[11] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent wavenet vocoder.," in *INTERSPEECH*, 2017, pp. 1118–1122.

[12] Xiaohai Tian, Eng-Siong Chng, and Haizhou Li, "A speaker-dependent wavenet for voice conversion with non-parallel data," in *Interspeech*, 2019.

[13] Chung-Han Lee and Chung-Hsien Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Ninth International Conference on Spoken Language Processing*, 2006.

[14] Tomoki Toda, Yamato Ohtani, and Kiyohiro Shikano, "Eigenvoice conversion based on gaussian mixture model," 2006.

[15] Jie Wu, Zhizheng Wu, and Lei Xie, "On the use of i-vectors and average voice model for voice conversion without parallel data," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[16] Xiaohai Tian, Junchao Wang, Haihua Xu, Eng Siong Chng, and Haizhou Li, "Average modeling approach to voice conversion with non-parallel data.," in *Odyssey*, 2018, pp. 227–232.

[17] Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou, and Li-Rong Dai, "Wavenet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, 2018, pp. 1983–1987.

[18] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," *Interspeech*, pp. 1978–1982, 2018.

[19] Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Maximum likelihood voice conversion based on gmm with straight mixed excitation," 2006.

[20] Hongqiang Du, Xiaohai Tian, Lei Xie, and Haizhou Li, "Wavenet factorization with singular value decomposition for voice conversion," in *ASRU2019*, 2019.

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

[22] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[23] Douglas B Paul and Janet M Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[24] ITUR Recommendation, "Method for the subjective assessment of intermediate sound quality (mushra)," *ITU, BS*, pp. 1543–1, 2001.