# Online Speaker Adaptation for WaveNet-based Neural Vocoders

Qiuchen Huang, Yang Ai, Zhenhua Ling
National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
E-mail: {qchuang, ay8067}@mail.ustc.edu.cn, zhling@ustc.edu.cn

***Abstract*—In this paper, we propose an online speaker adaptation method for WaveNet-based neural vocoders in order to improve their performance on speaker-independent waveform generation. In this method, a speaker encoder is first constructed using a large speaker-verification dataset which can extract a speaker embedding vector from an utterance pronounced by an arbitrary speaker. At the training stage, a speaker-aware WaveNet vocoder is then built using a multi-speaker dataset which adopts both acoustic feature sequences and speaker embedding vectors as conditions. At the generation stage, we first feed the acoustic feature sequence from a test speaker into the speaker encoder to obtain the speaker embedding vector of the utterance. Then, both the speaker embedding vector and acoustic features pass the speaker-aware WaveNet vocoder to reconstruct speech waveforms. Experimental results demonstrate that our method can achieve a better objective and subjective performance on reconstructing waveforms of unseen speakers than the conventional speaker-independent WaveNet vocoder.**

***Index Terms*—WaveNet, neural vocoder, speech synthesis, speaker adaptation, speaker embedding vector**

## I. Introduction

In recent years, speech synthesis has become an essential technique for intelligent speech applications, such as audiobook, customer service, speech translation, etc. At present, speech synthesis also faces more and more challenges, such as high quality, high efficiency, and better generalization ability toward multi-speakers.

Statistical parametric speech synthesis (SPSS) is one of the mainstream speech synthesis approaches, which is achieved by acoustic modeling and vocoder-based waveform generation. It has advantages of smoothness, flexibility and coherence. Acoustic models predict acoustic features from input linguistic features and can be built based on hidden Markov models (HMM) [1], neural networks or other deep learning methods. Then, vocoders reconstruct speech waveforms from the predicted acoustic features. Traditional vocoders usually adopt the source-filter signal processing model, i.e., passing a spectrally flat excitation (impulse train or noise) through a linear vocal tract filter, to reconstruct speech waveforms. Represented by STRAIGHT [2] and WORLD [3], these vocoders are convenient and practical but have some deficiencies. For example, the process of real speech production contains nonlinear effects, which can not be reflected by linear filtering. In addition, there is the loss of spectral details and phase information in these vocoders.

Recently, deep learning models have been widely applied to various signal processing tasks. In the filed of speech synthesis, vocoders based on neural networks have also been studied. WaveNet [4], a non-linear autoregressive waveform generation model, has been proposed and WaveNet-based neural vocoders [5] outperformed traditional vocoders on the naturalness of generated speech. Some variants, including WaveRNN [6], FloWaveNet [7], ClariNet [8] and WaveGlow [9], have also been proposed to improve the performance and efficiency of WaveNet vocoders. However, all of the above neural vocoders rely on speaker-dependent model training. For some applications such as personalized and expressive speech synthesis, the training data of a target speaker is usually limited. Besides, separate models need to be stored for different speakers which increases the footprint of speech synthesis systems when adding speakers.

The methods of acoustic model adaptation have been well-studied in traditional SPSS [10]. For building neural vocoders, speaker-dependent training methods have also been proposed [11], [12] to avoid the demand for large speaker-dependent training datasets. Liu et al. [11] initialized the WaveNet model with a multi-speaker corpus and then fine-tuned it with the small amount of data from the target speaker. In this method, speaker embedding vectors were learnt simultaneously with WaveNet parameters at the initialization stage. Besides, a quasi-periodic WaveNet vocoder (QPNet) was also proposed [12], whose dilated convolution structure is adjusted to the fundamental frequency to enhance pitch controllibiliy for better speaker adaptation.

These speaker adaptation methods of neural vocoders still require a certain amount of adaptation data and an extra adaption process for each target speaker. Thus, they can not achieve fully speaker-independent generation of speech waveforms which is the advantage of traditional source-filter-based vocoders [2], [3]. Aiming at avoiding this deficiency, the methods of building speaker-independent neural vocoders have been studied using WaveNet [13] and WaveRNN [6] or combining speech production mechanisms [14]. Speaker-independent WaveNet vocoder [13] used a multi-speaker corpus to train a conditional WaveNet directly. However,

its performance degraded significantly on unseen speakers comparing with speaker-dependent counterpart [5]. The reason is that we are unable to cover the voice characteristics of all possible unseen speakers with the training set of vocoders, which makes the built vocoder models prone to overfit to the speaker characteristics in the training set.

On the other hand, the techniques of representing speaker-specific information using speaker identity embeddings have also been investigated. There are mainly three types of speaker identity embeddings: a speaker-code vector (e.g., one-hot vector), an acoustic-driven vector extracted using external models such as i-vector [15] or d-vector [16], [17], and an acoustic-driven vector based on encoders jointly trained with acoustic models [17], [18]. These speaker identity embeddings have been widely used in speaker identification and speech recognition tasks [19]–[22]. They have also achieved good performance on building acoustic models for multi-speaker speech synthesis [15], [16], [23]–[25].

In this paper, we integrate d-vectors into neural vocoders and propose an online speaker adaptation method in order to improve the performance of speaker-independent WaveNet vocoders when dealing with unseen speakers. First, a speaker encoder is trained using a large speaker-verificatioin dataset. For each utterance in the multi-speaker vocoder training set, a speaker embedding vector, i.e., d-vector, is extracted using the built speaker encoder. Then, these speaker embedding vectors are utilized as auxiliary features to train a speaker-aware WaveNet vocoder. At the generation stage, we send the acoustic feature sequence of a test utterance into the speaker encoder to extract its speaker embedding vector, which is combined with acoustic features and passed through the speaker-aware WaveNet vocoder for waveform reconstruction. Experiment results show that our proposed method can synthesize speech with better objective and subjective quality than the traditional speaker-independent WaveNet vocoder.

Our paper is organized as follows. Section II introduces the details of our proposed method. Section III describes the vocoders we built for comparison and the evaluation results. Section IV is the conclusion.

## II. Methods

As shown in Fig. 1, our proposed model is composed of two separately trained neural networks: a recurrent speaker encoder which computes the speaker embedding vector from the given acoustic features of each utterance, and an auto-regressive WaveNet vocoder which utilizes the concatenation of the speaker embedding and acoustic features as the condition for waveform reconstruction.

### A. Speaker Encoder

The speaker encoder is employed to extract speaker embedding vectors from the acoustic features of target speakers which be used to assist the vocoder network in generating waveforms. The extracted speaker embedding vectors are expected to reflect the speaker characteristics of given acoustic features rather than the text contents or the background noise.

For building the speaker encoder, we refer to the previous study [21] which proposed an efficient and accurate text-independent speaker verification model based on the generalized end-to-end (GE2E) loss. This model mapped the acoustic features of an utterance into a fixed-dimensional speaker embedding vector, known as d-vector [16], [17]. By optimizing the GE2E loss, the d-vectors of the training utterances from the same speaker achieved a high cosine similarity, while those of utterances from different speakers remained far apart in the embedding space.

In our implementation, the speaker encoder is a 3-layer LSTM network including 768 units with projection size of 256. The embedding vector (d-vector) is defined as the network output at the last frame and its dimension is the same as the projection size of the LSTM network. At the inference stage of processing each utterance, we apply a sliding window of fixed 160 frames with 50% overlapping. The d-vector of each window is first computed and the final utterance-wise d-vector is generated by averaging the L2 normalization of network output. Different from previous work for speaker verification [21], the acoustic features used to train the speaker encoder here are consistent with the given features of our WaveNet vocoder, which includes 40-dimensional mel-cepstra, an energy, an F0 and a voiced/unvoiced (V/UV) flag for each frame. STRAIGHT [2] is applied for natural acoustic features extraction. The window size is 400 samples (25ms) and the window shift is 80 samples (5ms).

### B. Speaker-Aware WaveNet Vocoder

WaveNet [4] is a deep autoregression-based convolutional neural network which can directly generate high-fidelity audio signal sample-by-sample. WaveNet vocoder [5] models the joint distribution of waveform samples given auxiliary acoustic features which can be factorized as a product of conditional probabilities as

$$p(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}, \boldsymbol{h}), \tag{1}$$

where $x_t$ is the waveform value of the $t$-th sample, $T$ is the waveform length, and the condition $\boldsymbol{h}$ is the sequence of acoustic features. For modeling the conditional probabilities, WaveNet employs a stack of dilated causal convolution layers. The history waveforms and condition features pass through these convolution layers with gated activation functions, and predict the posterior probability of the current waveform sample with the $\mu$-law quantization [26] using a softmax output layer.

In our proposed method, a speaker-aware WaveNet vocoder is built, which means that the condition $\boldsymbol{h}$ contains not only the acoustic features of input utterance but also the speaker embedding vector extracted from these acoustic features by the speaker encoder. The speaker encoder is trained with a large enough speaker identification dataset. By introducing speaker embedding vectors, we expect that they contain implicit information which is not capable to be extracted from acoustic features during the waveform reconstruction by using the

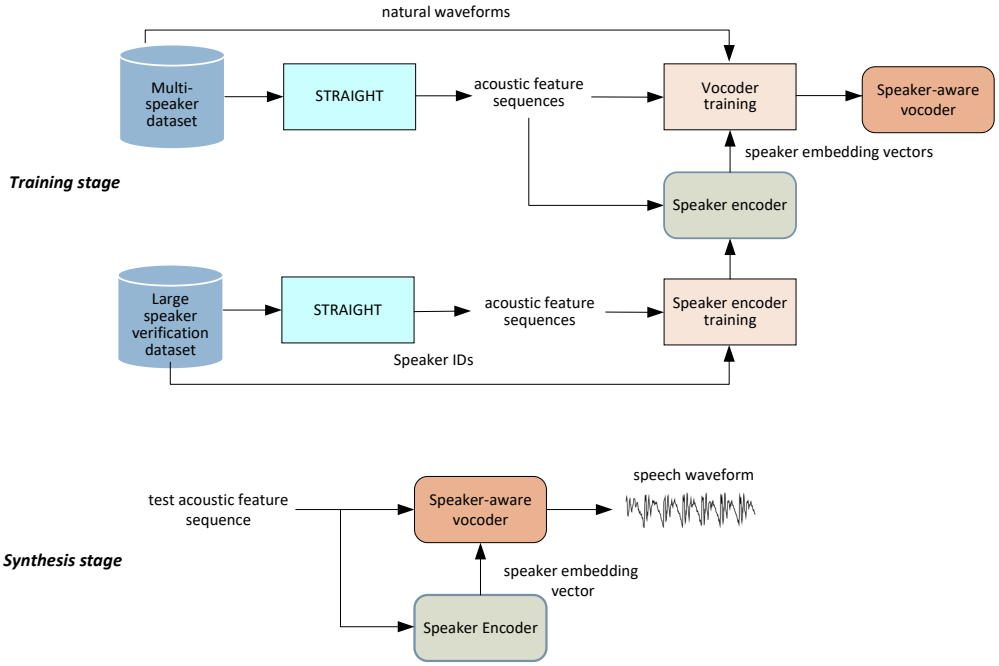


Fig. 1. The training and synthesis procedures of our proposed model.

traditional conditional WaveNet vocoder, thus can improve the speaker-independency and the speaker-generalization ability of the WaveNet model. In our implementation, the speaker embedding vector is concatenated with the acoustic features in each frame. Then, they pass through a conditional network consisting of a $1 \times 1$ convolution layer, a stack of 4 dilated convolution layers and an upsampling layer before acting as the sample-wise local conditions of the WaveNet model.

As shown in Fig. 1), a multi-speaker dataset is adopted to train the speaker-aware WaveNet vocoder. Acoustic features are first extracted from all training utterances by STRAIGHT [2]. Then, a speaker embedding vector is extracted from each utterance using the built speaker encoder. It is expected that the speaker embedding vectors can capture speaker-related information in the training set. Finally, the parameters of the speaker-aware WaveNet model are estimated under cross-entropy criterion using waveforms, acoustic features together with speaker embedding vectors of the training set.

At the synthesis stage, given the acoustic features of a test utterance, we first extract its speaker embedding vector as shown in Fig. 1). Then, the extracted speaker embedding vector are concatenated with the acoustic features for waveform construction. Here, the online adaptation of WaveNet vocoder is achieved because the speaker embedding vector is calculated for each input utterance from an arbitrary speaker and it is not necessary to conduct model adaptation offline by using pre-collected data.

## III. Experiments

### A. Datasets

The VCTK corpus [27] was adopted to build vocoders in our experiments. This dataset was downsampled to 16kHz. It contained 44-hours utterances recorded from 109 native speakers of Engilish with various accents. For the experiments, the corpus was split into three disjoint sets. A total of 34,977 utterances from 99 speakers and 288 utterances from 10 unseen speakers were chosen to construct the training set and the test set of vocoders. The remaining 3,028 utterances from the 10 unseen speakers were used as the offline adaptation set.

For building the speaker encoder, a dataset with more speakers is necessary in order to deal with diversified speakers. Thus, the subsets of the Librispeech and Voxceleb1 corpora used in the previous study on speaker diarization [22] was adopted here. The train-other-500 subset of Librispeech [28] contained 148,688 utterances from 1,166 speakers, and the dev subset of Voxceleb1 [29] contained over 147,935 utterances from 1,211 speakers. And we used the same test set as that of vocoder to evaluate the performance of speaker encoder.

### B. Model Construction

In order to investigate the effectiveness of our proposed method, we built three types of WaveNet-based vocoders for comparison using the same VCTK training set. The configurations of them are described below.

*1) Speaker Independent (SI) Vocoder:* This vocoder was built by training a unified WaveNet model without speaker embeddings [13] and acted as the baseline in our experiments.

TABLE I
SPEAKER VERIFICATION EERS (%) OF TWO SPEAKER ENCODERS ON UNSEEN SPEAKERS.

| System | Speaker Encoder Training Datasets | EER(%) |
|---|---|---|
| OSA1 | Librispeech,Voxceleb1 | 2.96 |
| OSA2 | Librispeech,Voxceleb1,VCTK | 1.07 |

The WaveNet configurations were the same as the ones of our proposed models.

*2) Offline Speaker Adaptation (SA) Vocoders:* For each speaker in the test set, five speaker-dependent WaveNet vocoders were built by fine-tuning the SI vocoder using 20%, 40%, 60%, 80%, and 100% adaptation data of this speaker respectively.

*3) Online Speaker Adaptation (OSA) Vocoders:* As shown in Table I, two proposed vocoders were built by using the speaker encoders estimated with different training sets. For the OSA1 vocoder, we used the Librispeech and Voxceleb1 datasets introduced above to train the speaker encoder. For the OSA2 vocoder, we further added 99 speakers in the VCTK training set to train the speaker encoder. The outputs of both speaker encoders were 256-dimensional speaker embedding vectors. The built WaveNet model had 4 convolutional blocks. Each block had 10 dilated casual convolution layers whose filter width was 2 and dilation coefficients were $\{1, 2, \ldots, 2^9\}$. In the gated activation units, the number of gate channels was 100. The numbers of residual channels and skip channels in the residual architectures were 100 and 256 respectively. The waveform samples were quantized by 8-bit $\mu$-law. In the conditional network, the 299-dimensional condition first passed through a $1\times1$ convolution layer and a stack of 4 dilated convolution layers. The channels of all the convolutional layers were 80. The filter size was 3 and the dilation coefficients were $\{1, 2, 4, 8\}$ for the dilated convolution layers. Finally the 80-dimensional output was connected to the gated activation units after upsampling. The upsampling was performed by repeating the output within each frame. The target of the training was to minimize the cross-entropy and an *Adam* optimizer [30] was adopted to update the model parameters. The initial learning rate was 0.0001, which was halved every 100000 steps. The model was trained in a total of 400000 steps. The Models were trained and evaluated on a single Nvidia 1080Ti GPU.

### C. Performance of Speaker Encoders

To measure the performance of built speaker encoders, we computed the equal error rates (EERs) of speaker verification on the VCTK test set and the results are shown in Table I. We enrolled the utterances from the 10 speakers in the VCTK test set. For both speaker encoders, the enrolled and verification speakers were unseen at the training stage. EER was calculated by pairing each test utterance with each enrolled speaker. From Table I, we can see that the EERs of both vocoders were relatively low which indicated that both speaker encoders can extract speaker-related information from acoustic feature sequences effectively. Besides, the speaker encoder of OSA2 achieved lower EER than that of OSA1 because the VCTK training set was also utilized to train the speaker encoder of OSA2, which may reduce the mismatch between the training and test data of the encoder.

### D. Objective Evaluation Results

Five objective metrics comparing the waveforms generated by vocoders with natural references, including signal-to-noise ratio (SNR), root mean square error of log amplitude spectra (RMSE-LAS), mel-cepstral distortion (MCD), root mean square error of F0 (RMSE-F0), and voice/unvoiced error rate (V/UV Error), were used in our objective evaluation. The detailed formulae for calculating these metrics can be found in previous studies [5], [13].

The test set evaluation results are shown in Table II. Regarding with the five SA vocoders, we can see that their performances on SNR, RMSE-F0 and V/UV error rate were improved when using more adaptation data, while their spectral distortions (i.e., RMSE-LAS and MCD) were almost the same. Comparing the proposed OSA vocoders with SA vocoders, we can see that the spectral distortions of OSA models were comparable with or better than that of SA models, while the SNRs of OSA models were worse than that of SA models, no matter how much adaptation data was used by SA models. The F0-RMSEs of OSA models were close to that of the SA model using 40% adaptation data, and the V/UV error rate of OSA1 was comparable with that of the SA model using 20% adaptation data. Furthermore, although OSA2 achieved lower EER than OSA1 as shown in Table I, its objective performances were not clearly better than OSA1, except on the metric of F0-RMSE.

### E. Subjective Evaluation Results

For comparing the subjective performance of the speech generated by SI, SA and OSA2 vocoders, four groups of ABX preference tests were conducted on the crowdsourcing platform of Amazon Mechanical Turk with anti-cheating considerations. In each test, 20 utterances generated by two comparative vocoders were randomly picked out from the test set. Each pair of voice was evaluated in random order. At least 30 English native speakers were asked to judge which utterance in each pair had better naturalness or sounded more similar to the natural reference. In addition to average preference scores, the p-value of t-test was also calculated to measure the significance of the difference between two comparative vocoders.

The subjective evaluation results are shown in Table III. We can see that the OSA2 vocoder achieved better naturalness of reconstructed speech than the SI vocoder significantly ($p < 0.01$). Meanwhile, we can find that the SA vocoder outperformed both SI and OSA2 vocoders on both naturalness and similarity. The preference score differences between SA and OSA2 were smaller than those between SA and SI. All these results indicate that our proposed method is able to improve the subjective quality of reconstructed waveforms comparing with the traditional speaker-independent WaveNet vocoder. Although OSA1 also achieved higher preference

TABLE II
OBJECTIVE EVALUATION RESULTS ON TEST SET. HERE, SA STANDS FOR THE VOCODER USING 100% ADAPTATION DATA OF EACH SPEAKER.

| Model | SNR(dB) | RMSE-LAS(dB) | MCD(dB) | RMSE-F0(cent) | V/UV error rate(%) |
|---|---|---|---|---|---|
| SI | 3.05 | 8.60 | 2.03 | 69.88 | 7.71 |
| SA_20% | 3.45 | 8.49 | 2.00 | 62.90 | 6.27 |
| SA_40% | 3.56 | 8.51 | 2.00 | 53.97 | 6.04 |
| SA_60% | 3.69 | 8.47 | 2.00 | 49.82 | 5.90 |
| SA_80% | 3.76 | 8.48 | 2.00 | 47.27 | 5.96 |
| SA | **3.83** | 8.49 | 2.00 | **45.97** | **5.86** |
| OSA1 | 3.29 | **8.43** | **1.96** | 55.28 | 6.23 |
| OSA2 | 3.30 | 8.46 | 1.98 | 51.17 | 6.52 |

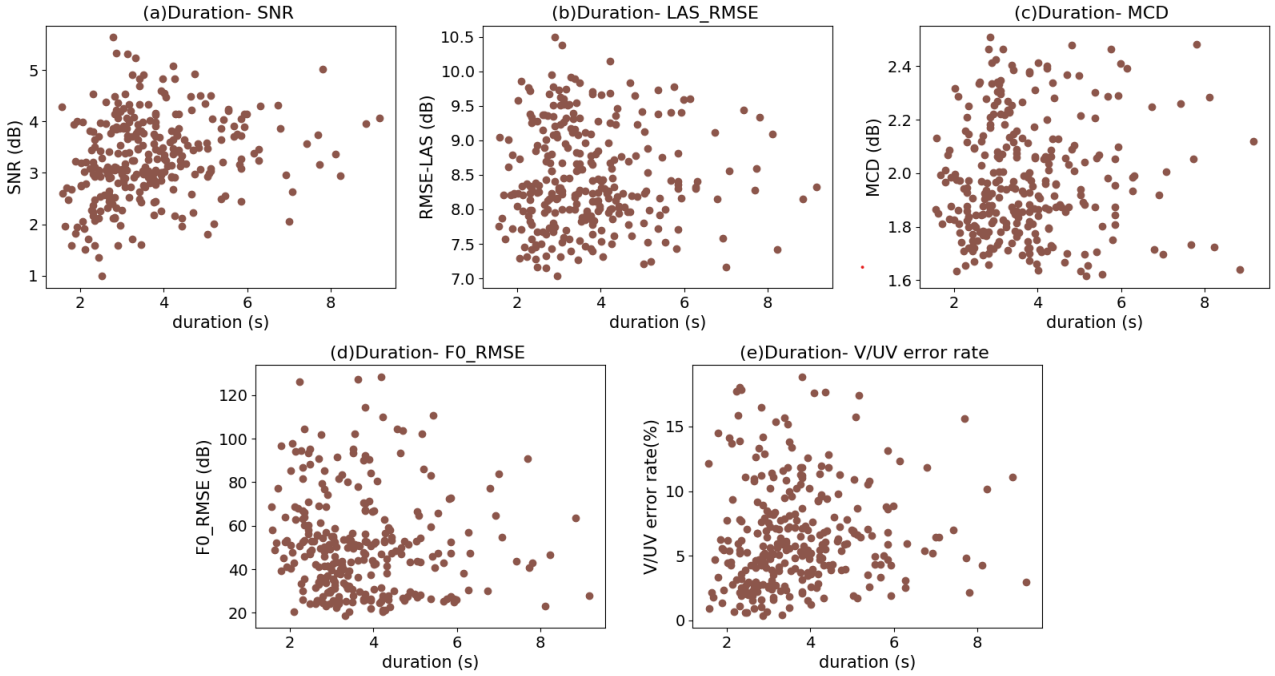


Fig. 2. The scatter diagrams between utterance durations and objective evaluation metrics on 288 test utterances.

TABLE III
AVERAGE PREFERENCE SCORES (%) ON NATURALNESS AND SIMILARITYNESS AMONG SYSTEMS, WHERE N/P STANDS FOR "NO PREFERENCE" AND P DENOTES THE P-VALUE OF A T-TEST BETWEEN TWO VOCODERS.

| | SI | SA | OSA1 | OSA2 | N/P | $p$ |
|---|---|---|---|---|---|---|
| Naturalness | 26.00 | - | - | **40.83** | 33.17 | <0.01 |
| | 33.64 | - | 38.64 | - | 27.72 | 0.13 |
| | 21.88 | **54.22** | - | - | 23.90 | <0.01 |
| | - | **35.00** | - | 28.06 | 36.94 | <0.05 |
| Similarity | 23.50 | - | - | 26.83 | 49.66 | 0.50 |
| | 27.57 | - | 31.82 | - | 40.61 | 0.16 |
| | 15.16 | **38.28** | - | - | 46.56 | <0.01 |
| | - | **26.25** | - | 20.69 | 53.06 | <0.05 |

TABLE IV
THE PEARSON'S CORRELATION COEFFICIENTS BETWEEN UTTERANCE DURATIONS AND EVALUATION METRICS, WHERE $C$ STANDS FOR THE COEFFICIENTS AND $p$ STANDS FOR $p$-VALUE OF SIGNIFICANCE TEST.

| | SNR | RMSE-LAS | MCD | RMSE-F0 | V/UV Error Rate |
|---|---|---|---|---|---|
| C | 0.23 | 0.01 | 0.06 | -0.08 | 0.09 |
| p | <0.01 | 0.80 | 0.30 | 0.20 | 0.13 |

scores than the SI vocoder on both naturalness and similarity, their differences were insignificant ($p > 0.05$), which implies that the subjective performance of OSA1 was still not as good as OSA2. One possible reason is the mismatch between the training and test data of the speaker encoder used by OSA1, which led to the higher EER and F0-RMSE of OSA1 as shown in Table I and II.

### *F. Correlation Analysis between Model Performance and Utterance Duration*

A correlation analysis was conducted to investigate the relationship between the objective performance of the OSA2 vocoder and the duration of test utterances. 288 utterances were randomly selected from the VCTK test set and their durations varied from less than 2 seconds to more than 8 seconds. These utterances were reconstructed into waveforms using the OSA2 vocoder and the five metrics used in Section III.D were calculated for each utterance. The scatter diagrams and the Pearson's correlation coefficients between utterance durations and evaluation metrics are shown in Fig. 2 and Table IV

respectively. We can see that though there were no correlations between utterance durations and most other evaluation metrics, the SNR metric has a significant weak correlation relationship to duration. This indicates the performance of our model is influenced somewhat by duration of test utterances and the longer utterance generated is more likely to achieve better quality. which demonstrates the robustness of our proposed method on the duration of test utterances.

## IV. Conclusions

In this paper, we have proposed an online speaker adaptation method based on a discriminatively-trained speaker encoder and a speaker-aware WaveNet vocoder to improve the performance of traditional speaker-independent neural vocoders. To demonstrate the effectiveness of our proposed model, we also built speaker-independent model (SI) and offline speaker adaptation (SA) vocoders for comparison in our experiments. Experimental results have demonstrated that our method can achieve lower distortion and better naturalness of reconstructed waveforms than the SI vocoder when dealing with unseen speakers. Although this paper focuses on WaveNet vocoders, it is also possible to apply our proposed online adaptation method to other neural vocoders, which will be a task of our future work.

## Acknowledgment

This work was supported by the National Nature Science Foundation of China (Grant No. 61871358).

## References

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[2] 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 sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[3] 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.

[4] 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 *9th ISCA Speech Synthesis Workshop*, pp. 125–125.

[5] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder." in *Interspeech*, vol. 2017, 2017, pp. 1118–1122.

[6] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *ICML*, 2018.

[7] S. Kim, S.-g. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet: A generative flow for raw audio," *arXiv preprint arXiv:1811.02155*, 2018.

[8] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2018.

[9] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[10] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[11] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet vocoder with limited training data for voice conversion." in *Interspeech*, 2018, pp. 1983–1987.

[12] Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-periodic WaveNet vocoder: A pitch dependent dilated convolution model for parametric speech generation," *arXiv preprint arXiv:1907.00797*, 2019.

[13] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 712–718.

[14] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, P. Alku *et al.*, "Speaker-independent raw waveform model for glottal excitation," in *Interspeech*, 2018.

[15] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in DNN-based speech synthesis using d-vectors." in *INTERSPEECH*, 2017, pp. 3404–3408.

[17] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[19] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7942–7946.

[20] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[21] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[22] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[23] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[24] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[25] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *International Conference on Learning Representations*, 2018.

[26] C. Recommendation, "Pulse code modulation (PCM) of voice frequencies," in *ITU*, 1988.

[27] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.