Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions

Dipjyoti Paul[1]*, Yannis Pantazis*[2] *and Yannis Stylianou*[1]

[1]Computer Science Department, University of Crete
[2]Inst. of Applied and Computational Mathematics, Foundation for Research and Technology - Hellas

dipjyotipaul@csd.uoc.gr, pantazis@iacm.forth.gr, yannis@csd.uoc.gr

Abstract

Recent advancements in deep learning led to human-level performance in single-speaker speech synthesis. However, there are still limitations in terms of speech quality when generalizing those systems into multiple-speaker models especially for unseen speakers and unseen recording qualities. For instance, conventional neural vocoders are adjusted to the training speaker and have poor generalization capabilities to unseen speakers. In this work, we propose a variant of WaveRNN, referred to as speaker conditional WaveRNN (SC-WaveRNN). We target towards the development of an efficient universal vocoder even for unseen speakers and recording conditions. In contrast to standard WaveRNN, SC-WaveRNN exploits additional information given in the form of speaker embeddings. Using publicly-available data for training, SC-WaveRNN achieves significantly better performance over baseline WaveRNN on both subjective and objective metrics. In MOS, SC-WaveRNN achieves an improvement of about 23% for seen speaker and seen recording condition and up to 95% for unseen speaker and unseen condition. Finally, we extend our work by implementing a multi-speaker text-to-speech (TTS) synthesis similar to zero-shot speaker adaptation. In terms of performance, our system has been preferred over the baseline TTS system by 60% over 15.5% and by 60.9% over 32.6%, for seen and unseen speakers, respectively.

Index Terms: Universal Vocoder, Speech Synthesis, WaveRNN, Text-to-Speech, Zero-shot TTS.

1. Introduction

Speech synthesis has received attention in the research community as voice interaction systems have been implemented in various applications, such as personalized Text-to-Speech (TTS) systems, voice conversion, dialogue systems and navigations [1, 2, 3, 4]. In the past, conventional statistical parametric speech synthesis (SPSS) exhibited high naturalness under best-case conditions [5, 6]. Hybrid synthesis was also proposed as a way to take advantage of both SPSS and unit-selection approach [7, 8]. Most of these TTS systems consist of two modules: the first module converts textual information into acoustic features while the second one, i.e., the vocoder, generates speech samples from the previously generated acoustic information.

Traditional vocoder approaches mostly involved source-filter model for the generation of speech parameters [9, 10, 11, 12]. The parameters were defined by voicing decisions, fundamental frequency (F0), spectral envelope or band aperiodicities. Algorithms like Griffin-Lim utilized spectral representation to generate speech [13, 14]. However, the speech quality of such vocoders was restricted by the inaccuracies in parameter estimation. Recently, the naturalness of vocoders has been significantly improved by benefiting from direct waveform modeling approach. Neural vocoders like WaveNet utilize a autoregressive generative model that can reconstruct waveform from intermediate acoustic features [15, 16]. To overcome the time complexity at inference, parallel wave generation approach was adopted to generate speech in real time [17, 18]. Wave Recurrent Neural Networks (WaveRNN) which employs recurrent layers increases the efficiency of sampling without compromising their quality [19]. In particular, it can realize real-time high-quality synthesis by introducing a gated recurrent unit (GRU). Although, WaveRNN has been suggested focusing on text-to-speech synthesis, our work exercises it as a vocoder while changing the conditioning criteria from linguistic information to acoustic information. Other recent works have been also found in literature, notable among them are SampleRNN [20], WaveGlow [21], LPCNet [22] and MelNet [23].

Techniques in neural vocoders involve data-driven learning and are prone to specialize to the training data which leads to poor generalization capabilities. Moreover, in multi-speaker scenarios, it is practically impossible to cover all possible in-domain (or seen) and out-of-domain (or unseen) cases in the training database. Previous studies also attempted to improve adaptation capabilities of vocoders [24], either with or without providing speaker information [25, 26]. However, these studies did not address the generalization capabilities for unseen out-of-domain data. In [27], a potential universal vocoder was introduced claiming that speaker encoding is not essential to train a high-quality neural vocoder.

Inspired by the performance and computational aspects of WaveRNN, we propose a novel approach for designing a universal WaveRNN vocoder. The proposed universal vocoder-speaker conditional WaveRNN (SC-WaveRNN) explores the effectiveness of explicit speaker information, i.e., speaker embeddings as a condition and improves the quality of generated speech across broadest possible range of speakers without any adaptation or retraining. Even though conventional WaveRNN is capable of modeling good temporal structure for a single speaker, it fails to capture the dynamics of multiple speakers. We have experimentally demonstrated that our proposed SC-WaveRNN overcomes such limitation by modeling temporal structure from a large variability of data, making it possible to generate high-quality synthetic voices. Our work involves independent training of a speaker-discriminative neural encoder on a speaker verification (SV) task using a state-of-the-art generalized end-to-end loss [28]. The SV model, trained on a large amount of disjoint data, can attain robust speaker representations that are independent of channel conditions and captures large space of speaker characteristics. Coupling such speaker information with the speech synthesis training also reduces the need to obtain ample high-quality multi-speaker training data. At the same time, it increases the model's ability to generalize. Experimental results based on both objective and subjective

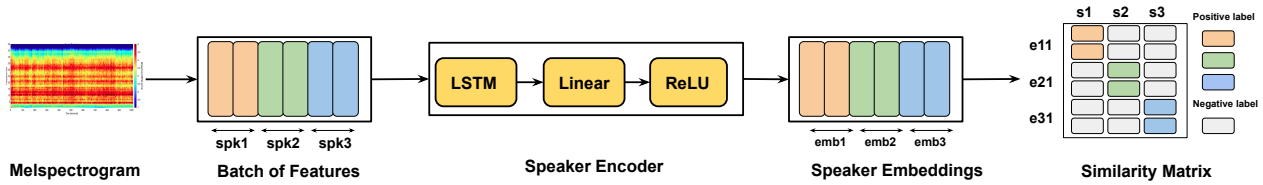

Figure 1: *System overview of speaker encoder [28]. Features, speaker embeddings and similarity scores from different speakers are represented by different color codes. 'spk' denotes speakers and 'emb' represents embedding vectors.*

evaluation confirms that the proposed method achieves better speaker similarity and perceptual speech quality than baseline WaveRNN in both seen and unseen speakers.

In parallel with the above-mentioned studies on universal vocoder, there has been substantial development in multi-speaker TTS where speaker encoder is jointly trained with TTS [29, 30]. These jointly-trained speaker encoders lead to poor inference performance when applied on data which are not included in the training dataset. Fine-tuning pretrained TTS model in combination with speaker embeddings was addressed in [31, 32, 33]. Such approaches always require transcribed adaptation data along with more computational time and resources to adapt to a new speaker. To overcome this, TTS models can be adapted from a few seconds of target speaker's voice in a zero-shot manner by solely using speaker embedding without retraining the entire model. [34, 35, 36].

Unfortunately, limitations still exist and human-level naturalness is not achieved yet. Additionally, prosody information was mismatched especially for unseen speakers. To address those issues, we first train a multi-speaker Tacotron which is conditioned on the speaker embeddings obtained from the independently-trained speaker encoder. Tacotron [37] is a sequence-to-sequence network which predicts mel-spectrograms from text. Next, we incorporate the proposed SC-WaveRNN as a vocoder using the same speaker encoder and synthesize the temporal waveform from the sequence of Tacotron's mel-spectrograms. We compare our system with the baseline TTS method [36] which studies the effectiveness of several neural speaker embeddings in the context of zero-shot TTS. Our results demonstrate that the proposed zero-shot TTS system outperforms baseline zero-shot TTS in [36] in-terms of both speech quality and speaker similarity on both seen and unseen conditions.

2. Neural Speaker Encoder

Our work highlights the importance of speaker encoder in universal vocoders through the application of generalized end-to-end (GE2E) SV task trained on thousands of speakers [28]. The encoder network initially computes frame-level feature representation and then summarizes them to utterance-level fixed-dimensional speaker embeddings. Next, the classifier operates on GE2E loss, where embeddings from the same speaker have high cosine similarity and embeddings from different speakers are far apart in the embedding space. As depicted in Fig. 2, Uniform Manifold Approximation and Projection (UMAP) shows that the speaker embeddings are perfectly separated with large inter-speaker distances and very small intra-speaker variance.

2.1. Training Encoder Network

Speaker encoder structure is depicted in Figure 1. The log mel-spectrograms are extracted from speech utterance of arbitrary window length. The feature vectors are then assembled in the form of a batch that contains S different speakers, and each speaker has U utterances. Each feature vector $\mathbf{x}_{ij}$ ($1 \leq i \leq S$ and $1 \leq j \leq U$) represents the features extracted from speaker i utterance j. The features $\mathbf{x}_{ij}$ are then passed to an encoder architecture. The final embedding vector $\mathbf{e}_{ij}$ is L2 normalized and they are calculated by averaging on each window separately.

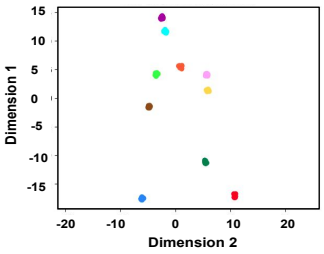

Figure 2: *UMAP projection of 10 utterances for each of the 10 speakers. Different colors represent different speakers.*

2.2. Generalized End-to-End Loss

During training, embedding of all utterance for a particular speaker should be closer to the centroid of that particular speaker's embeddings, while far from other speakers' centroids. The similarity matrix $\mathbf{SM}_{ij,k}$ is defined as the scaled cosine similarities between each embedding vector $\mathbf{e}_{ij}$ to all speaker centroids $\mathbf{c}_k$ ($1 \leq i, k \leq S$ and $1 \leq j \leq U$).

$$\mathbf{SM}_{ij,k} = \begin{cases} w \cdot cos(\mathbf{e}_{ij}, \mathbf{c}_i^{-j}) + b & \text{if } k = i \\ w \cdot cos(\mathbf{e}_{ij}, \mathbf{c}_k) + b & \text{otherwise} \end{cases}$$

$$where\ \mathbf{c}_i^{-j} = \frac{1}{U-1} \sum_{u=1;u\neq j}^{U} \mathbf{e}_{iu}\ and\ \mathbf{c}_k = \frac{1}{U} \sum_{u=1}^{U} \mathbf{e}_{ku}$$

Here, w and b are trainable parameter. The ultimate GE2E loss L is the accumulative loss over similarity matrix ($1 \leq i \leq S$ and $1 \leq j \leq U$) on each embedding vector $\mathbf{e}_{ij}$:

$$L(\mathbf{x}; \mathbf{w}) = \sum_{i,j} L(\mathbf{e}_{ij}) = -\mathbf{SM}_{ij,i} + \log \sum_{k=1}^{S} \exp(\mathbf{SM}_{ij,k})$$

The use of softmax function on similarity matrix makes the output equals to 1 iff $k = i$, otherwise the output is 0.

3. Speaker conditional WaveRNN

In literature, convolutional models have been thoroughly explored and achieved excellent performance in speech synthesis [15, 18] yet they are prone to instabilities. Recurrent neural network (RNN) is expected to provide a more stable high-quality speech due to the persistence of the hidden state.

3.1. Preliminaries

Our WaveRNN implementation is based on the repository[1] which is heavily inspired by WaveRNN training [19]. This architecture is a combination of residual blocks and upsampling network, followed by GRU and FC layers as depicted in Fig. 3. The architecture can be divided into two major networks: conditional network and recurrent network. The conditioning

[1] https://github.com/fatchord/WaveRNN

network consists of a pair of residual network and upsampling network with three scaling factors. At the input, we first map the acoustic features i.e., mel-spectrograms to a latent representation with the help of multiple residual blocks. The latent representation is then split into four parts which will later be fed as input to the recurrent network. The upsampling network is implemented to match the desired temporal size of input signal. The outputs of these two convolutional networks i.e., residual and upsampling networks along with speech are fed into the recurrent network. As part of the recurrent network, two uni-directional GRUs are employed with a few fully-connected (FC) layers at the end. By design, the overhead complexity is reduced with less parameters and takes advantage of temporal context for better prediction.

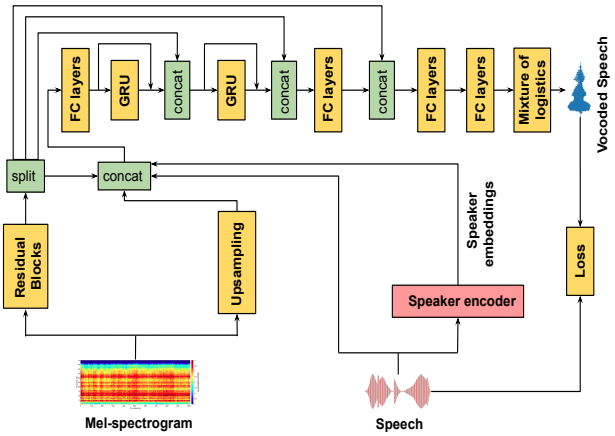

Figure 3: *Block diagram of proposed SC-WaveRNN training.*

3.2. Training WaveRNN with Speaker Embeddings

The above auto-regressive model can generate state-of-the-art natural sounding speech, however, it needs large amounts of training data to train a stable high-quality model and scarcity of data remains a core issue. Moreover, a key challenge is its generalization ability. We observe degradation in speech quality and speaker similarity when the model generates waveforms from speakers that are not seen during training.

In order to assist the development of a stable universal vocoder and remove data dependency, we propose in this paper an alternative training module referred to as speaker conditional WaveRNN (SC-WaveRNN). In SC-WaveRNN, the output of the speaker encoder is used as additional information to control the speaker characteristics during both training and inference. The additional information plays a pivotal role in generating more stable high-quality speech across all speaker conditions. The direct estimation of raw audio waveform $\mathbf{y} = \{y_1, y_1, \cdots, y_N\}$ is described by the conditional probability distribution:

$$sc\text{-}wavernn(\mathbf{y}) = p(y_t|y_{t-1}; \mathbf{h}_t; \mathbf{e}; \lambda)$$

where $\mathbf{e}$ is the 256 dimension speaker embeddings vector. The speaker encoder is independently trained using large diversity of multi-speaker data that can generalize sufficiently to produce meaningful embeddings. The embedding vector $\mathbf{e}$ is computed in a utterance-wise manner. For each utterance, the final embedding vector is averaged over all frames and hence it is fixed for any utterance. The embedding vector is concatenated with the conditional network output and speech samples to form the conditional network. The details of the SC-WaveRNN algorithm is presented in Figure 3. In addition, we apply continuous univariate distribution constituting a mixture of logistic distributions [17] which allows us to easily calculate the probability on the observed discretized value y. Finally, discretized mix logistic loss is applied on the discretized speech.

4. Zero-shot Text-to-Speech

The use of the auxiliary speaker encoder enables us to propose a TTS system capable of generating high-fidelity synthetic voice for unseen speakers without retraining Tacotron and vocoder model. Such speaker adaptation to completely new speakers is called zero-shot learning. This speaker-aware TTS system mimics voice characteristics from a completely unseen speaker with only a few seconds of speech sample.

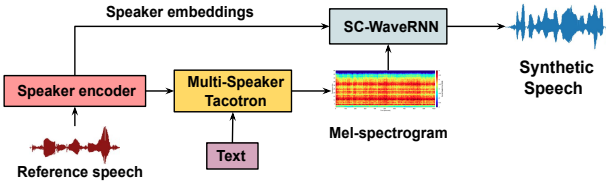

Figure 4: *Block diagram of the proposed zero-shot TTS.*

Our proposed system is composed of three separately trained networks, illustrated in Figure 4: (a) a neural speaker encoder, based on GE2E training, (b) a multi-speaker Tacotron architecture [37], which predicts a mel-spectrogram from text, conditioned on speaker embedding vector, and (c) the proposed speaker conditional WaveRNN, which converts the spectrogram into time domain waveforms. First, the speaker embeddings are extracted from each target speakers' utterance using the speaker encoder. At each time step, the embedding vector for the target speaker is then concatenated with the embeddings of the characters before fed into encoder-decoder module. The final output is mel-spectrograms. To convert the predicted mel-spectrograms into audio, we use SC-WaveRNN which is independently trained by conditioning on the additional speaker embeddings. Due to generalization capabilities of the models, combining multi-speaker Tacotron with SC-WaveRNN can achieve efficient zero-shot adaptation for unseen speakers. We compare the proposed zero-shot system with a recently proposed zero-shot TTS [36] as baseline system. There, the best performing system uses multi-speaker Tacotron with gender-dependent WaveNet vocoders as TTS system and x-vector with learnable dictionary encoding as speaker encoder network.

5. Experimental Setup

The speaker encoder training has been conducted on three public dataset: LibriSpeech, VoxCeleb1 and VoxCeleb2 containing utterances from over 8k speakers [34]. The log mel-spectrograms are first extracted from audio frames of width 25ms and step 10ms. Voice Activity Detection (VAD) and a sliding window approach is used. The GE2E model consists of 3 LSTM layers of 768 cells followed by a projection to 256 dimensions. While training, each batch contains S = 64 speakers and U = 10 utterances per speaker.

Tacotron and WaveRNN models are trained using VCTK English corpus [38] from 109 different speakers. To evaluate generalization performance, we consider three scenarios: seen speakers-seen sound quality (SS-SSQ), unseen speakers-seen sound quality (UNS-SSQ) and unseen speakers-unseen sound quality (UNS-USQ). Seen speakers refers to the speakers that are already present in the training and unseen speakers are the new speakers during testing. Sound quality refers to the recording condition such as recording equipment, reverberation etc. We train the network using 100 speakers leaving 9 speakers for UNS-SSQ scenarios that are chosen to be a mix of genders and

having enough unique utterances per speaker. CMU-ARCTIC database [39] is used for UNS-USQ scenario having 2 male and 2 female speakers. Moreover, to overcome the limited linguistic variability in VCTK data, we initially train Tacotron model on LJSpeech database as a "warm-start" training approach similar to [36]. Code and sound samples can be found in [2].

6. Results and Discussion

6.1. Universal vocoder

In this section, we evaluate the performance of vocoded speech shown in Table 1. To assess the effectiveness of speaker embeddings in SC-WaveRNN, PESQ and STOI objective measures are computed from 50 random samples. We carry out evaluations on three conditions: SS-SSQ, UNS-SSQ and UNS-USQ. The purpose of each condition is to evaluate the proposed vocoder not only on seen or unseen speakers but also for the quality of the recordings. As expected, seen scenarios perform better with respect to unseen samples. However, we observe that SC-WaveRNN significantly improves both the objective scores when compared to baseline WaveRNN for all scenarios.

Table 1: *Objective evaluation tests.*

Methods	SS-SSQ		UNS-SSQ		UNS-USQ	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
WaveRNN	2.2575	0.8173	2.1497	0.7586	1.4850	0.8620
SC-WaveRNN	**2.7948**	**0.9049**	**2.8657**	**0.8984**	**1.8063**	**0.9195**

Concerning the perceptual assessment of speech quality and speaker similarity, two separate listening tests are reported: mean opinion score (MOS) and 'ABX' preference test. The subjects are asked to rate the naturalness of generated utterances on a scale of five-point (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Excellent). In the 'ABX' test, experimental subjects have to decide whether a given reference sentence 'X' is closer in speaker identity to one of 'A' and 'B' sentences, which are samples obtained either from the proposed or the baseline method, not necessarily in that order. Fifteen native and non-native English listeners participated in our listening tests. The evaluation results of both MOS and 'ABX' tests are demonstrated in Figure 5. Error bars represent 95% confidence intervals. For all seen and unseen scenarios, the MOS scores for the proposed SC-WaveRNN are much higher than the baseline WaveRNN (between 14% to 95% relative improvement). Under the same sound quality conditions (SS-SSQ and UNS-SSQ), although, the proposed technique is preferred in terms of speaker similarity preference test, a majority of preference is given to 'same preference' option which indicates similar speaker characteristics for both methods. In contrast, experimental analysis shows a significant preference score (92%) in unseen sound quality for proposed SC-WaveRNN. We conclude that additional speaker information in the form of embeddings is effective for improvements in naturalness and speaker similarity especially for unseen data and capable of achieving a truly universal vocoder. This is attributed by the fact that unseen scenarios are handled more efficiently by the model since additional embeddings are able to capture broad spectrum of speaker characteristics. Moreover, SC-WaveRNN does not compromise the performance in seen conditions.

6.2. Zero-shot TTS Synthesis

To evaluate the performance of the proposed zero-shot TTS, MOS and 'ABX' test are employed, as depicted in Figure 6. We subjectively evaluate both baseline [36] and our methods by synthesizing sample utterances from seen speakers and unseen speakers. Different sound qualities are not considered in the evaluation experiments of zero-shot TTS. As expected, a gap between seen and unseen speakers are visible: seen speakers' synthetic speech has slightly higher quality to unseen speakers. MOS scores indicate that proposed TTS is superior in quality with 19.2% and 14.5% relative improvement for seen and unseen speakers respectively. We also found that our proposed TTS mimic better speaker characteristics and shows significant improvement under both conditions. With regard to speaker similarity, the proposed TTS obtains the majority of preferences with 60% and 60.9% compared to 15.5% and 32.6% of the baseline TTS for seen and unseen speakers, respectively.

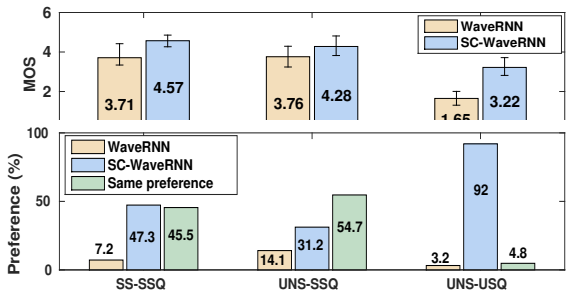

Figure 5: *Vocoder Subjective listening test (MOS) for speech quality and preference test in (%) for speaker similarity.*

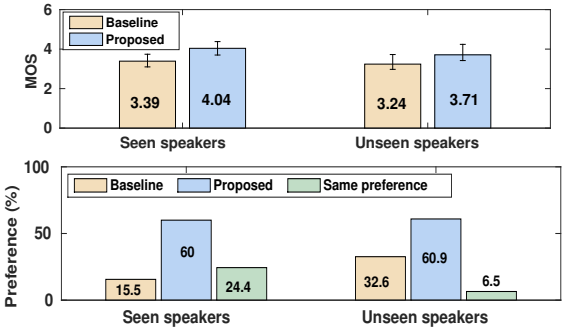

Figure 6: *Zero-shot TTS Subjective listening test (MOS) for speech quality and preference test for (%) for speaker similarity.*

7. Conclusions

In this paper, we proposed a robust universal SC-WaveRNN vocoder that is capable of synthesizing high-quality speech. The system was conditioned on extracted speaker embeddings which cover a very diverse range of seen and unseen conditions. The main advantage of SC-WaveRNN is its high controllability, since it improves multi-speaker vocoder training along with better generalization ability by allowing reliable transfer to unseen speaker characteristics. Furthermore, speaker conditioning is typically more data efficient and computationally less expensive than training separate models for each speaker. Subjective and objective evaluation revealed that the proposed method generated higher sound quality and speaker similarity than the baseline method. In addition, we extended our approach in devising an efficient zero-shot TTS system. We demonstrated that the proposed zero-shot TTS with universal vocoder can improve speaker similarity and naturalness of synthetic speech for seen and unseen speakers. In future, we list more experimentation on the construction of speaker embeddings and its effectiveness in other applications with unseen data.

Acknowledgements: The work has received funding from the EUs H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu).

[2] https://dipjyoti92.github.io/SC-WaveRNN/

8. References

[1] T. Dutoit, *An introduction to text-to-speech synthesis*. Springer Science & Business Media, 1997, vol. 3.

[2] P. Taylor, *Text-to-speech synthesis*. Cambridge University Press, 2009.

[3] 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.

[4] D. Paul, Y. Pantazis, and Y. Stylianou, "Non-parallel voice conversion using weighted generative adversarial networks." in *Proc. Interspeech*, 2019, pp. 659–663.

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

[6] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

[7] Y. Qian, F. K. Soong, and Z.-J. Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 280–290, 2012.

[8] T. Merritt, R. A. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *Proc. ICASSP*, 2016, pp. 5145–5149.

[9] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[10] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[11] 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.

[12] 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.

[13] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[14] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin-Lim algorithm," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[15] 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.

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

[17] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 3918–3926.

[18] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2019.

[19] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*, 2018, pp. 2410–2419.

[20] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *International Conference on Learning Representations*, 2017.

[21] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, 2019, pp. 3617–3621.

[22] J. M. Valin and J. Skoglund, "LPCnet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.

[23] S. Vasquez and M. Lewis, "MelNet: A generative model for audio in the frequency domain," *arXiv preprint arXiv:1906.01083*, 2019.

[24] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder." in *Interspeech*, 2018, pp. 1978–1982.

[25] 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.

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

[27] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards achieving robust universal neural vocoding," in *Proc. Interspeech*, 2019, pp. 4879–4883.

[28] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*, 2018, pp. 4879–4883.

[29] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie *et al.*, "Sample efficient adaptive text-to-speech," in *International Conference on Learning Representations*, 2018.

[30] J. Park, K. Zhao, K. Peng, and W. Ping, "Multi-speaker end-to-end speech synthesis," *arXiv preprint arXiv:1907.04462*, 2019.

[31] Y. Deng, L. He, and F. Soong, "Modeling multi-speaker latent space to improve neural TTS: Quick enrolling new speaker and enhancing premium voice," *arXiv preprint arXiv:1812.05253*, 2018.

[32] Q. Hu, E. Marchi, D. Winarsky, Y. Stylianou, D. Naik, and S. Kajarekar, "Neural text-to-speech adaptation from low quality public recordings," in *Speech Synthesis Workshop*, vol. 10, 2019.

[33] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 019–10 029.

[34] 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.

[35] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," in *Proc. Interspeech*, 2019, pp. 2105–2109.

[36] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. ICASSP*, 2020, pp. 6184–6188.

[37] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint:1703.10135*, 2017.

[38] V. Christophe, Y. Junichi, and M. Kirsten, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *The Centre for Speech Technology Research (CSTR)*, 2016.

[39] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *5th ISCA workshop on speech synthesis*, 2004.