Interspeech 2018
2-6 September 2018, Hyderabad

Multi-task WaveNet: A Multi-task Generative Model for Statistical Parametric Speech Synthesis without Fundamental Frequency Conditions

Yu Gu, Yongguo Kang

Baidu Speech Department
Baidu Technology Park, Beijing, 100193, China
guyu04@baidu.com, kangyongguo@baidu.com

Abstract

This paper introduces an improved generative model for statistical parametric speech synthesis (SPSS) based on WaveNet under a multi-task learning framework. Different from the original WaveNet model, the proposed Multi-task WaveNet employs the frame-level acoustic feature prediction as the secondary task and the external fundamental frequency prediction model for the original WaveNet can be removed. Therefore the improved WaveNet can generate high-quality speech waveforms only conditioned on linguistic features. Multi-task WaveNet can produce more natural and expressive speech by addressing the pitch prediction error accumulation issue and possesses more succinct inference procedures than the original WaveNet. Experimental results prove that the SPSS method proposed in this paper can achieve better performance than the state-of-the-art approach utilizing the original WaveNet in both objective and subjective preference tests.

Index Terms: WaveNet, multi-task learning, statistical parametric speech synthesis

1. Introduction

Text-to-speech (TTS) synthesis involves generating intelligible and natural sounding synthetic speech waveforms given the input text messages. At present, TTS synthesis technique is an indispensable basic component in various applications with speech interface such as car navigation systems, speech-to-speech translation, voice assistant and screen readers, etc. Thus the demands and expectations for high-quality, high-naturalness, more stable and expressive speech waveform synthesis algorithms will be increasing more and more in future.

Statistical parametric speech synthesis (SPSS) [1] is one of the mainstream speech synthesis techniques in which statistical models are employed to model the complex mapping relationship between the input linguistic information and the corresponding acoustic features. Decision-tree-clustered context-dependent hidden Markov model (HMM) speech synthesis [2] with single Gaussian state-output distribution has dominated SPSS in the past decade. In recent years, deep learning technology has been intensively studied and adopted in many speech generation tasks [3]. Different kinds of stochastic neural networks with various deep structures have also been utilized as the acoustic models to replace the decision trees and the probability density functions over acoustic features in conventional HMM-based SPSS methods. These models can model the relationship between input features and acoustic features more accurately than conventional HMMs and Gaussian mixture models and they can be used to model high-dimensional spectra directly. Unidirectional or bidirectional recurrent neural networks (RNNs) incorporating long short-term memory (LSTM) cells of inherently strong ability in capturing long range temporal dependencies were also exploited on speech synthesis systems to produce higher quality and smoother speech trajectories than conventional deep neural networks [4, 5] .

Comparing with unit-selection speech synthesis, SPSS is more flexible and has the advantages on TTS tasks for a small-scale and mildly-curated speech corpus and speech synthesis on mobile devices. The synthesized speech by SPSS is much more stable and can effectively alleviate the discontinuity which is a representative drawback of speech synthesis based on unit selection and waveform concatenation. Nevertheless, due to the limits of some factors such as vocoder quality, modeling accuracy and over-smoothing effect, the quality and similarity of generated speech from SPSS are still far from those of natural speech. WaveNet [6] proposed recently by DeepMind is one of the state-of-the-art generative models in speech generation area. WaveNet and its variants with similar dilated convolutional network structures have achieved huge success on multiple audio generation tasks besides speech synthesis such as speech enhancement [7], voice conversion [8], singing synthesis [9, 10], speech bandwidth extension [11] and neural vocoders [12, 13, 14]. Distinguished from conventional frame-based SPSS approaches, WaveNet can model the speech waveforms directly using dilated causal convolutional neural networks (CNNs) instead of vocoders. It has been proved that WaveNet was capable of producing significantly more natural sounds than conventional SPSS approaches [6, 15], therefore WaveNet is an effective and promising solution to fill the gap of quality between natural and synthesized speech in SPSS.

Original WaveNets for the SPSS task were locally conditioned on the logarithmic fundamental frequency (log F0) values in addition to the linguistic features. The pitch or fundamental frequency information of speech is crucial for waveform generation because it represents the periodicity of speech waveforms. The absence of the log F0 condition information could degrade the naturalness of generated speech and lead in severe intonation mistakes [6], therefore a different auxiliary model for predicting log F0 contours from linguistic features was quite essential for waveform generation in WaveNet. However the involvement of an external F0 prediction model made the WaveNet inference procedures much more complicated. The error of the pitch determinations and the mistakes for voiced/unvoiced decisions in the F0 prediction model could also bring in inaccurate or wrong input condition information for WaveNet, which directly alleviated the naturalness and quality of the synthesized speech, even though WaveNet model itself was accurate. In this paper, an improved WaveNet for SPSS under the multi-task learning framework is proposed to address these issues. The proposed Multi-task WaveNet can get rid of the redundant F0 prediction model by deploying the F0 prediction as a secondary task. Both objective and subjective tests

10.21437/Interspeech.2018-1506

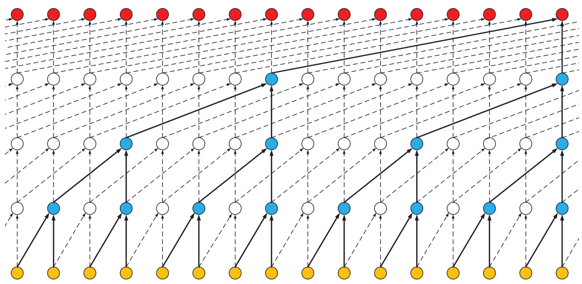


Figure 1: *Dilated causal convolutional network structures in WaveNet.*

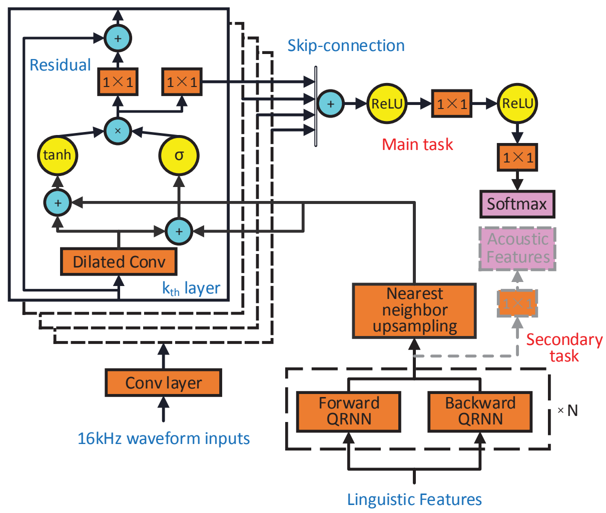


Figure 2: *Network structures of Multi-task WaveNet.*

show that the proposed Multi-task WaveNet can outperform the original WaveNet and can obtain smaller objective distortion and better subjective preference results.

This paper is organized as follows. Section 2 gives a brief review of the model structure of WaveNet. Section 3 introduces the proposed Multi-task WaveNet in this paper and the constructed SPSS system in detail. The experimental conditions and results are described in Section 4 and finally Section 5 concludes this paper.

2. WaveNet

WaveNet [6] is a generative model which was proposed for TTS synthesis and other general audio generation tasks. WaveNet performed autoregressive speech sample generation using an acoustic model with stacked dilated causal convolutional layers instead of depending on vocoders. The architectures of exploited dilated CNNs are illustrated on Figure 1. The causal convolutional layers have various dilation factors that allow their receptive field to grow exponentially in terms of the depths of networks as opposed to linearly, and can therefore cover the input history information from thousands of timesteps ahead. The dilated causal CNN can be regarded as a statistical model and the conditional distribution of the output sample sequence $\boldsymbol{x} = \{x_1, x_2, \cdots, x_T\}$ given the input local condition sequence $\boldsymbol{c}$ is factorised as the product of conditional probabilities as follows:

$$p\left(\boldsymbol{x} \mid \boldsymbol{c}\right) = \prod_{i=1}^{T} p\left(x_i \mid x_{i-N+1}, x_{i-N+2}, \cdots, x_{i-1}, \boldsymbol{c}\right), \quad (1)$$

where N is the length of the receptive field.

Residual learning strategies [16] were also applied on the dilated CNNs in WaveNet to address the issues of training accuracy degradation and slow convergence. Each convolutional layer in WaveNet is wrapped in a residual block which contains gated activation units and two additional convolutional layers towards the following and output layers respectively with convolution filters of size 1. The residual and parameterized skip-connections are deployed throughout the network to capacitate training deeper networks and to accelerate convergence. The gated activation units with condition sequence $\boldsymbol{c}$ in k-th layer are expressed as:

$$\begin{aligned}\hat{\boldsymbol{h}}_k = &\tanh(\boldsymbol{W}_{f,k} * \boldsymbol{h}_k + \boldsymbol{V}_{f,k} * \boldsymbol{c}) \odot \\ &\sigma(\boldsymbol{W}_{g,k} * \boldsymbol{h}_k + \boldsymbol{V}_{g,k} * \boldsymbol{c}),\end{aligned} \quad (2)$$

where f and g denote the filter and gate parts respectively, σ is the sigmoid non-linearity function, $\odot$ is the element-wise product and $*$ is the convolution operator. The output layer is cascaded with a softmax layer and thus the model can describe the categorical distribution of waveform sample values encoded by μ-law algorithm [17].

3. Multi-task WaveNet

3.1. Multi-task learning

Multi-task learning (MTL) [18] is a very useful machine learning strategy which aims at improving model generalization ability and performance by jointly learning several different but related tasks. The main task usually shares a part of the representation with other secondary tasks and the secondary tasks can contribute to the model training of the corresponding main task by supplementing information, transferring knowledge and increasing the amount of training data. MTL approaches can be easily deployed on stochastic neural networks by sharing certain hidden layers cross different tasks. MTL has obtained great achievements on multiple speech signal processing areas such as speech synthesis [19] and automatic speech recognition [20].

In this paper, the MTL strategy is also employed on WaveNet to assist the network training. For the SPSS method based on the original WaveNet, the core task is to generate the speech samples in an autoregressive manner given the input linguistic features and log F0 values, meanwhile the task of the conventional frame-based SPSS approach is to predict the frame-level acoustic features extracted by vocoders from the input linguistic features. These two different tasks share a part of the same inputs and the outputs of the two tasks are highly correlated, therefore the task of acoustic feature prediction can be treated as the secondary task for WaveNet. As exhibited in Figure 2, the secondary task shares the same conditional network with WaveNet. The output acoustic features of the secondary task consist of spectral features and fundamental frequency information, thus the conditional network of the trained Multi-task WaveNet possesses the representation capacity of fundamental frequencies corresponding with the target speech waveforms. Due to the pitch information complement from the secondary task, the linguistic features can be directly utilized as the condition information instead of concatenating with log F0 values in Multi-task WaveNet. The proposed Multi-task WaveNet owns several advantages over the original WaveNet. Firstly, the F0 prediction model is no longer needed, which will greatly simplify the inference procedures of WaveNet-based SPSS. Secondly, the inconsistency of using natural F0s for training and using predicted F0s for testing in the original WaveNet doesn't exist in the proposed model and the error accumulation problem from the F0 prediction model is also tackled. Finally, the secondary task is also conducive to accelerating the convergence of WaveNet training.

3.2. Quasi-recurrent network condition

Similar with the network architectures in Deep Voice [15], quasi-recurrent neural networks (QRNNs) [21] with a stack of bidirectional RNN layers are employed in Multi-task WaveNet as the conditional networks to encode the linguistic input information. A unidirectional QRNN layer with fo-pooling [21] is defined by the following equations:

$$\hat{\boldsymbol{h}} = \tanh(\boldsymbol{W}_h * \boldsymbol{x} + \boldsymbol{B}_h), \tag{3}$$

$$\boldsymbol{o} = \sigma(\boldsymbol{W}_o * \boldsymbol{x} + \boldsymbol{B}_o), \tag{4}$$

$$\boldsymbol{f} = \sigma(\boldsymbol{W}_f * \boldsymbol{x} + \boldsymbol{B}_f), \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{f}_t \cdot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{f}_t) \cdot \hat{\boldsymbol{h}}_t, \tag{6}$$

$$\boldsymbol{z}_t = \boldsymbol{o}_t \cdot \boldsymbol{h}_t, \tag{7}$$

where $*$ denotes the convolution operator. Bidirectional QRNN layer is computed by running two unidirectional QRNNs, one on the input sequence and one on a reverse of the input sequence. Following the bidirectional QRNN layers, the encoded information is upsampled to the same time resolution with native audio frequency by repetition. QRNNs allow for parallel computation across the timestep dimension and have faster training and testing speeds than conventional RNNs.

3.3. SPSS using Multi-task WaveNet

At the training stage of the SPSS method based on Multi-task WaveNet, a conventional decision tree-clustered context-dependent HMM-based SPSS model is trained to acquire the alignment information of the corpus. And then a duration model based on a bidirectional RNN with stacked LSTM layers is trained using the obtained alignment information. The input linguistic features for WaveNet include some one-hot features for categorical linguistic contexts (e.g. phonemes identities, stress marks) and some numerical features for numerical linguistic contexts (e.g. the number of syllables in a word, the position of the current frame in the current phoneme). The output acoustic features for the secondary task including mel-cepstral coefficients (MCCs) and F0 values are extracted from natural speech by vocoders. Similar with the original WaveNet, all the input and output waveforms are quantized to 256 discrete values using μ-law and one-hot coding is pursued on the quantized waveforms as the input waveform sequence. For the main task, the network training is based on cross-entropy criterion to iteratively improve the classification accuracy of output samples with the target output sample sequences in training set. For the secondary task, minimum mean squared error criterion is applied to estimate the MCCs, log F0 and voice/unvoice flag.

At the stage of inference, the phone-level linguistic features generated by front-end text analysis are delivered to the duration prediction neural network to produce the duration information. Then combining with the duration information, WaveNet model can perform autoregressive speech sample generation conditioned on the frame-level linguistic features and the secondary task part as shown by the grey dashed line in Figure 2 is abandoned at the speech synthesis phase.

Table 1: *Comparison of distortion between acoustic features of natural speech and synthesized speech from different systems. V/UV means frame-level voiced/unvoiced error. BAP and Corr. represents the BAP prediction error and correlation coefficients respectively.*

(a) Results for *Corpus A*

System	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	F0 Corr.	V/UV (%)
LSTM	2.265	2.131	30.485	0.860	5.264
WaveNet-lin	1.524	2.798	39.413	0.756	4.745
WaveNet	1.548	2.787	32.206	0.845	4.796
MTL-WaveNet	**1.519**	2.712	**22.396**	**0.922**	**4.298**

(b) Results for *Corpus B*

System	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	F0 Corr.	V/UV (%)
LSTM	1.641	2.053	37.021	0.787	4.328
WaveNet-lin	1.515	2.683	41.745	0.756	3.588
WaveNet	1.500	2.681	37.790	0.775	3.592
MTL-WaveNet	**1.481**	2.653	**33.801**	**0.821**	3.682

4. Experiments

4.1. Experimental conditions

To evaluate the performance of the proposed Multi-task WaveNet, two different scales of speech database pronounced by two Chinese female speakers respectively were used in the experiments. The large-scale corpus contained 16.2 hours of speech data, which was sufficient for unit-selection synthesis and was named as " *Corpus A* ". The small-scale corpus named as "*Corpus B*" only included 2.1 hours of speech data whose corresponding transcripts were mainly designed for car navigation. The sampling rate of the speech data was 16 kHz and 50 test sentences those were not present or similar to those in the training set were used as the test set to measure the performance of different speech synthesis systems for evaluation. For the main task, the WaveNet model consisted of 40 layers, which were grouped into 4 dilated residual block stacks of 10 layers. In every stack, the dilation rate exponentially increased by a factor of 2 in every layers which started with rate 1 and ended with the maximum dilation of 512. For the secondary task, 25-dimensional MCCs were extracted from the smoothed spectral envelopes obtained by STRAIGHT analysis [22] and the F0 values were extracted by RAPT algorithm [23].

The following five speech synthesis systems were established for comparison.

- ***LSTM***: The SPSS system using bidirectional RNNs with stacked LSTM layers as duration and acoustic models;
- ***Concatenative***: The HMM-driven unit selection concatenative speech synthesis;
- ***WaveNet***: The original WaveNet SPSS system with F0s and linguistic features as WaveNet conditions;
- ***WaveNet-lin***: The original WaveNet SPSS system only conditioned on linguistic features;
- ***MTL-WaveNet***: The proposed SPSS system using Multi-task WaveNet.

4.2. Objective evaluation

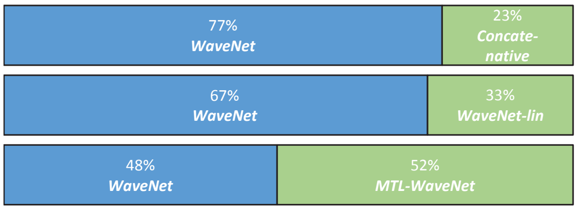
Objective tests were conducted to evaluate different synthesis systems. Because the duration of synthetic speech in the ***Concatenative*** system was intractable to adjust, the ***Concatenative*** system was excluded from the objective evaluations and the other SPSS systems remained the same ground-truth duration as the target natural speech for the convenience of comparison. Mel-cepstral distortion (MCD), distortion of band aperiodicities (BAP), voiced/uvoiced prediction error, root-mean-square error (RMSE) and correlation coefficients of F0 values on a linear scale between natural speech and synthesized speech by different SPSS systems are presented in Table 1. It is worth mentioning that for the ***WaveNet***, ***WaveNet-lin*** and ***MTL-WaveNet*** systems, the compared acoustic features were re-extracted from the generated waveforms, while those were directly the model outputs for the ***LSTM*** system.

The objective results for both two speech databases show that the synthesized speech from the ***MTL-WaveNet*** system can acquire more accurate F0 values and smaller spectral distortion. The F0 RMSE of the ***WaveNet-lin*** system is the largest among all the systems, which indicates the log F0 conditional information is quite essential for the original WaveNet. The F0 prediction error of the ***WaveNet*** system is also larger than the conventional ***LSTM*** system which means the inaccuracy of the F0 prediction model can be accumulated into the waveform generation step of WaveNet and further impact the synthesized speech quality in the ***WaveNet*** system.

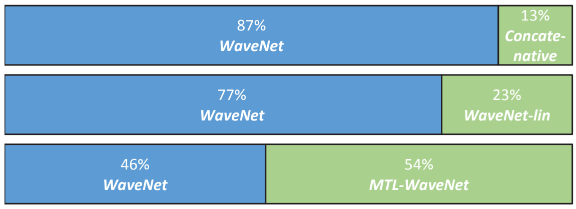
4.3. Subjective evaluation

Several preference tests were performed to assess the subjective perceptual quality of the speech generated from different speech synthesis systems. In each preference test, 20 test utterances randomly selected from the test set were synthesized by two different systems and evaluated in random order by five listeners. Because the speech quality of the ***LSTM*** system was far from other systems, we only compared the preference performances among the ***Concatenative***, ***WaveNet-lin***, ***WaveNet*** and ***MTL-WaveNet*** systems. The listeners were asked to choose their preference for each pairwise utterances in terms of speech quality.[1] The preference scores of listening tests conducted in two databases of different scales are exhibited in Figure 3 respectively with the p-values from t-test. The comparisons of the conventional ***Concatenative*** system and the ***WaveNet*** systems demonstrate the WaveNet based method using waveform modeling and causal dilated CNNs can successfully improve the quality of synthesized speech. Although the candidate and selected units for the ***Concatenative*** system are natural speech segments in the corpus, the discontinuity of the synthesized speech in unit-selection synthesis seriously degrades the perceptual quality and naturalness. The gap between the ***Concatenative*** and ***WaveNet*** systems for *Corpus B* is much larger than that for *Corpus A*, which is because the data volume of *Corpus B* is not sufficient for building a unit-selection synthesis system and there are more bad cases than *Corpus A*. The data size can also affect the synthesized speech quality for the approaches based on WaveNets, while the differences are not as obvious as those for unit-selection synthesis. The comparisons between the ***WaveNet-lin*** and ***WaveNet*** systems can also prove the importance of using log F0 as a part of condition information in original WaveNet SPSS methods. We find that WaveNet only conditioned on linguistic features can synthesize natural waveforms for the most part of one utterance but sometimes it has unnatural phones and syllables by pronouncing wrong and strange tones. The superiority of the ***MTL-WaveNet*** system over the ***WaveNet*** system on preference scores indicates the effectiveness of utilizing the acoustic feature prediction as the secondary task. The improvement of preference scores for the ***MTL-WaveNet*** is less significant than those of objective tests. In fact, the generated speech from the ***WaveNet*** system is generally good enough and the advantages of the proposed Multi-task WaveNet are embodied on some certain speech segments and some speech details which are easily ignored by listeners.

[1]Examples of synthesized speech by different systems are available at https://ttsdemos.github.io.



(a) Results for *Corpus A*. The p-values of t-test in these comparisons are 5.6×10^{-9}, 8.8×10^{-7} and 0.69 respectively.



(b) Results for *Corpus B*. The p-values of t-test in these comparisons are 9.5×10^{-19}, 5.6×10^{-9} and 0.43 respectively.

Figure 3: *Preference test scores among different TTS systems.*

5. Conclusions

In this paper, we propose an improved WaveNet model for SPSS exploring the multi-task learning strategy. The frame-level acoustic feature prediction is introduced as the auxiliary secondary task to supplement the requisite pitch information. Comparing with the original WaveNet based SPSS approach, the proposed Multi-task WaveNet can get rid of the redundant F0 prediction model and increase the inference efficiency. This model can also solve the pitch prediction error accumulation problems. Both objective and subjective evaluation results show that the SPSS method proposed in this paper have the advantages over the original WaveNet. To achieve more natural prosody and more expressive speech, some end-to-end speech synthesis models will be tried to further replace the duration models for WaveNets in our future work. We will deploy such multi-task learning structure on Parallel WaveNet to further accelerate the inference speed and apply the proposed algorithms on online products.

6. Acknowledgements

The authors thank Xian Li for his early constructive contribution to this work, as well as Zhenyu Wang, Hao Li, Xiaohui Sun, Feiya Li of Baidu Speech Department for their assistance with data preprocessing and running subjective evaluations. The authors also thank Wei Ping and Kainan Peng of Baidu Silicon Valley AI Lab for their helpful discussions and advices.

7. References

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

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[3] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep Learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 35–52, May 2015.

[4] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *INTERSPEECH 2014*, September 2014, pp. 1964–1968.

[5] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4470–4474.

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

[7] D. Rethage, J. Pons, and X. Serra, "A WaveNet for speech denoising," *arXiv preprint arXiv:1706.07162*, 2017.

[8] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," *INTERSPEECH 2017*, pp. 1138–1142, 2017.

[9] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," *INTERSPEECH 2017*, pp. 4001–4005, 2017.

[10] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," *arXiv preprint arXiv:1704.01279*, 2017.

[11] Y. Gu and Z.-H. Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension," *INTERSPEECH 2017*, pp. 1123–1127, 2017.

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

[13] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep Voice 2: Multi-speaker neural text-to-speech," *arXiv preprint arXiv:1705.08947*, 2017.

[14] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: 2000-speaker neural text-to-speech," *arXiv preprint arXiv:1710.07654*, 2017.

[15] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep Voice: Real-time neural text-to-speech," in *International Conference on Machine Learning*, 2017, pp. 195–204.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] R. G. ITU-T and I. Switzerland, "711. pulse code modulation (PCM) of voice frequencies," *International Telecommunication Union, Geneva, Switzerland*, 1988.

[18] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[19] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4460–4464.

[20] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[21] J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural networks," *arXiv preprint arXiv:1611.01576*, 2016.

[22] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[23] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.