SPEAKER-ADAPTIVE NEURAL VOCODERS FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS SYSTEMS

Eunwoo Song[1,2], *Jinseob Kim*[1], *Kyungguen Byun*[2] *and Hong-Goo Kang*[2]

[1]NAVER Corp., Seongnam, Korea
[2]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

ABSTRACT

This paper proposes speaker-adaptive neural vocoders for statistical parametric speech synthesis (SPSS) systems. Recently proposed WaveNet-based neural vocoding systems successfully generate a time sequence of speech signal with an autoregressive framework. However, building high-quality speech synthesis systems with limited training data for a target speaker remains a challenge. To generate more natural speech signals with the constraint of limited training data, we employ a speaker adaptation task with an effective variation of neural vocoding models. In the proposed method, a speaker-independent training method is applied to capture universal attributes embedded in multiple speakers, and the trained model is then fine-tuned to represent the specific characteristics of the target speaker. Experimental results verify that the proposed SPSS systems with speaker-adaptive neural vocoders outperform those with traditional source-filter model-based vocoders and those with WaveNet vocoders, trained either speaker-dependently or speaker-independently.

Index Terms— Statistical parametric speech synthesis, WaveNet, ExcitNet, neural vocoder, speaker adaptation

1. INTRODUCTION

Waveform generation systems using WaveNet have attracted a great deal of attention in the speech signal processing community thanks to their high performance and ease of use in various applications [1, 2]. In a system of this kind, the time domain speech signal is represented as a sequence of discrete symbols, and its probability distribution is autoregressively modeled by stacked convolutional neural networks (CNNs). By appropriately conditioning the acoustic features to the input, WaveNet-based systems have also been successfully adopted in a neural vocoder structure for statistical parametric speech synthesis (SPSS) [3–7].

To further improve the perceptual quality of the synthesized speech, more recent neural excitation vocoders (e.g. ExcitNet [8]) take advantages of the merits from both the parametric LPC vocoder and the WaveNet structure [9, 10]. In this framework, an adaptive predictor is used to decouple the formant-related spectral structure from the input speech signal, and the probability distribution of its residual signal (i.e. the excitation signal) is then modeled by the WaveNet network. As variation in the excitation signal is only constrained by vocal cord movement, the training and generation processes become much more efficient. As such, SPSS systems with neural excitation vocoders reconstruct more accurate speech signals than conventional parametric or WaveNet vocoders [9, 10].

However, this approach to wave generation still requires large amounts of training data to faithfully represent the complex mechanics of human speech production. As a result, unnatural outputs are generated when the training data for the target speaker is inadequate (e.g. a database comprising less than ten minutes' speech). The speaker-independent training method of utilizing multiple speakers for a single unified network shows the feasibility of generating multiple voices by conditioning the target speaker's acoustic features [4]. However, our preliminary experiments verify that this approach still generates discontinuous speech segments if the target speaker's data is not included in the training process. This problem is more prominent under an SPSS framework where prediction errors in estimating auxiliary parameters are inevitable; prediction errors are propagated throughout the autoregressive generation process.

To alleviate this problem, we propose a speaker-adaptive training method for neural vocoding systems. In this framework, to address the lack of speaker-specific information caused by limited training data for a target speaker, a model is trained independently of the target speaker to extract universal attributes from multiple speakers [4]. This model is then used to initialize training of the target speaker's model, and all weights are fine-tuned to represent the distinctive characteristics within the target's database. Because this adaptation process helps the deep neural network capture speaker-dependent characteristics, it is also advantageous in reducing the discontinuity problems that occur in conventional speaker-independent models.

We investigate the effectiveness of the proposed method by conducting objective and subjective evaluations comparing systems designed both dependently and independently of the target speaker. In speech analysis and synthesis, and under SPSS conditions, the experimental results show that the speaker-adaptive neural vocoder significantly improves the perceptual quality of synthesized speech compared to conventional methods.

2. RELATIONSHIP TO PRIOR WORK

The idea of using WaveNet in neural vocoders for speech synthesis frameworks is not new. By effectively representing sample-by-sample correspondence between acoustic features and speech waveform, the WaveNet vocoder has successfully replaced the role of traditional parametric vocoders [3–7]. Based on the human auditory system [11], neural vocoders using frequency-dependent spectral shaping filters have further improved the perceptual quality of synthesized speech [9, 10, 12]. However, the speech signals generated by neural vocoders often suffer from unnatural outputs when the amount of training data for the target speaker is insufficient. Although employing a multi-speaker training method enables the general speech signal characteristics to be represented [4], capturing the specific nature of the target speaker's own speech is limited.

To ameliorate these issues, our aim is to use an adaptation task to improve the modeling accuracy of WaveNet-based neural vocoders. Although prior work in using speaker-adaptive WaveNet vocoders in voice conversion applications has been undertaken [13, 14], our research differs from these studies in several ways: First, we focus on the effect of the speaker adaptation in SPSS tasks. In this study, we verify the effectiveness of the proposed method not only in speech analysis/synthesis but also within an SPSS framework. Second, our experiments seek to verify the superior performance of speaker-adaptive training methods as compared with conventional speaker-dependent and speaker-independent approaches. Furthermore, the synthesis quality of each training method is investigated across various types of vocoder, for example, a plain WaveNet framework and a neural excitation vocoder, namely ExcitNet. Both the objective and subjective test results provide helpful guidelines for the design of similarly configured vocoding systems. Third, in terms of the perceptual quality of the vocoding, the proposed method shows superiority over the other approaches under the same SPSS model structure. And finally, we also explore the effectiveness of the proposed method in a pitch or fundamental frequency (F0) modification task. Experiments in arbitrary changes to F0 contours confirm that the proposed speaker-adaptive training method synthesizes the modified F0 sound very reliably compared to conventional speaker-dependent approaches.

3. NEURAL VOCODERS

3.1. WaveNet-based neural vocoding frameworks

The basic WaveNet framework is an autoregressive network which generates a probability distribution of waveforms from a fixed number of past samples [2]. By conditioning the model on additional input variables, the output can be guided to produce waveforms with required characteristics. Recent *WaveNet vocoders* directly utilize acoustic features as the conditional input where these features are extracted from conventional parametric vocoders [3–7, 12]. This enables the WaveNet system to automatically learn the relationship between acoustic features and speech samples which results in superior perceptual quality over traditional parametric vocoders [3, 15].

However, due to the inherent structural limitations of CNNs in terms of capturing the dynamic nature of speech signals, the WaveNet approach often generates noisy outputs caused by distortion in the spectral valley regions. To improve the perceptual quality of synthesized speech, several frequency-dependent noise-shaping filters have been proposed [9, 10, 12]. In particular, the neural excitation vocoder *ExcitNet* (described in Fig. 1a) exploits linear prediction (LP)-based adaptive predictor to decouple the spectral formant structure from the input speech signal. The WaveNet-based generation model is then used to train the residual LP component (i.e. the excitation signal). As variation in the excitation signal is only constrained by vocal cord movement, the training process becomes much more effective.

In the speech synthesis step shown in Fig. 1b, the acoustic parameters of the given input are first generated by an acoustic model designed with a conventional deep learning SPSS system [16]. Those parameters are used as auxiliary conditional features for the WaveNet model to generate the corresponding time sequence of the excitation signal. Ultimately, the speech signal is reconstructed by passing the generated excitation signal through the LP synthesis filter. In this way, the quality of the synthesized speech signal is further improved because the spectral component is well represented by the deep learning framework and the residual component is efficiently generated by the WaveNet framework.

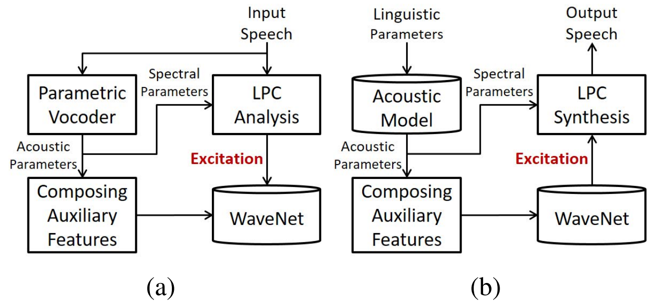

Fig. 1: ExcitNet vocoder framework for an SPSS system: (a) training and (b) synthesis.

3.2. Speaker-adaptive neural vocoders

The superiority of neural vocoding systems over traditional parametric vocoders has been explained above but it is still challenging to build a high-quality speech synthesis system when the training data for a target speaker is inadequate, for example with just ten minutes of speech. One of the approaches to address this problem is to speaker-independently train the neural vocoder [4]. This approach is capable of teaching global characteristics embedded in the speech of multiple speakers. However, its output often becomes discontinuous when the unseen speaker's acoustic features are used as conditional features. The quality loss becomes severe when the input acoustic parameters are estimated by the SPSS framework which is necessary in a text-to-speech configuration.

To generate a more natural speech signal with limited training data, we employ an adaptation task in training the neural vocoder. In the proposed framework, a speaker-independently trained multi-speaker model is used as the initializer, and then all weights are updated in training the target speaker's model. Additional merits of the proposed method can be found in its robust performance in a pitch modification task because its initial model shares the different attributes from multiple speech databases. Furthermore, the fine-tuning mechanism of the adaptation process enables capture of speaker-specific characteristics from the target's data set; the proposed method therefore improves the vocoding performance despite insufficient training data.

4. EXPERIMENTS

4.1. Experimental setup

To investigate the effectiveness of the proposed algorithm, we trained neural vocoding models using three different methods:

- SD: speaker-dependent training model
- SI: speaker-independent training model
- SA: speaker-adaptive training model

In the SD and SA models, speech corpora recorded by a Korean female speaker were used. The speech signals were sampled at 24 kHz, and each sample was quantized by 16 bits. In total, 90

(10 min), 40 (5 min), and 130 (15 min) utterances were used for training, development, and testing, respectively. To train the SI model, speech corpora recorded by five Korean male and five Korean female speakers not included in training the SD and SA models were used. For this, 6,422 (10 h) and 1,080 (1.7 h) utterances were used for training and development, respectively. The testing set in the SD and SA models was also used to evaluate the SI model.

To compose the acoustic feature vectors needed for auxiliary input information, the spectral and excitation parameters were extracted using a previously proposed parametric ITFTE vocoder [16]. In this way, 40-dimensional LP coefficients were extracted and converted to line spectral frequencies (LSFs) for the spectral parameters. In addition, 32-dimensional slowly evolving waveform (SEW) and 4-dimensional rapidly evolving waveform (REW) coefficients were extracted for the excitation parameters. The F0, gain, and v/uv information were also extracted. The frame and shift lengths were set to 20 ms and 5 ms, respectively.

In the WaveNet training step, all acoustic feature vectors were duplicated from a frame to the samples to match the length of the input speech signals. Before training, they were normalized to have zero mean and unit variance. The corresponding speech signal was normalized in a range between -1.0 and 1.0 and encoded by 8-bit-μ compression. The WaveNet architecture comprised three convolutional blocks, each with ten dilated convolution layers with dilations of 1, 2, 4, and so on up to 512. The number of channels of dilated causal convolution and the 1×1 convolution in the residual block were both set to 512. The number of 1×1 convolution channels between the skip-connection and the softmax layer was set to 256. The learning rate was set to 0.0001, and the batch size was set to 30,000 (1.25 sec).

To train the SI-WaveNet model, all data from the multiple number of different speakers were used; the sequence of each batch was randomized across all speakers before input to the training process. The weights were initialized using *Xavier* initialization and *Adam* optimization was used [17,18]. The training methods of the SD- and SA-WaveNets were similar but the initialization process was different in each case the SD model was initialized by Xavier initialization whereas the SA-WaveNet was initialized using the SI-WaveNet model whose weights were optimized toward the target speaker's database to represent speaker-specific characteristics.

To construct a baseline SPSS system, we employed a shared hidden layer (SHL) *acoustic model* [19]. In this approach, the hidden layers were trained to model the universal attributes among the different speakers, and the output regression layers were trained to represent speaker-specific characteristics [20]. We trained the SHL model with the SI data and fine-tuned the weights using the target dataset. The linguistic feature vectors that were input included 356-dimensional contextual information consisting of 330 binary features of categorical linguistic contexts and 26 features of numerical linguistic contexts. The corresponding output vectors consisted of all the acoustic parameters together with their time dynamics [21]. Before training, both input and output features were normalized to have zero mean and unit variance. The SHL consisted of three feedforward layers with 1,024 units and one long short-term memory layer with 512 memory blocks. The weights were trained using a *backpropagation through time* algorithm with Adam optimization [22]. The learning rate was set to 0.02 for the first 10 epochs, 0.01 for the next 10, and 0.005 for the remaining epochs.

In the synthesis step, first the means of all acoustic feature vectors were predicted by the SHL model, and a speech parameter generation algorithm was then applied[1] to generate smooth trajectories for the acoustic parameters [24]. To enhance spectral clarity, an LSF-sharpening filter was also applied to the spectral parameters [16]. To reconstruct the speech signal, the generated acoustic features were used to compose the input auxiliary features. By inputting these acoustic features, the WaveNet vocoder generated discrete symbols corresponding to the quantized speech signal, and its dynamic was recovered via μ-law expansion.

The setups for training the SI-, SD-, and SA-ExcitNets were the same as for the WaveNets although the ExcitNet was able to predict the distribution of the excitation signal, obtained by passing the speech signal through an LP analysis filter. Similar to the WaveNet vocoder, the ExcitNet vocoder generated the excitation sequence in the synthesis step. Ultimately, the speech signal was reconstructed through an LP synthesis filter.

Table 1. LSD (dB) and F0 RMSE (Hz) results in an A/S framework: the smallest errors are in bold.

System	WaveNet		ExcitNet	
	LSD	F0 RMSE	LSD	F0 RMSE
SD	3.65	38.27	2.10	19.83
SI	2.00	10.64	1.32	10.48
SA	**1.79**	**10.70**	**1.12**	**9.66**

Table 2. LSD (dB) and F0 RMSE (Hz) results of an SPSS task: the smallest errors are in bold.

System	WaveNet		ExcitNet	
	LSD	F0 RMSE	LSD	F0 RMSE
SD	4.78	48.75	4.50	39.14
SI	4.51	35.53	4.42	36.28
SA	**4.45**	**35.45**	**4.36**	**35.47**

4.2. Objective test results

4.2.1. Speech analysis and synthesis

To verify the performance of the vocoder, we first analyzed it in an analysis/synthesis (A/S) framework where the acoustic features extracted from the recorded speech were used to directly compose the auxiliary input features. In the test, distortions between the original speech and the synthesized speech were measured by log-spectral distance (LSD; dB) and F0 root mean square error (RMSE; Hz). The LSD and F0 RMSE test results shown in Table 1 confirm that the proposed SA training methods more accurately produce speech signals than the SD and SI methods in both the WaveNet and ExcitNet vocoders.

4.2.2. Statistical parametric speech synthesis

Distortions between the original and the generated speech were measured in an SPSS task. In this case, the acoustic features estimated by the SHL acoustic model were used to compose the input auxiliary features. Table 2 presents the test results with

[1]Because the acoustic model could not predict the variance for the speech parameter generation algorithm, we used the pre-computed global variances of the acoustic features from all training data [23].

respect to the different types of training methods. The findings can be outlined as follows: (1) The proposed SA training method still reconstructs much more accurate speech signals than the SD and SI models in both WaveNet and ExcitNet vocoders; (2) Among the different vocoding systems, ExcitNet performed significantly better than WaveNet in terms of spectral distortion since the adoption of an adaptive spectral filter in the ExcitNet vocoder helped reconstruct more accurate speech signals.

4.2.3. Speech modification

To further verify the effectiveness of the proposed SA training method, we investigated the performance variation of neural vocoders when F0 is manually modified. It has already been shown that the SI model is effective in generating pitch-modified synthesized speech [4]. As the SA approach in the present study was adapted from an SI model, it was expected to further improve performance as compared to conventional SD approaches.

In this experiment, the F0 trajectory was first generated by the SPSS framework and then multiplied by a scaling factor $(0.6, 0.8, 1.0,$ and $1.2)$ to modify the auxiliary feature vectors. Finally, the speech signal was synthesized using the neural vocoding systems. Fig. 2 illustrates the F0 RMSE (Hz) test results with respect to the different values of scaling factor. The results demonstrate that the proposed SA training model contained much smaller modification errors than the conventional SD approach. Compared with the SI method, the SA-ExcitNet in particular maintained equivalent qualities even though all weights had been optimized toward the target speaker's database. In addition, the results show that the ExcitNet performed significantly better than the WaveNet system. Because only the ExcitNet model was instructed to learn the variations of vocal cord movement, we may conclude that it was therefore able to reconstruct the F0-modified speech segments more flexibly than the WaveNet-based approach.

4.3. Subjective test results

To evaluate the perceptual quality of the proposed system, mean opinion score (MOS)[2] tests were performed[3]. In the tests, 12 native Korean listeners were asked to make quality judgments about the synthesized speech using the following five possible responses: 1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; and 5 = Excellent. In total, 20 utterances were randomly selected from the test set and were then synthesized using the different neural vocoders. To verify vocoding performance, speech samples synthesized by conventional vocoders such as ITFTE and WORLD (D4C edition [25]) were also included.

As presented in Fig. 3, the subjective test results confirm the effectiveness of each system in several ways. Within the A/S framework, the SD-WaveNet performed worst because it was difficult to learn the target speaker's characteristics with such limited training data. The SI-WaveNet performed similarly to the ITFTE vocoder but performed much better than the WORLD system. Across all WaveNet systems, the SA version achieved the best quality which confirms that adapting the multi-speaker model to the target speaker's database was beneficial to the vocoding performance. The ExcitNet results show similar

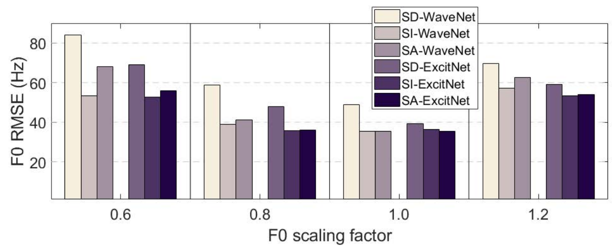

Fig. 2: F0 RMSE (Hz) results with respect to different values of the scaling factor.

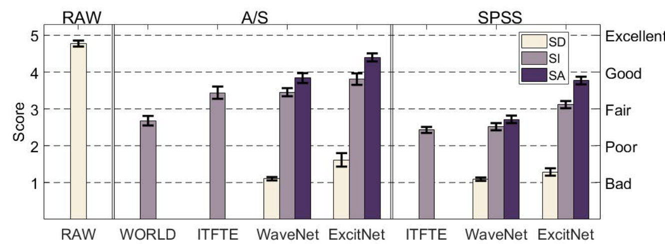

Fig. 3: MOS results with 95% confidence intervals for conventional and proposed systems. Acoustic features extracted from recorded speech and generated from an acoustic model were used to compose the input auxiliary features in the A/S and SPSS tests, respectively.

tendencies to those from the WaveNet systems but the ExcitNet performed much better overall, confirming that decoupling the formant component of the speech signal via an LP inverse filter improves the modeling accuracy of the residual signal. Consequently, the proposed SA-ExcitNet vocoder achieved a 4.40 MOS within the A/S framework.

In terms of SPSS, both the SI-WaveNet and SI-ExcitNet vocoders provided better perceptual quality than the parametric ITFTE system. The results confirm that the proposed SA training model significantly improved the quality of synthesized speech as compared with existing speaker-dependent and -independent methods. As in the A/S analysis, the ExcitNet performed better than the WaveNet in the SPSS task. Although the model generated overly smoothed speech parameters, ExcitNet was then able to alleviate the smoothing effect by directly estimating time domain excitation signals. Consequently, the SPSS system with the proposed SA-ExcitNet vocoder achieved a 3.77 MOS.

5. CONCLUSION

This paper proposed speaker-adaptive neural vocoders for statistical parametric speech synthesis systems (SPSS) where insufficient target speaker data exists. Using an initial speaker-independent trained model, the system first captured universal attributes from the speech of multiple speakers. This model was then fine-tuned with the target speaker's database to successfully represent speaker-specific characteristics using only ten minutes of training data. The experimental results verified that an SPSS system with the proposed speaker-adaptive neural vocoder performed significantly better than traditional versions with vocoders based on linear predictive coding and than systems with similarly configured neural vocoders trained both speaker-dependently and speaker-independently. Future research will include integrating the entire framework into speech synthesis systems that use an end-to-end approach.

[2]Generated audio samples are available at the following url: https://soundcloud.com/eunwoo-song-532743299/sets/speaker-adaptive-neural-vocoders

[3]The tests were performed in an acoustically isolated room using a Sennheiser HD650 headphone.

6. REFERENCES

[1] A. Van Den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with PixelCNN decoders," in *Proc. NIPS*, 2016, pp. 4790–4798.

[2] 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," *CoRR abs/1609.03499*, 2016.

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

[4] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Proc. ASRU*, 2017, pp. 712–718.

[5] Y.-J. Hu, C. Ding, L.-J. Liu, Z.-H. Ling, and L.-R. Dai, "The USTC system for blizzard challenge 2017," in *Proc. Blizzard Challenge Workshop*, 2017.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[7] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a statistical vocoder," in *Proc. ICASSP*, 2018, pp. 5674–5678.

[8] E. Song, K. Byun, and H.-G. Kang, "A neural excitation vocoder for statistical parametric speech synthesis systems," *Submitted to Pattern Recogn. Lett.*, 2018.

[9] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 26, no. 7, pp. 1173–1180, 2018.

[10] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," 2018, pp. 2012–2016.

[11] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of Acoust. Soc. of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

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

[13] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted WaveNet vocoder," *Proc. INTERSPEECH*, pp. 1978–1982, 2018.

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

[15] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. ICASSP*, 2018, pp. 4804–4808.

[16] E. Song, F. K. Soong, and H.-G. Kang, "Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 11, pp. 2152–2161, 2017.

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[19] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. ICASSP*, 2015, pp. 4475–4479.

[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 *Proc. ICASSP*, 2013, pp. 7304–7308.

[21] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[22] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural computat.*, vol. 2, no. 4, pp. 490–501, 1990.

[23] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[24] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[25] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech commun.*, vol. 84, pp. 57–65, 2016.