INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis

Shinji Takaki[1], *Hirokazu Kameoka*[2], *Junichi Yamagishi*[1,3]

[1]National Institute of Informatics, Japan.
[2]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan.
[3]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom.
takaki@nii.ac.jp, hirokazu.kameoka@lab.ntt.co.jp, jyamagis@nii.ac.jp

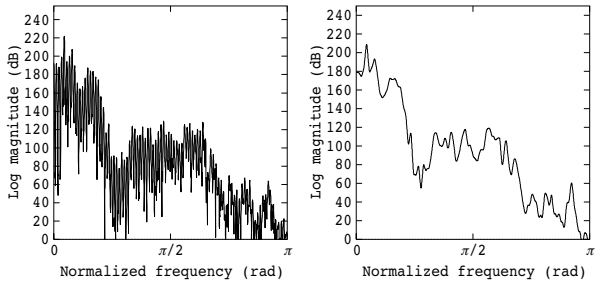
Abstract

In statistical parametric speech synthesis (SPSS) systems using the high-quality vocoder, acoustic features such as mel-cepstrum coefficients and F0 are predicted from linguistic features in order to utilize the vocoder to generate speech waveforms. However, the generated speech waveform generally suffers from quality deterioration such as buzziness caused by utilizing the vocoder. Although several attempts such as improving an excitation model have been investigated to alleviate the problem, it is difficult to completely avoid it if the SPSS system is based on the vocoder. To overcome this problem, there have recently been attempts to directly model waveform samples. Superior performance has been demonstrated, but computation time and latency are still issues. With the aim to construct another type of DNN-based speech synthesizer with neither the vocoder nor computational explosion, we investigated direct modeling of frequency spectra and waveform generation based on phase recovery. In this framework, STFT spectral amplitudes that include harmonic information derived from F0 are directly predicted through a DNN-based acoustic model and we use Griffin and Lim's approach to recover phase and generate waveforms. The experimental results showed that the proposed system synthesized speech without buzziness and outperformed one generated from a conventional system using the vocoder.
Index Terms: Statistical parametric speech synthesis, DNN, FFT spectrum, Phase reconstruction, Vocoder

1. Introduction

Research on statistical parametric speech synthesis (SPSS) has been advancing recently due to deep neural networks (DNNs) with many hidden layers. For systems where waveform signals are generated using a high-quality vocoder such as STRAIGHT [1], WORLD [2, 3], or sinusoidal models, DNNs, recurrent neural networks, long-short term memories, etc. have been used to learn the relationship between input texts and vocoder parameters [4, 5, 6, 7]. Recently, generative adversarial networks [8] have also been used as a post-process module to refine the outputs of the speech synthesizers, and the resulting synthetic speech has become statistically less significant compared to analysis-by-synthesis samples [9]. In addition, there have been new attempts for directly modeling waveform signals using neural networks such as WaveNet [10] and SampleRNN [11].

In this work, we investigate the direct modeling of *frequency spectra* that contains both spectral envelopes and harmonic structures together obtained by a simple deterministic frequency transform such as ordinary short-term Fourier transform (STFT). Figure 1 shows examples of a STFT spectral amplitude and spectral envelope obtained by a simple frequency transform and WORLD spectral analysis, respectively. Compared to our previous work, where we concentrated on the extraction of low-dimensional features from the frequency spectra by using a deep auto-encoder [12], the focus of the present work is about waveform generation using the frequency spectra predicted by DNNs (but without using a vocoder).



(a) STFT spectral amplitude (b) WORLD spectral envelope

Figure 1: *A STFT spectral amplitude and WORLD spectral envelope obtained via a simple frequency transformation or WORLD spectral analysis.*

The advantages of the proposed waveform generation are that a) the representation is much "closer" to original waveform signals compared to vocoder parameters and b) DNNs need to be used per frame, whereas for direct waveform modeling, DNNs need to be used per waveform sample. To enable the proposed waveform generation, it is necessary to build DNNs that can accurately predict high-dimensional frequency spectra including harmonic structures. Note that the dimension of frequency spectra is typically much higher than the vocoder parameters. We also need to recover phase information if we model amplitude spectra only.

For constructing such a high-quality acoustic model for STFT spectral amplitudes, we investigate 1) the use of F0 information as well as linguistic features as the input, 2) an objective criterion based on Kullback-Leibler divergence (KLD), and 3) peak enhancement of predicted STFT spectral amplitudes. For the phase recovery and waveform generation, we use a well-known conventional phase reconstruction algorithm proposed by Griffin and Lim [13]. We compared synthetic speech based on the proposed waveform generation with ones based on the vocoder.

The rest of this paper is organized as follows. Section 2 of this paper presents a DNN-based acoustic model and objective criteria to train it. Section 3 describes the procedure of waveform generation used in the proposed systems. The experimental results are presented in Section 4. We conclude in Section 5 with a brief summary and mention of future work.

http://dx.doi.org/10.21437/Interspeech.2017-488

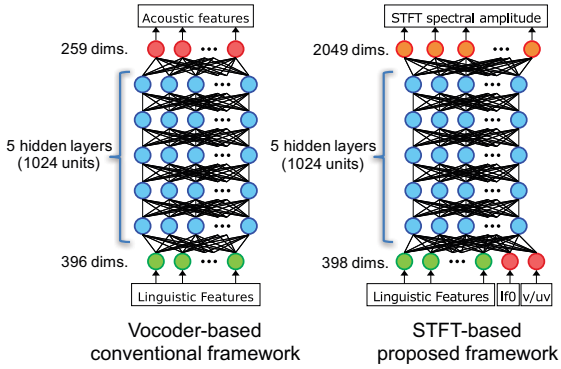


Figure 2: *DNN architectures for the proposed waveform generation*

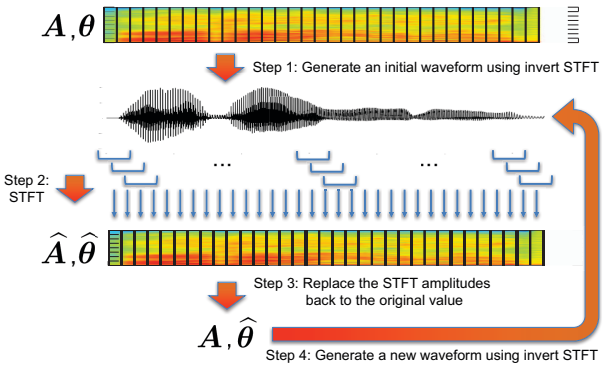


Figure 3: *Flow chart of phase reconstruction algorithm. Here, $\boldsymbol{A}$, $\widehat{\boldsymbol{A}}$, $\boldsymbol{\theta}$, and $\widehat{\boldsymbol{\theta}}$ represent predicted and new spectral amplitudes, initial and new phase values, respectively.*

2. Direct modellig of frequency spectra

2.1. Architecture

The left part in Fig. 2 shows the framework of the conventional DNN-based acoustic model used for the vocoder. The DNN-based acoustic models are normally used to represent the relationship between linguistic and vocoder features [4, 14, 5, 15].

The right part in the figure shows a new DNN architecture to be used for the proposed waveform generation. High-dimensional STFT spectral amplitudes are the outputs and we explicitly use F0 information, i.e., log F0 and voiced/unvoiced values, in addition to linguistic features as the inputs. We expect that spectral envelopes can be predicted by linguistic features and that harmonic structures can be predicted by the F0 information.

2.2. KLD based training

In general, least square error (LSE) is used as an objective criterion to train a DNN-based acoustic model. An objective criterion using LSE is defined as

$$\hat{\boldsymbol{\lambda}}_{LSE} = \arg\min_{\boldsymbol{\lambda}} \frac{1}{2} \sum_{t=1}^{T} \sum_{d=1}^{D} (o_{t,d} - y_{t,d})^2, \tag{1}$$

where $o_{t,d}$, $\boldsymbol{l}_t$, t, d, and $\boldsymbol{\lambda}$ represent an observation (i.e., an acoustic feature), an input (i.e., a linguistic feature and F0 information), a frame index, a dimension and the model parameters of a DNN, respectively. Also, $y_{t,d} = g_d^{(\boldsymbol{\lambda})}(\boldsymbol{l}_t)$ and a function $g^{(\boldsymbol{\lambda})}(\cdot)$ is non-linear transformation represented by a DNN.

In contrast to [4], in this paper we use the high-dimensional STFT spectral amplitudes directly as the output references to train a DNN-based acoustic model. To utilize the benefit of direct use of the STFT spectral amplitudes and construct a more appropriate model, we define an objective criterion based on Kullback-Leibler divergence (KLD), which has been successfully used for source separation with non-negative matrix factorization [16, 17], as

$$\hat{\boldsymbol{\lambda}}_{KL} = \arg\min_{\boldsymbol{\lambda}} \sum_{t=1}^{T} \sum_{d=1}^{D} o_{t,d} \log \frac{o_{t,d}}{\tilde{y}_{t,d}} - o_{t,d} + \tilde{y}_{t,d}, \tag{2}$$

$$\tilde{y}_{t,d} = s_d y_{t,d} + b_d, \tag{3}$$

where s_d and b_d represent fixed values calculated from training data for performing unnormalization. For using a KLD-based objective criterion, observations and $\tilde{y}_{t,d}$ have to be positive. However, there is no guarantee about output range if the DNN directly outputs $\tilde{y}_{t,d}$ using a linear output layer. To avoid this problem, we adopted the sigmoid function for an output layer to predict normalized values ranged from 0 to 1 so that an objective criterion is defined on the basis of KLD.

By using pairs of input and output features obtained from the training dataset, the parameters of a DNN can be effectively trained by using SGD [18] with derivative w.r.t. $y_{t,d}$ as

$$\frac{\partial E_{LSE}}{\partial y_{t,d}} = y_{t,d} - o_{t,d}, \tag{4}$$

$$\frac{\partial E_{KL}}{\partial y_{t,d}} = s_d \left(1 - \frac{o_{t,d}}{s_d y_{t,d} + b_d} \right). \tag{5}$$

2.3. Post-filter of predicted STFT spectral amplitudes

Although the accuracy of the STFT spectral amplitudes predicted by the DNNs is good, we saw that refinement of the amplitudes gains the final performance. We therefore apply a signal processing-based post-filter (PF) [19] for enhancing the spectral peaks of predicted STFT spectral amplitudes. The process is as follows: 1) predicted STFT spectral amplitudes are converted into linear-scale cepstrum vectors that have the same dimensions as the STFT amplitudes, 2) the post-filter is applied to the cepstrum vectors for the peak enhancement, and 3) the cepstrum vectors after post-filtering are converted back into the spectral amplitudes.

3. Waveform generation based on phase recovery

This section describes the speech waveform generation algorithm based on phase recovery. In this work, we adapted the well-known phase reconstruction algorithm proposed by Griffin and Lim [13], the flow chart of which is shown in Fig. 3. The algorithm consists of the following iterative steps.

1. Generate initial speech waveforms using inverse STFT of predicted STFT spectral amplitudes $\boldsymbol{A}$ with or without postfilter and random phase $\boldsymbol{\theta}$ at each frame, followed by overlap-add operations.
2. Window the speech waveforms and apply STFT at each frame to obtain new spectral amplitude $\widehat{\boldsymbol{A}}$ and phase values $\widehat{\boldsymbol{\theta}}$.
3. Replace the STFT spectral amplitudes $\widehat{\boldsymbol{A}}$ with the original values $\boldsymbol{A}$ at each frame.

Table 1: *Inputs, output references, and objective criteria for training each acoustic model are listed in this table. Here, v/uv and bap represent voiced/unvoiced values and band aperiodicity measures, respectively.*

Model name	Input	Output	Criterion	Post-filter	Waveform generation
Baseline	linguistic features	mel-cep. log F0, v/uv, bap	LSE		vocoder
Baseline+PF	linguistic features	mel-cep. log F0, v/uv, bap	LSE	$\surd$	vocoder
LSE	linguistic features	SFTF spectral amplitude	LSE		phase recovery
KLD	linguistic features	STFT spectral amplitude	KLD		phase recovery
LSE+F0	linguistic features, log F0, v/uv	STFT spectral amplitude	LSE		phase recovery
KLD+F0	linguistic features, log F0, v/uv	STFT spectral amplitude	KLD		phase recovery
LSE+F0+PF	linguistic features, log F0, v/uv	STFT spectral amplitude	LSE	$\surd$	phase recovery
KLD+F0+PF	linguistic features, log F0, v/uv	STFT spectral amplitude	KLD	$\surd$	phase recovery

4. Generate a new speech waveform using inverse STFT of original STFT spectral amplitudes A and updated phases $\widehat{\boldsymbol{\theta}}$, followed by overlap-add operations.

5. Go back to step 2 until convergence.

4. Experiments

4.1. Experimental conditions

We used the database that was provided for the Blizzard Challenge 2011 [20], which contains approximately 17 hours of speech data comprising 12K utterances. All data were sampled at 48 kHz. Two hundred sentences that are not included in the database were used as a test set.

We constructed two baseline and six proposed systems listed in Table 1. In addition to investigate the effectiveness of the objective criterion based on KLD, post-filter, and waveform generation, we also look into the effectiveness of using F_0. For the baseline system, the WORLD analysis was used for obtaining spectral envelopes that were then converted into mel-cepstrum coefficients. The WORLD vocoder was used to generate a waveform from the predicted acoustic features.

For the proposed systems, STFT spectral amplitudes were used as the output references. The KLD-based objective criterion was used for training systems notated KLD, KLD+F0, and KLD+F0+PF, while the LSE-based objective criterion was used for other proposed systems. The F0 information was added as the input for training systems LSE+F0, LSE+F0+PF, KLD+F0, and KLD+F0+PF, while only the linguistic features were used as the input for the other proposed system. We have applied PF for systems LSE+F0+PF and KLD+F0+PF and used the results of the post-filter as the initial values of phase recovery. For other proposed systems, we used the outputs of DNNs as the initial values of phase recovery. For the baseline system, we used the conventional cepstral-based post-filter [19] to ensure fair comparison. For each waveform, we extracted its frequency spectra with 2049 STFT points. The feature vectors for the baseline system comprised 259 dimensions: 59 dimensional bark-cepstral coefficients (plus the 0th coefficient), log F0, 25 dimensional band aperiodicity measures, their dynamic and acceleration coefficients, and voice/unvoiced values. The context-dependent labels were built using the pronunciation lexicon Combilex [21]. The linguistic features for DNN acoustic models comprised 396 dimensions. Five hidden layer feed-forward neural networks with a sigmoid-based activation function were used for acoustic models. In the synthesis phase, we used log F0 and voiced/unvoiced values predicted by using the baseline system as the inputs of the LSE+F0, LSE+F0+PD, KLD+F0, and KLD+F0+PF.

For subjective evaluation, MUSHRA tests were conducted to evaluate the naturalness of synthesized speech. Natural speech was used as a hidden top anchor reference. Fourteen native subjects participated in the experiments. Twenty sentences were randomly selected from the test set for each participant. The experiments were carried out using headphones in quiet rooms.

4.2. Experimental results

4.2.1. Synthetic spectrogram

Fig. 4 shows the low-frequency parts (8 kHz) of synthetic spectral amplitude in each system. First, we can see from the figures that harmonic information was clearly predicted when F0 information was explicitly used for inputs of the DNN-based acoustic models (LSE+F0 and KLD+F0). Systems based on LSE and KLD, in which F0 information was not used as inputs, could not sufficiently predict harmonic information, though parts of the harmonics were faintly generated compared them with ones generated by the baseline system.

Second, when we compare the synthetic spectral amplitudes obtained by training with objective criteria based on LSE and KLD, we can see that peaks of harmonics parts were enhanced by using the criterion based on KLD. This demonstrates that an objective criterion based on KLD is more appropriate to model the STFT spectral amplitude including harmonic information.

Finally, we can see from the figures that using the post-filter (PF) further enhanced the peaks of the harmonic information. These results indicates that using F0 information as inputs, an objective criterion based on KLD for training, and the post-filter would be effective for generating STFT spectra including harmonic information.

4.2.2. Subjective results

Figure 5 shows the subjective results with 95% confidence intervals in the experiments. The result for natural speech is excluded from the figures to make the comparison easier. For the subjective tests, we additionally trained an acoustic model called KLD+F0+PF (32 kHz) with down sampled data (32 kHz) using the same strategy as KLD+F0+PF. This is because the original speech quality between audios sampled at 32 kHz and 48 kHz are comparable, but the number of STFT points can be reduced and training a DNN would then become easier. At synthesis time, this means computationally efficient and low latency. Therefore, STFT spectra with 1025 points were used for KLD+F0+PF(32kHz). The speech generation speed of this system was 5 times faster than that using 48kHz. We used six systems constructed on the basis of Baseline, Baseline+PF, LSE+F0, KLD+F0, KLD+F0 +PF, and KLD+F0+PF (32 kHz) for the listening test.

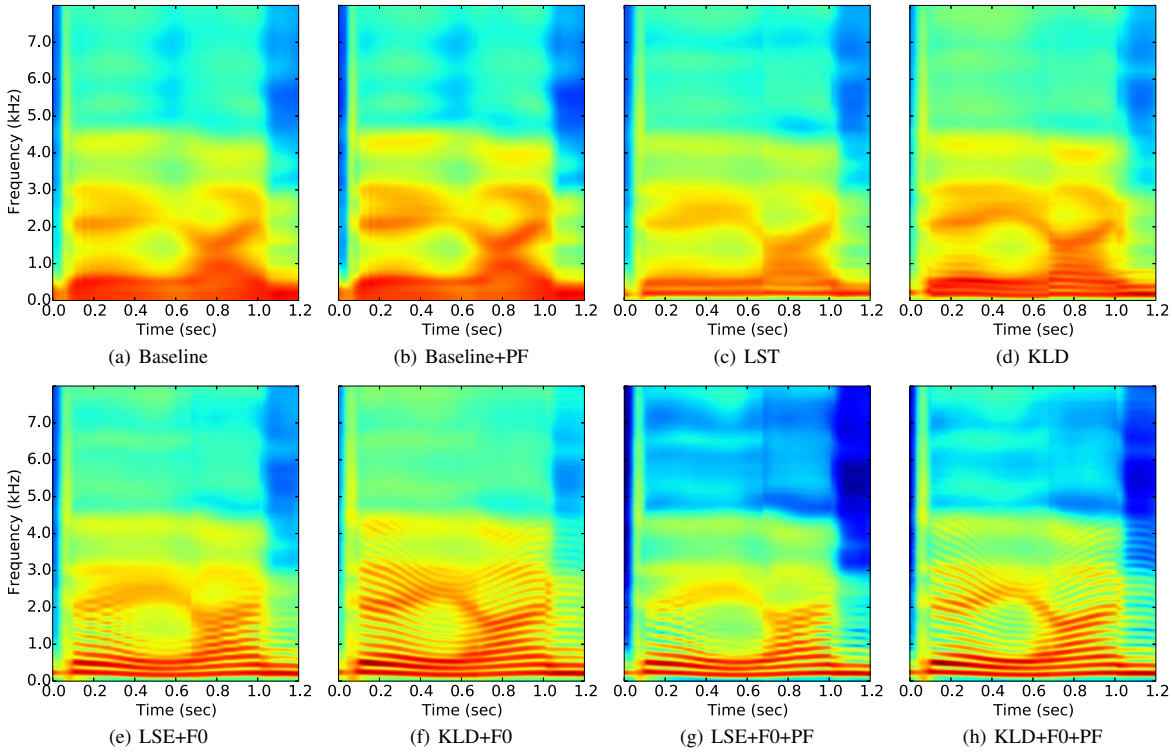


Figure 4: *Low-frequency parts (8 kHz) of synthetic spectral amplitudes in each system. PF means the post-filter.*

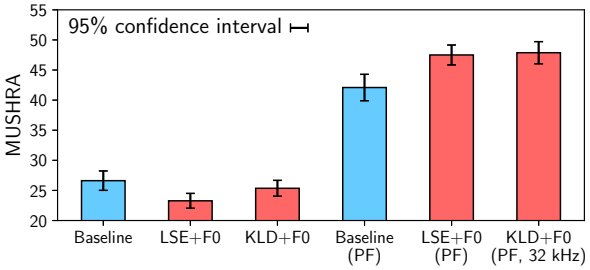


Figure 5: *Subjective results.*

First, among the systems without the post-filter, we can see from the figure that the system using the KLD-based objective criterion (KLD+F0) statistically outperformed the one using the LSE-based objective criterion (LSE+F0). This indicates that the KLD based objective criterion was more appropriate to use for modeling the STFT spectral amplitudes than using the LSE based objective criterion. However, these systems using the STFT spectral amplitudes without the post-filter (LSE+F0, KLD+F0) outputs worse quality of synthetic speech than ones synthesized by the baseline system based on the WORLD vocoder.

Second, we can see from the figure about the proposed systems with and without the post-filter that the quality of speech synthesized by the systems with the post-filter, i.e., KLD+F0+PF and KLD+F0+PF(32kHz), were significantly improved from one synthesized by the systems without the post-filter (KLD+F0). The proposed system with the post-filter outputs synthetic speech with less noise caused by reconstructing inappropriate phase compared to those generated from the systems without the post-filter. This means that enhancing the STFT spectral amplitudes using the post-filter was effectively utilized to perform phase recovery and waveform generation. The computationally efficient system using audio sampled at 32 kHz was as good as the one using audio sampled at 48 kHz because the difference between these two systems was not statistically significant.

Finally, it can be seen from the figure that the proposed systems with the post-filter, i.e., KLD+F0+PF and KLD+F0+PF(32 kHz), outperforms the baseline system based on the post-filter, i.e., Baseline+PF.

5. Conclusion

We presented our investigation of direct modeling of frequency spectra and waveform generation based on phase recovery towards constructing another type of DNN-based speech synthesis system without a vocoder. Experimental results demonstrated that explicit use of F0 information as the input of a DNN-based acoustic model and an objective criterion defined using KLD were effective to model STFT spectral amplitudes that include harmonic information. Also, the results of a subjective listening test showed that although the prediction accuracy of STFT spectral amplitudes from the DNN-based acoustic model was still insufficient, the post-filter could enhance the spectral peaks, and the proposed systems with the post-filter outperformed the conventional DNN-based synthesizer using a vocoder with the post-filter.

We have also attempted to replace the signal processing post-filter with a generative adversarial nets (GAN)-based model [8] for further improvement, which will be reported in our another paper [22].

6. Acknowledgements

This work was partially supported by ACT-I from the Japan Science and Technology Agency (JST), by MEXT KAKENHI Grant Numbers (26280066, 15H01686, 16K16096, 16H06302), and by The Telecommunications Advancement Foundation Grants.

7. References

[1] H. Kawahara, I. Masuda-Katsuse, and A. 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, pp. 187–207, 1999.

[2] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," *the Stockholm Music Acoustics Conference 2013 (SMAC2013)*, pp. 289–292, 2015.

[3] ——, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,," *Speech Communication*, vol. 67, pp. 1–7, 2015.

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

[5] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," *Proceedings of Interspeech*, pp. 1964–1968, 2014.

[6] Q. Hu, Z. Wu, K. Richmond, J. Yamagishi, Y. Stylianou, and R. Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," *Proceedings of Interspeech*, pp. 854–858, 2015.

[7] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," *Proceedings of ICASSP*, pp. 5120–5124, 2016.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proceedings of NIPS*, pp. 2672–2680, 2014.

[9] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," *Proceedings of ICASSP*, pp. 4910–4914, 2017.

[10] 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," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[11] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *CoRR*, vol. abs/1612.07837, 2016. [Online]. Available: http://arxiv.org/abs/1612.07837

[12] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," *Proceedings of ICASSP*, pp. 5535–5539, 2016.

[13] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, pp. 236–243, 1984.

[14] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 2129–2139, 2013.

[15] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," *Proceedings of Interspeech*, pp. 2268–2272, 2014.

[16] H. S. S. D. D. Lee, "Algorithms for nonnegative matrix factorization," *Proceedings of Adv. Neural Inform. Process. Syst.*, pp. 556–562, 2001.

[17] B. R. P. Smaragdis and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proceedings of 7thInt. Conf. Ind. Compon. Anal. Signal Separat.*, pp. 414–421, 2007.

[18] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science 28*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP coding based on mel-cepstral analysis," *Proceedings of ICASSP*, pp. 33–33, 1995.

[20] S. King and V. Karaiskos, "The Blizzard Challenge 2011," *Blizzard Challenge 2011 Workshp*, 2011. [Online]. Available: http://festvox.org/blizzard/bc2011/summary_Blizzard2011.pdf

[21] K. Richmond, R. Clark, and S. Fitt, "On generating Combilex pronunciations via morphological analysis," *Proceedings of Interspeech*, pp. 1974–1977, 2010.

[22] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," *Proceedings of Interspeech*, 2017.