



# Text-to-speech synthesis using spectral modeling based on non-negative autoencoder

Takeru Gorai, Daisuke Saito, Nobuaki Minematsu

The University of Tokyo, Japan

{gorai, dsk.saito, mine}@gavo.t.u-tokyo.ac.jp

## Abstract

This paper proposes a statistical parametric speech synthesis system that uses non-negative autoencoder (NAE) for spectral modeling. NAE is a model that extends non-negative matrix factorization (NMF) as neural networks. In the proposed method, we employ latent variables in NAE as acoustic features. Reconstruction of spectral information and estimation of latent variables are simultaneously trained. The non-negativity of latent variables in NAE is expected to contribute to dimensionality reduction such that the fine structure of the spectral envelopes is preserved. Experimental results demonstrate the effectiveness of the proposed framework. We also study multi-speaker modeling where each of NAEs corresponds to each single speaker. In addition, a neural source-filter (NSF) model was applied to the waveform generation. When a neural vocoder is trained with natural acoustic features and tested with synthesized features, quality degradation occurs due to the mismatch between training and test data. In order to mitigate the mismatch, this system uses features obtained by reconstructing natural speech using NAE for training. Experimental results show that reconstructed features are similar to synthesized features, and as a result, the quality of the synthesized speech is improved.

**Index Terms:** text-to-speech synthesis, non-negative autoencoder, neural source-filter

## 1. Introduction

In the area of text-to-speech synthesis (TTS), end-to-end speech synthesis models have achieved a high degree of naturalness in recent years [1, 2]. On the other hand, statistical parametric speech synthesis (SPSS) is a speech synthesis system divided into several modules. In SPSS, vocoder parameters are estimated from linguistic features and speech waveforms are generated using vocoder. When sufficient amount of training data is available, the quality of synthesized speech of SPSS is relatively lower than that of end-to-end approaches. However it has advantages such as the ease of changing speaker characteristics and the ability to work with relatively small amounts of data. From an acoustic modeling perspective, we discuss two factors that cause quality degradation.

First, there is a problem of degradation due to dimensionality reduction of spectral envelopes. Since spectral envelope is a high-dimensional and redundant feature, it is known that it can be represented more efficiently by compressing it into appropriate low-dimensional features in the first stage [3]. The most common low-dimensional feature that can be obtained data-independently is Mel-cepstrum coefficients [4]. Although Mel-cepstrum is known to efficiently represent spectral envelopes with low-dimensional parameters, it has the disadvantage of excessively smoothing the fine structure of spectral envelopes. Several statistical methods have been proposed to refine spec-

tral modeling. One of them is the deep autoencoder (DAE) method [3]. By using latent variables of DAE as acoustic features, this method successfully achieves dimensionality reduction more suitable for speech synthesis than the method based on Mel-cepstrum. A method using non-negative matrix factorization (NMF) [5] has also been proposed. This method uses NMF to decompose spectral envelopes into the product of a basis matrix and activation weights. The basis matrix and the activation weights correspond to a set of spectral templates and their usage weights, respectively. By using the activation weights as acoustic features, it is possible to perform synthesis without losing the fine structure of spectral envelopes.

Second, there is a quality degradation due to vocoders. Vocoders based on signal processing use various approximations, such as constant analysis window lengths and time-invariant linear filters, which cause degradation during speech analysis and synthesis. One of the breakthroughs for the problem is to use a neural vocoder such as WaveNet vocoder [6, 7]. However, it has been reported that a neural vocoder does not perform well due to the mismatch between the training and test acoustic features [8].

In order to approach these two factors that cause quality degradation of SPSS, in this paper, we propose a novel TTS system using spectral modeling by a non-negative autoencoder (NAE) [9]. NAE is a model that extends NMF as a neural network and it can be incorporated into a DNN, which has the advantage of simultaneously extracting low-dimensional features and training a DNN-based acoustic model. In this paper, NAE-based acoustic models for both the cases of single speaker and multi speakers are investigated. In addition, in particular to tackle the problem of the mismatch, this paper describes a model that uses neural source-filter model (NSF), with helps of NAE-based reconstruction of acoustic features.

## 2. Proposed method

### 2.1. Overview

The proposed TTS system is a cascade of three modules; front-end processing, estimation of acoustic features based on NAE and NSF-based waveform generation. Section 2.2 describes estimation of acoustic features using NAE-based spectral modeling. Section 2.3 describes NSF-based waveform generation.

### 2.2. NAE-based spectral modeling

#### 2.2.1. Non-negative autoencoder

Non-negative matrix factorization (NMF) [5] is a method to approximate a nonnegative matrix  $Y \in \mathbb{R}^{\geq 0, K \times N}$  by multiplication of two non-negative matrices  $H \in \mathbb{R}^{\geq 0, K \times M}$  and  $U \in \mathbb{R}^{\geq 0, M \times N}$ .

$$Y \simeq HU \quad (1)$$

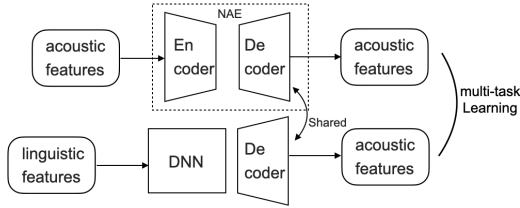


Figure 1: Overview of multi-task learning of NAE and DNN-based acoustic model.

where  $H$  is called *dictionary* and  $U$  is called *activation*. When  $M \ll K$  and  $M \ll N$ , NMF can be interpreted as low-rank approximation, and it is known that non-negative constraints make activation sparse. When NMF is applied to amplitude spectrograms of speech signals, it can be interpreted as a decomposition into a product of spectral templates and their sparse weights. Dictionary is strongly dependent on the nature of the dataset, such as speaker information. Activation is a low-dimensional feature that depends strongly on linguistic information and not so much on speaker characteristics.

Non-negative autoencoder (NAE) [9] is an extension of NMF to neural networks (NN). In NMF, decomposition into two non-negative matrices and reconstruction can be interpreted as a linear autoencoder and rewritten as follows

$$\text{Encoder} : U = H^\dagger Y, \quad \text{Decoder} : Y = HU, \quad (2)$$

where  $H^\dagger$  is the pseudo-inverse of  $H$ .  $H^\dagger$  and  $H$  are the encoder and decoder, respectively, and  $U$  represents the latent variables of autoencoder.

In the NAE model, non-negativity of activation is guaranteed by adoption of non-negative activation functions. On the other hand, for the encoder and decoder, their constraints for non-negativity are removed. Encoding and decoding processes are as follows;

$$\text{Encoder} : \mathbf{z} = g(W_1 \mathbf{y}), \quad \text{Decoder} : \hat{\mathbf{y}} = g(W_2 \mathbf{z}), \quad (3)$$

where  $W_1$  and  $W_2$  are weight matrices and  $g$  is an activation function that guarantees non-negativity of the output, such as softplus. In [9], Kullback-Leibler Divergence (KLD) is used to calculate the reconstruction error.

### 2.2.2. TTS based on NAE

In this paper, we propose a statistical speech synthesis method using NAE for spectrogram dimensionality reduction. As shown in Figure 1, multi-task learning of NAE reconstruction and estimation of acoustic features from linguistic features is performed.

The input and output features of NAE are spectral envelopes normalized to have an L1-norm of 1. Latent variables estimated by the DNN-based acoustic model are input to the decoder, and the TTS error is calculated based on its output. The overall error function is given by Equation (4);

$$D_{KL}(\mathbf{y}, d(\mathbf{z}_{enc})) + D_{KL}(\mathbf{y}, d(\mathbf{z}_{tts})) + D_{KL}(p, \hat{p}), \quad (4)$$

where  $D_{KL}$  is KLD,  $\mathbf{y}$  is the normalized amplitude spectral envelope,  $\mathbf{z}_{enc}$  is the output of the encoder when  $\mathbf{y}$  is input to NAE,  $\mathbf{z}_{tts}$  is the latent vector estimated by the DNN-based acoustic model,  $d$  is the decoder network,  $p$  is the L1-norm of the original spectral envelope.

It is expected that by training NAE and DNN-based acoustic models simultaneously, dimensionality reduction of spectral envelopes more suitable for TTS tasks is performed.

### 2.2.3. multi-speaker modeling based on NAE

In this section, we introduce multi-speaker modeling based on the proposed method described in the previous section. Due to the non-negative constraint of NAE, latent variables are considered less dependent on speaker information. Therefore, we investigate a framework in which the DNN-based acoustic model estimates speaker-independent acoustic features and performs speaker adaptation using NAEs learned for each speaker as in [10]. Thanks to the non-negative constraint of NAE, it is expected that speaker information and speaker-independent latent representations are efficiently separated.

In addition, it is possible to add speaker representations such as speaker codes [11] to the input of the DNN-based acoustic model. Adding speaker representations is expected to compensate for the cases where expressive capability of NAE is not sufficient for speaker adaptation.

## 2.3. Waveform generation

### 2.3.1. Neural source-filter model

Neural Source Filter (NSF) [12, 13] is a neural waveform model based on the source-filter model. It consists of a source module that models the excitation source, a filter module that consists of dilated convolution networks, and a condition module that preprocesses input features. The loss is calculated based on multi-resolution amplitude spectrogram. Because neither autoregression nor knowledge distillation is used, training and inference are relatively fast. It also has the advantage of easy control of prosody.

In the original proposal of the NSF model, audio samples whose sampling frequency is 16 kHz were used. On the other hand, in this study, we have struggled waveform generation of audio samples whose sampling frequency is 48 kHz. In this expansion, naive NSF models have not worked well very much. Some unnatural noises were observed in the generated samples. To mitigate the problem, we finally adopted the helps of conventional vocoder. In our method, after wave generation by NSF, WORLD [14] (D4C edition[15]) resynthesis are applied. We have investigated the proposed cascade experimentally.

### 2.3.2. Mismatch reduction in NSF

Generally speaking, neural waveform generators such as NSF are independently trained of acoustic models for TTS. Hence, when the neural vocoders are applied to TTS, there tends to be a mismatch between actual and expected inputs for the vocoders. It is assumed that quality degradation occurs because acoustic features extracted from natural speech are conditioned during training while synthesized features are conditioned during inference.

A simple solution is to use acoustic features estimated from texts instead of natural speech when the vocoder is trained. However, this approach is less effective than expected because mismatches exist in the linguistic information and temporal structure of natural and synthesized features. Actually, preliminary experiments confirmed that this approach showed little improvement.

Therefore, in this paper, we propose a framework in which spectral envelopes reconstructed by NAE are used for NSF training. Through the encoding-decoding process of NAE, the acoustic properties are expected to be similar to synthesized features while preserving linguistic information and temporal structures. In addition, to further reduce the mismatch, we also studied a method for training the NAE decoder to reconstruct the spectral envelope estimated by the acoustic model rather

than that extracted from natural speech. Finally, the loss function is as follows;

$$D_{KL}(d(\mathbf{z}_{tts}), d(\mathbf{z}_{enc})) + D_{KL}(\mathbf{y}, d(\mathbf{z}_{tts})) + D_{KL}(p, \hat{p}). \quad (5)$$

We denote the learning method described in section 2.2.2 “DECNAT”, and the method described here “DECTTS”.

## 2.4. Related works

Statistical parametric speech synthesis using NMF has been proposed [16]. In this method, NMF is first applied to spectral envelopes to extract activations. Next, a DNN-based acoustic model is used to learn mapping from linguistic features to activations. Since spectral envelopes are decomposed into the product of templates and sparse weights of them, dimensionality reduction is performed such that fine structure of spectral envelopes is preserved. In addition, since dictionary can be interpreted as speaker-independent and activations as containing speaker-independent linguistic information, a multi-speaker modeling is proposed, in which speaker adaptation is performed by switching the dictionary for each speaker [17].

As a previous study that addresses the reduction of mismatches in neural vocoders, in voice conversion based on variational autoencoder (VAE), WaveNet vocoder is fine-tuned using features reconstructed by VAE in [18].

The proposed method differs from NMF-based TTS in that we use NAE and generation of dictionary is incorporated into training of DNN. In addition, while NMF-based multi-speaker modeling requires parallel speech data to generate dictionary for each speaker, the NAE-based model can be trained with nonparallel data. In addition, NAE is also utilized for solution of the mismatch problem in wave generation in the proposed system.

## 3. Experiments

### 3.1. Overview

To evaluate the proposed system, three experiments were carried out; (a) NAE-based SPSS for *single speaker*, (b) NAE-based SPSS for *multi-speaker*, and (c) *NSF wave generation* with NAE-based acoustic modeling. In experiment (a), the effectiveness of the method described in section 2.2.2 was evaluated and compared with a baseline method based on DAE [3]. The experiment was performed on a single speaker dataset. In experiment (b), a multi-speaker model based on NAE was constructed and compared with a speaker-dependent model. In experiment (c), a multi-speaker model followed by NSF was evaluated, and the effect of NAE reconstruction on mismatch reduction was verified.

### 3.2. Experimental conditions

In experiment (a), 450 sentences (0.5 hours) of the ATR Japanese sentence database [19] were used for training and 53 for evaluation. Speech samples were taken from the male speech data in HTS-demo [20]. In experiment (b), 100 speakers (42 males and 58 females) from the VCTK corpus [21] and in experiment (c), 10 speakers (3 males and 7 females) from the VCTK corpus were used for both the acoustic model and NSF training. For each speaker, 300 utterances were used for training and 20 for evaluation. The sampling frequency for both the ATR and VCTK datasets was 48kHz.

The speech analysis was performed using WORLD and the dimensionality of spectral envelopes was set to 1025. Harvest [22] was used for  $F_0$  extraction. Since this study focuses on the estimation of spectral envelopes,  $F_0$  and aperiodicity

Table 1: The configurations of multi-resolution STFT loss.

FFT size (pt)	4096	2048	1024	512	256
frame shift (pt)	400	200	100	50	25
window size (pt)	1600	800	400	200	100
Dim of Mel bands	640	320	160	80	40

measures extracted from natural speech samples were used in the test phase. Linguistic features for the input were computed frame by frame based on full context labels and normalized to take values of [0.01,0.99]. The dimension of the linguistic features was set to 675 for experiment (a) and 425 for experiments (b) and (c). For experiment (c), a speaker code (a one-hot vector) was input to the acoustic model.

The input and output of NAE were normalized spectral envelopes that sum to 1. The dimension of the latent variable was 200. The unit numbers of NAE were 1025-200-1025 in experiment (a) and (b), and 1025-500-200-1025 in experiment (c). The activation function in the hidden and output layers was softplus, and the hidden layer was normalized to sum to 1.

Feedforward networks were used for the acoustic model. The hidden layers were 6-layered and 1024 dimensional, and the activation function of the hidden layers was tanh. The output was 201 dimensions and the activation functions were softmax for the latent variables and softplus for the power coefficient.

For the DAE-based method in experiment , the input and output features of the DAE were logarithmic spectral envelopes normalized to a mean of 0 and variance of 1. The latent variables were set to 200 dimensions as in the proposed method. The decoder weights were the transpose matrix of the encoder weight matrix. The output layer of the DNN-based acoustic model was set to 200 dimensions and the activation function was tanh. The training algorithm was as follows. First, only the reconstruction of the DAE is trained, and then the DNN-based acoustic model is trained with the weights of the trained DAE fixed. Finally, fine-tuning of the entire networks is performed.

Stochastic gradient descent (SGD) was used for the optimization. The learning rate was set to 0.01 in NAE-based model and 0.0002 in DAE-based model.

For the NSF used in experiment (c), the network configuration and parameter setting were based on the released code of [13]. The frame length, frame shift length, and window length of multi-resolution STFT loss were set in the Table 1 with reference to the experimental conditions in [23]. The squared error of the Mel-log spectrum was added to the loss function to enhance the low-frequency bandwidth.  $F_0$  and logarithmic spectral envelopes were used for condition features.

As subjective experiments, preference AB tests on speech quality were conducted. In each test, 10 pairs of utterances from the two focused models were randomly selected and 25 subjects answered the questions. The user selects the one he or she perceives to be the better choice. Subjects were recruited via a crowdsourcing system.

### 3.3. Experimental results

#### 3.3.1. NAE-based SPSS for single speaker

In this experiment, in order to verify the effectiveness of the multi-task learning and the non-negative constraints in the proposed method, we compared the proposed method with the three methods (NAE\_FIX, JOINT and DAE). In NAE\_FIX, only the NAE reconstruction was trained at first, and the DNN-based acoustic model was trained with the decoder weights fixed. In JOINT, training of NAE reconstruction was not performed, and only TTS training was performed using the same

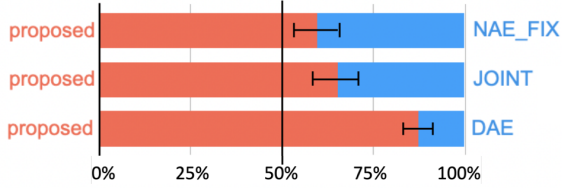


Figure 2: Results of subjective evaluations about naturalness. Error bars denote 95% confidential intervals.

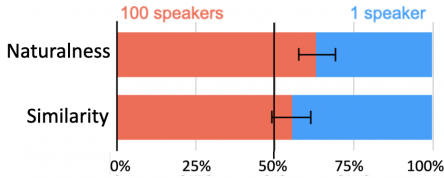


Figure 3: Results of subjective evaluations about naturalness and speaker similarity. Error bars denote 95% confidential intervals.

architecture as the proposed method.

The results of the subjective experiment are shown in Figure 2. The synthesized speech by the proposed method showed higher naturalness than that of **NAE\_FIX**. This indicates that the multi-task learning has resulted in features that are more suitable for TTS training. The proposed method also outperformed **JOINT**. This seems to be due to the efficient dimensionality reduction achieved by adding the constraint that the NAE is reconstructed. In addition, the proposed method performed better than **DAE**. This is presumably due to the non-negative constraints of NAE, which reduces dimensionality in such a way that fine structure of spectral envelopes is not lost.

### 3.3.2. NAE-based SPSS for multi-speaker

In this experiment, to confirm the effectiveness of learning NAE separately for each speaker, the quality of synthesized speech from the multi-speaker model and the speaker-dependent model was compared.

The results of the subjective evaluation experiment are shown in Figure 3. The superiority of the multi-speaker model was demonstrated for naturalness. There was not significant difference for speaker similarity.

The experimental results show that in the multi-speaker model of the proposed method, estimation of the speaker-independent latent variables and speaker adaptation by NAE were properly performed, resulting in improvements in naturalness compared to training on single-speaker data alone.

### 3.3.3. NSF wave generation

Figure 4 shows the comparison of the three models (**WLD**, **NSF** and **NSF\_WLD**) with AB tests on naturalness. WORLD vocoder was used in **WLD**, only NSF was used in **NSF** and NSF followed by WORLD resynthesis was performed in **NSF\_WLD** for waveform generation. **NSF\_WLD** performed better than **WLD**. This seems to be because in **WLD**, aperiodicity measures, which are highly dependent on spectral envelopes, were given independently of estimated spectral envelopes in the test phase. It was also shown that the waveform generation by NSF does not provide adequate phase control, and that the quality can be improved by minimum phase conversion by WORLD.

Table 2 shows the analysis results of mismatches in acoustic features.  $\mathbf{y}$  is the spectral envelope extracted from natural speech,  $\mathbf{y}_{rec}$  is that self-reconstructed from  $\mathbf{y}$  by NAE, and  $\mathbf{y}_{tts}$  is that estimated from the linguistic feature. Mel-cepstral

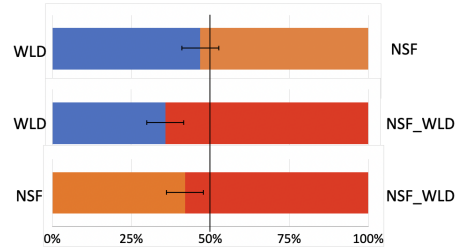


Figure 4: Results of subjective evaluations about naturalness and speaker similarity. Error bars denote 95% confidential intervals.

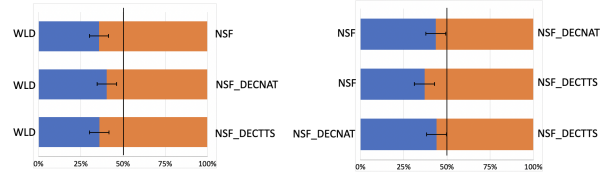


Figure 5: Results of subjective evaluations about naturalness and speaker similarity. Error bars denote 95% confidential intervals.

Table 2: Mel-cepstral distances between  $\mathbf{y}$ ,  $\mathbf{y}_{rec}$  and  $\mathbf{y}_{tts}$ .

Method	MCD( $\mathbf{y}$ , $\mathbf{y}_{rec}$ )	MCD( $\mathbf{y}_{rec}$ , $\mathbf{y}_{tts}$ )	MCD( $\mathbf{y}$ , $\mathbf{y}_{tts}$ )
<b>DECNAT</b>	1.62 dB	5.79 dB	6.20 dB
<b>DECTTS</b>	5.69 dB	2.62 dB	6.32 dB

distortions (MCD) between the three features were calculated. In **DECNAT**, it was confirmed that reconstructing  $\mathbf{y}$  with NAE decreased the distance to  $\mathbf{y}_{tts}$ . In **DECTTS**, MCD( $\mathbf{y}_{tts}$ ,  $\mathbf{y}_{rec}$ ) was much smaller, although the performance of acoustic feature estimation was reduced.

Figure 5 shows the comparison of the synthesized speech of four systems, **WLD**, **NSF\_WLD**, **NSF\_WLD (DECNAT)**, and **NSF\_WLD (DECTTS)**. **NSF\_WLD(DECNAT)** showed better naturalness than **NSF\_WLD**, and **NSF\_WLD(DECTTS)** showed even better naturalness. This is due to the reduction of acoustic mismatches during training and synthesis phase.

## 4. Conclusions

This paper introduced a TTS system using NAE-based spectral modeling. The experiment on the single-speaker dataset showed that the non-negative constraint and the simultaneous learning of dimensionality reduction and estimation of acoustic features helped to extract low-dimensional features more suitable for TTS, and improved the quality of the synthesized speech. We also studied multi-speaker modeling by training NAE separately for each speaker. Subjective experiments showed that the multi-speaker model outperformed the speaker-dependent model, and that the decoder weights of each NAE are trained to appropriately reflect speaker information. In the system of NAE-based TTS with NSF, the experimental results showed that minimum-phase conversion by WORLD resynthesis after waveform generation by NSF improved naturalness of synthesized speech. In addition, to reduce the mismatches between training and testing data of NSF, we proposed a framework in which spectral envelopes reconstructed by NAE were used for condition features in training phase. The subjective experiments showed the effectiveness of the approach.

## 5. Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 21H04900.

## 6. References

- [1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, and S. Bengio, "Tacotron: Towards end-to-end speech synthesis," *INTERSPEECH*, pp. 4006–4010, 2017.
- [2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, and R. Skerry-Ryan, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783, 2018.
- [3] S. Takaki, S. Kim, J. Yamagishi, and J. Kim, "Multiple feed-forward deep neural networks for statistical parametric speech synthesis," *INTERSPEECH*, pp. 2242–2246, 2015.
- [4] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 389–392, 1996.
- [5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems 13*, pp. 556–562, 2001.
- [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] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," *INTERSPEECH*, pp. 1118–1122, 2017.
- [8] Y.-C. Wu, P. L. Tobing, K. Yasuhara, N. Matsunaga, Y. Ohtani, and T. Toda, "A cyclical post-filtering approach to mismatch refinement of neural vocoder for text-to-speech systems," *INTERSPEECH*, 2020.
- [9] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 86–90, 2017.
- [10] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4475–4479, 2015.
- [11] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-based speech synthesis using speaker codes," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 2, pp. 462–472, 2018.
- [12] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2019.
- [13] X. Wang and J. Yamagishi, "Using cyclic noise as the source signal for neural source-filter-based speech waveform model," *arXiv preprint arXiv:2004.02191*, 2020.
- [14] 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.
- [15] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.
- [16] S. Goto, D. Saito, and N. Minematsu, "DNN-based statistical parametric speech synthesis incorporating non-negative matrix factorization," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 148–153, 2019.
- [17] S. Goto, Y. Shirahata, G. Kotani, H. Suda, D. Saito, and N. Minematsu, "The UTokyo speech synthesis system for Blizzard Challenge 2019," *Blizzard Challenge workshop 2019*, 2019.
- [18] W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Refined wavenet vocoder for variational autoencoder based voice conversion," *27th European Signal Processing Conference*, pp. 1–5, 2019.
- [19] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.
- [20] <http://hts.sp.nitech.ac.jp/>.
- [21] C. Veaux, J. Yamagishi, and K. MacDonald, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.
- [22] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," *INTERSPEECH*, pp. 2321–2325, 2017.
- [23] P.-C. Hsu and H.-Y. Lee, "WG-WaveNet: Real-time high-fidelity speech synthesis without GPU," *INTERSPEECH*, pp. 210–214, 2020.