THE EFFECT OF NEURAL NETWORKS IN STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

This paper investigates how to use neural networks in statistical parametric speech synthesis. Recently, deep neural networks (DNNs) have been used for statistical parametric speech synthesis. However, the specific way how DNNs should be used in statistical parametric speech synthesis has not been studied thoroughly. A generation process of statistical parametric speech synthesis based on generative models can be divided into several components, and those components can be represented by DNNs. In this paper, the effect of DNNs for each component is investigated by comparing DNNs with generative models. Experimental results show that the use of a DNN as acoustic models is effective and the parameter generation combined with a DNN improves the naturalness of synthesized speech.

Index Terms— Statistical parametric speech synthesis, deep neural network, hidden Markov model

1. INTRODUCTION

Statistical parametric speech synthesis based on hidden Markov models (HMMs) has grown in popularity in the last decade [1]. This approach offers various advantages over concatenative speech synthesis approach [2], e.g. the flexibility to change its voice characteristics [3, 4, 5, 6] and small footprint [7, 8]. In HMM-based speech synthesis, the spectrum, excitation, and duration of speech are simultaneously modeled with HMMs, and speech parameter sequences are generated from the HMMs themselves [9]. The speech parameter trajectory generated by HMM-based speech synthesis systems is fairly smooth. However, it is known that synthesized speech generated from HMM-based speech synthesis systems still sounds muffled and the quality of the synthesized speech still does not reached that of natural speech.

Recently, deep neural networks (DNNs) [10], which are feedforward artificial neural networks (ANNs) with many hidden layers, have achieved significant improvement in many machine learning areas. DNNs can represent high dimensional and correlated features efficiently. In addition, complex mapping functions can be modeled compactly by DNNs. Motivated by the success of DNNs in speech recognition, DNNs have been introduced to statistical parametric speech synthesis in order to improve the performance of speech synthesis [11, 12, 13]. In statistical parametric speech synthesis, a number of contextual features that affect speech, including phonetic, syllabic, and grammatical ones, have to be taken into account in acoustic modeling to achieve naturally sounding synthesized speech. Effective modeling of these complex context dependencies is one of the most critical problems for statistical parametric speech synthesis. In DNN-based acoustic modeling, a DNN is trained to represent the mapping function from contextual features to acoustic features, which are modeled by decision tree-clustered context dependent HMMs in HMM-based approach [9]. Zen *et al.* [11] showed that DNN-based acoustic models offer an efficient and distributed representation of complex dependencies between contextual and acoustic features. However, DNNs can be introduced to components other than acoustic modeling in statistical parametric speech synthesis and it should be further investigated how DNNs can be used in statistical parametric speech synthesis.

In this paper, we investigate how to use DNNs in statistical parametric speech synthesis. A generation process of statistical parametric speech synthesis based on generative models can be divided into several components, and those components can be represented by DNNs. By replacing DNNs with generative models in each component, the effect of DNNs in statistical parametric speech synthesis is investigated.

The rest of this paper is organized as follows. Section 2 describes statistical parametric speech synthesis based on HMMs and DNNs. The experimental conditions and results are shown in Section 3. Concluding remarks and future work are presented in Section 4.

2. STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING NEURAL NETWORKS

In statistical parametric speech synthesis, relation between contextual features and acoustic features is modeled by statistical models, which is generally called acoustic models. Hidden Markov models (HMMs), which are generative models that simulate the processes of generating observations, are usually used as acoustic models. In statistical parametric speech synthesis based on HMMs, the acoustic features, e.g. spectral and excitation features, and duration of speech are simultaneously modeled by HMMs. Since there are a number of contextual features that affect speech, they have to be taken into account in acoustic modeling. To effectively handle the contextual features, decision tree based context clustering [14] is widely used in HMM-based speech synthesis. The speech parameters are generated from the decision tree-clustered context dependent HMMs and given text to be synthesized. In order to generate smooth speech parameter trajectories, not only static features but also dynamic features are modeled by HMMs, and speech parameter trajectories are generated considering the relation between static and dynamic features by maximum likelihood parameter generation (MLPG) algorithm [15]. Figure 1 shows an overview of the generation procedures of HMM-based speech synthesis. From this figure, the generation process can be divided into two component:

- Component 1: generation of static and dynamic features from contextual features
- Component 2: generation of static features from static and dynamic features

In the generation process of HMM-based approach, the decision tree-clustered context dependent HMMs convert a contextual fea-

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

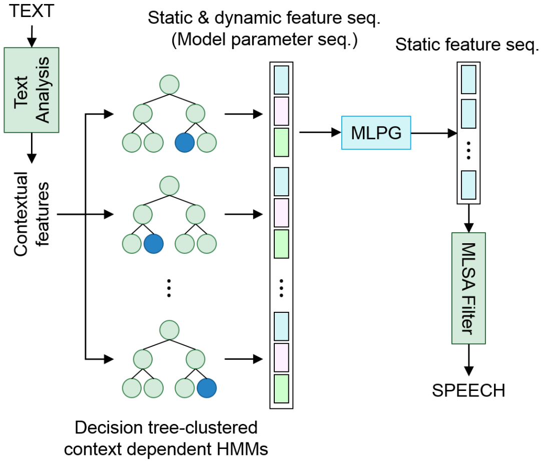

Fig. 1. An overview of the generation procedures in statistical parametric speech synthesis based on HMMs.

ture sequence into statistics of a static and dynamic feature sequence (component 1), and the parameter generation based on the MLPG algorithm convert a model parameter sequence into a smoothed static feature sequence (component 2).

In statistical parametric speech synthesis using DNN-based acoustic model [11, 12, 13], decision tree-clustered context dependent HMMs are replaced by a DNN. A single DNN is trained to represent a mapping function from contextual features to acoustic features including spectral and excitation parameters and their dynamic features. In the generation process, the contextual features extracted from given text to be synthesized are mapped to acoustic features by the trained DNN using forward propagation. Then, smooth speech parameters trajectories are generated by the MLPG algorithm in the same fashion as the HMM-based approach.

The use of DNNs as acoustic models is effective and potential to produce naturally-sounding synthesized speech have been shown. However, DNNs can be introduced to the other components in statistical parametric speech synthesis. DNNs can represent a mapping function from contextual features to static acoustic features and a mapping function from static and dynamic features to static features, which corresponds to parameter generation (component 2). Therefore, it should be further investigated how DNNs may be used in statistical parametric speech synthesis. In this paper, we compare speech synthesis systems that DNNs are used for each components mentioned above instead of generative models by objective and subjective evaluations. By comparing DNNs with generative models, the effect of DNNs for each component in statistical parametric speech synthesis is investigated.

3. EXPERIMENTS

3.1. Experiment 1

3.1.1. Experimental condition 1

Japanese 503 utterances, which can be downloaded from HTS web site[1], were used in the experiments. The contents of the data were

Table 1. Four speech synthesis systems for experiment 1.

	Component 1	Component 2
HMM+MLPG	HMM	MLPG
HMM+NN	HMM	NN
NN+MLPG	NN	MLPG
NN	NN	

the same as the B-set of the ATR phonetically balanced Japanese speech database [16]. The 450 utterances were used for training and the remaining 53 utterances were used for testing. Speech signals were sampled at a 48 kHz. Feature vectors were extracted with a 5 ms shift and the feature vector consisted of the 0th through 49th mel-cepstral coefficients and a log F_0 value. Mel-cepstral coefficients were extracted from the smoothed spectrum analyzed by the STRAIGHT [17].

Table 1 shows speech synthesis systems compared in this experiment. **HMM+MLPG** is a conventional HMM-based speech synthesis system [1]. The acoustic features modeled by HMMs consisted of mel-cepstral coefficients, log F_0 values and their dynamic features (delta and delta-delta). Five-state, left-to-right, no-skip hidden semi-Markov models (HSMMs) were used. To model log F_0 sequences consisting of voiced and unvoiced observations, a multi-space probability distribution (MSD) was used. **NN+MLPG** is the same structure as the system used in [11]. The input features for the DNN were 411 contextual features, including binary features and numerical features for contexts, and three duration features, including duration of the current phoneme and the position of the current frame. The output feature consisted of 154 acoustic features: mel-cepstral coefficients, a log F_0 value, their dynamic features (delta and delta-delta), and a voiced/unvoiced binary value. The input features for the DNN in **HMM+NN** were 1716 features that were 11-frame segmental features consisted of the HMM parameters: MSD weights, the mean vector for mel-cepstral coefficients, a log F_0 value, and their dynamic features. The output features were 52 acoustic features: mel-cepstral coefficients, a log F_0 value, and a voiced/unvoiced binary value. The input features for the DNN in **NN** were the same as those used in **NN+MLPG** and the output features were the same as those used in **HMM+NN**. The input and output features are time-aligned frame-by-frame by well-trained HMM models. The weights of the DNN were initialized randomly, then optimized to minimize the mean squared error between the output features of the training data and predicted values by a mini-batch stochastic gradient descent (SGD)-based back-propagation algorithm. The sigmoid activation function was used for hidden and output layers. A single network which modeled both spectral and excitation parameters was trained. The performance of the four systems were evaluated on objective and subjective measures.

3.1.2. Experimental results 1

To objectively evaluate the performance of the systems, mel-cepstral distortion (MCD) was used. The sizes of decision trees in **HMM+MLPG** were controlled by changing the tuning parameter α for the model complexity penalty term of the minimum description length (MDL) criterion [18] (α = 2, 1, 0.5, or 0.25). The DNNs used in **HMM+NN**, **NN+MLPG**, and **NN** had three hidden layers with different number of units per layer (256, 512, or 1024). Figure 2 shows the results of objective evaluation in MCD. All systems using DNNs showed similar performance regardless of

[1] http://hts.sp.nitech.ac.jp/

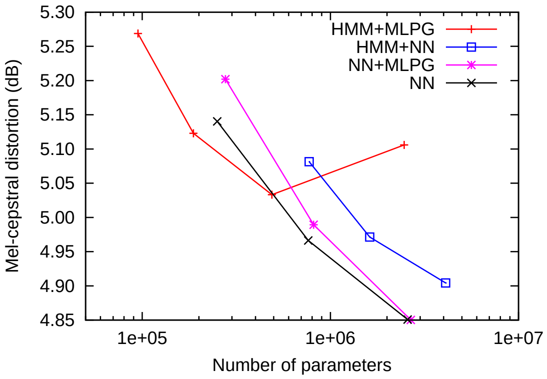

Fig. 2. Mel-cepstral distortions (dB) of the four speech synthesis systems. Note that the number of parameters for **HMM+NN** also includes the parameter of **HMM+MLPG** with the decision tree selected by the MDL criterion (α = 1).

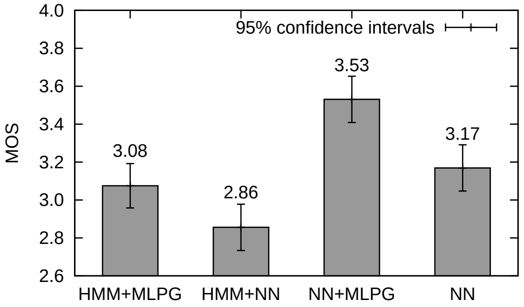

Fig. 3. Mean opinion scores of **HMM+MLPG**, **HMM+NN**, **NN+MLPG**, and **NN**.

the role of DNNs in the systems. As increasing the number of units per layer, the systems using DNNs improved the MCDs. Although the resultant MCDs of the systems using a DNN with 256 units per layer were worse than ones of **HMM+MLPG**, the systems using a DNN with many units per layer showed better performance than **HMM+MLPG**.

To evaluate the naturalness of the synthesized speech, a subjective listening test was conducted. In this evaluation, the decision tree structure selected by the MDL criterion (α = 1) was used for **HMM+MLPG** and **HMM+NN**, and the DNNs with 1024 units per layer were used for **HMM+NN**, **NN+MLPG**, and **NN**. The naturalness of the synthesized speech was assessed by the mean opinion score (MOS) test method. The subjects were eight Japanese students in our research group. Twenty sentences were chosen at random from the test sentences. Speech samples were presented in random order for each test sentence. In the MOS test, after listening to each test sample, the subjects were asked to assign the sample a five-point naturalness score (5: natural – 1: poor).

Figure 3 shows the subjective evaluation results. It can be seen from the figure that **NN+MLPG** outperform **HMM+MLPG**. This result indicates that replacing the decision tree-based clustered models into a DNN-based acoustic model is effective. On the other hand, **HMM+NN**, which used DNN instead of MLPG, showed the lowest score in all systems, though **HMM+NN** showed better performance than **HMM+MLPG** in the objective evaluation. Since a DNN represents a mapping function between input and output features based on a frame unit, **HMM+NN** cannot generate static features taking into account the neighboring features output from the DNN. As a result, **HMM+NN** generated discontinuous speech parameter trajectories. Similarly, **NN** also generated discontinuous speech parameter trajectories and thus **NN** showed worse score than **NN+MLPG**. These results indicate that the parameter generation based on MLPG, which can generate smooth parameter trajectories by considering the constraint between the static and dynamic features, is more appropriate than a DNN.

Table 2. Five speech synthesis systems for experiment 2.

	Component 1	Component 2
HMM+MLPG	HMM	MLPG
HMM+NN	HMM	NN
HMM+NN+MLPG	HMM	NN+MLPG
HMM+MLPG+NN	HMM	MLPG+NN
NN+MLPG	NN	MLPG

3.2. Experiment 2

3.2.1. Experimental condition 2

From the results of experiment 1, it can be seen that generating smooth speech parameter trajectories by the parameter generation with DNNs used in **HMM+NN** is difficult. To address this problem, we investigate the combination of MLPG-based parameter generation and DNN-based parameter generation. Table 2 shows speech synthesis systems compared in this experiment. In **HMM+MLPG+NN** and **HMM+NN+MLPG**, MLPG-based parameter generation is introduced before and after DNN-based parameter generation, respectively. The input features for the DNN in **HMM+NN+MLPG** were the same as those used in **HMM+NN** and the output features consisted of 154 acoustic features which were same as the output features for **NN+MLPG**. The input features for the DNN in **HMM+MLPG+NN** were 572 features that were 11-frame segmental features consisted of the mel-cepstral coefficients, a log F_0 value, and a voiced/unvoiced binary value, which were generated from MLPG-based parameter generation. The output features were the same as those used in **HMM+NN**. The other experimental conditions were the same as the experiment 1. The performance of the five systems were evaluated on objective and subjective measures.

3.2.2. Experimental results 2

The performance of the systems was objectively evaluated by mel-cepstral distortion (dB). Figure 4 shows the results of objective evaluation in mel-cepstral distortion (MCD). Experimental results show the similar trend to the results of the objective evaluation in experiment 1, i.e. the systems using a DNN with many units per layer showed better performance than **HMM+MLPG**.

The naturalness of the synthesized speech was evaluated by a subjective listening test. In this evaluation, the decision tree structure selected by the MDL criterion (α = 1) was used for acoustic models based on HMMs, and the DNNs with 1024 units per layer were used. The listening test setup was the same as the listening

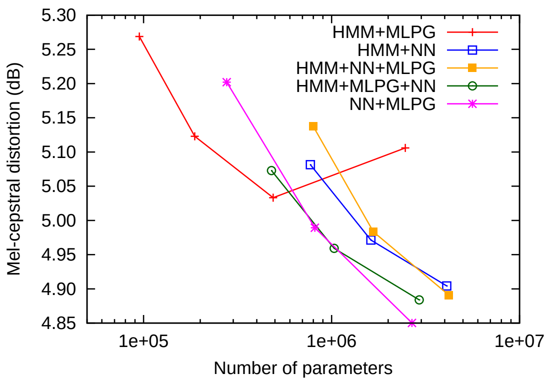

Fig. 4. Mel-cepstral distortions (dB) of the five speech synthesis systems. Note that the number of parameters for **HMM+NN**, **HMM+NN+MLPG** and **HMM+MLPG+NN** also include the parameter of **HMM+MLPG** with the decision tree selected by the MDL criterion ($\alpha = 1$).

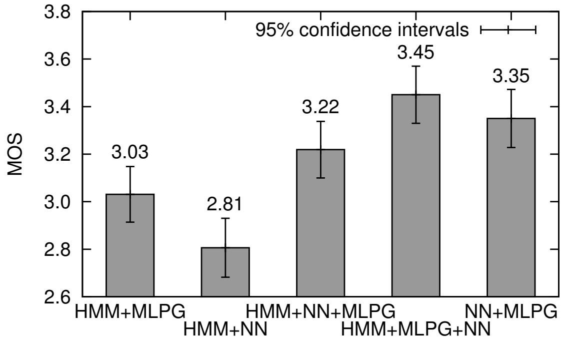

Fig. 5. Mean opinion scores of **HMM+MLPG**, **HMM+NN**, **HMM+NN+MLPG**, **HMM+MLPG+NN**, and **NN+MLPG**.

test in experiment 1. Figure 5 shows the subjective evaluation results. It can be seen from this figure that **HMM+NN+MLPG** and **HMM+MLPG+NN** outperformed **HMM+NN**. This is because the smooth speech parameter trajectories were generated by introducing MLPG-based parameter generation into **HMM+NN**. Comparing **HMM+MLPG+NN** with **HMM+NN+MLPG**, **HMM+MLPG+NN** showed a better score than **HMM+NN+MLPG**. Although speech parameter trajectories generated from MLPG are over-smoothed, the over-smoothing problem is alleviated by using DNN-based parameter generation after MLPG-based parameter generation. In addition, **HMM+MLPG+NN** showed better performance than **HMM+MLPG**. These results clearly show the effectiveness of the parameter generation combined with DNN.

The role of the DNN used in **HMM+MLPG+NN** can be regarded as DNN-based postfiltering. It is similar to the DNN-based postfiltering proposed by Chen *et al.* [19], but the DNN used in **HMM+MLPG+NN** represents the mapping function from acoustic features, which include mel-cepstral coefficients, log F_0 values, and voiced/unvoiced values, generated by the HMM-based system to acoustic features extracted from natural speech though the DNN used in the work by Chen *et al.* represents the mapping function from spectral envelopes of synthesized speech to one of natural speech. Recurrent neural networks (RNNs) are often used to model long-span sequential features and parameter generation process may be replaced by RNNs. RNN-based speech synthesis has been proposed and it shows good performance [20]. Therefore, the parameter generation based on RNNs should be compared with the parameter generation using DNNs, such as **HMM+MLPG+NN**. In statistical parametric speech synthesis based on HMMs, the parameter generation considering global variance (GV) [21] is widely used to enhance the dynamics within a speech utterance. Although this parameter generation method generates speech parameters taking account of the utterance-level features, i.e., the variance of acoustic features within a speech utterance, the DNN used in **HMM+MLPG+NN** cannot take into account the utterance-level features because the DNN represents the frame-level mapping function. Therefore, the parameter generation considering GV should be compared with the DNN-based parameter generation.

4. CONCLUSIONS

In this paper, we investigate the effect of DNNs in statistical parametric speech synthesis by comparing the speech synthesis systems that uses DNNs for each component on the objective and subjective measures. Experimental results show that replacing decision tree-clustered HMMs with a DNN is effective for modeling relation between contextual and acoustic features but replacing parameter generation based on MLPG with a DNN degrade the naturalness of synthesized speech due to generation of discontinuous speech parameter trajectories. To address this problem, the parameter generation combined with a DNN is used. It can be seen from experimental results that the parameter generation combined with a DNN can generate smooth parameter trajectories and improve the naturalness of synthesized speech.

In future work, we will investigate the effect of DNNs in statistical parametric speech synthesis on larger database. Additionally, future work also includes the comparison of the other parameter generation methods, such as RNN-based parameter generation and parameter generation considering GV.

5. ACKNOWLEDGEMENTS

The research leading to these results was partly funded by the Core Research for Evolutionary Science and Technology (CREST) program from the Japan Science and Technology Agency (JST).

6. REFERENCES

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[2] A. Hunt and A.W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” *Proceedings of ICASSP 1996*, pp. 373–376, 1996.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using mllr,” *Proceedings of ICASSP 2001*, pp. 805–808, 2001.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” *Proceedings of Eurospeech 1997*, pp. 2523–2526, 1997.

[5] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," *Proceedings of ICSLP 2002*, pp. 1269–1272, 2002.

[6] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information & Systems*, vol. E90-D, no. 2, pp. 533–543, 2007.

[7] S.J. Kim, J.J. Kim, and M.S. Hahn, "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, pp. 1384–1390, 2006.

[8] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, "Quantized HMMs for low footprint text-to-speech synthesis," *Proceedings of Interspeech 2010*, pp. 837–840, 2010.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proceedings of Eurospeech 1999*, pp. 2347–2350, 1999.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A.. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

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

[12] H. Lu, S. King, and O. Watts, "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis," *Proceedings of ISCA SSW8*, pp. 281–285, 2013.

[13] Y. Qian, Y. Fan, H. Wenping, and F.K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," *Proceedings of ICASSP 2014*, pp. 3857–3861, 2014.

[14] S. Young, J.J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *Proceedings of ARPA Workshop on Human Language Technology*, pp. 307–312, 1994.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proceedings of ICASSP 2000*, pp. 936–939, 2000.

[16] 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, pp. 357–363, 1990.

[17] 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.

[18] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," *Proceedings of Eurospeech 1997*, pp. 99–102, 1997.

[19] L.H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," *Proceedings of Interspeech 2014*, pp. 1954–1958, 2014.

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

[21] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information & Systems*, vol. E90-D, no. 5, pp. 816–824, 2007.