INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An Investigation of DNN-Based Speech Synthesis Using Speaker Codes

Nobukatsu Hojo[1], *Yusuke Ijima*[1], *Hideyuki Mizuno*[2]

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]Tokyo University of Science, Suwa
{hojo.nobukatsu, ijima.yusuke}@lab.ntt.co.jp, h.mizuno@rs.tus.ac.jp

Abstract

Recent studies have shown that DNN-based speech synthesis can produce more natural synthesized speech than the conventional HMM-based speech synthesis. However, an open problem remains as to whether the synthesized speech quality can be improved by utilizing a multi-speaker speech corpus. To address this problem, this paper proposes DNN-based speech synthesis using speaker codes as a simple method to improve the performance of the conventional speaker dependent DNN-based method. In order to model speaker variation in the DNN, the augmented feature (speaker codes) is fed to the hidden layer(s) of the conventional DNN. The proposed method trains connection weights of the whole DNN using a multi-speaker speech corpus. When synthesizing a speech parameter sequence, a target speaker is chosen from the corpus and the speaker code corresponding to the selected target speaker is fed to the DNN to generate the speaker's voice. We investigated the relationship between the prediction performance and architecture of the DNNs by changing the input hidden layer for speaker codes. Experimental results showed that the proposed model outperformed the conventional speaker-dependent DNN when the model architecture was set at optimal for the amount of training data of the selected target speaker.
Index Terms: Speech synthesis, acoustic model, deep neural network, speaker codes

1. Introduction

Recent studies have shown that deep neural network (DNN)-based speech synthesis [1, 2, 3] can produce more natural synthesized speech than the conventional hidden Markov model (HMM)-based speech synthesis. However, DNN-based speech synthesis requires a considerable amount of speech data uttered by the target speaker to obtain sufficient performance. The problem then becomes the high cost to generate speech from various speakers from DNN because we need a considerable amount of speech data from all the speakers the system uses, and annotations of phonetic and prosodic contextual information on them.

As for the field of HMM-based speech synthesis, many techniques have succeeded in generating speech from a smaller amount of the target speaker's data. A powerful method is an average-voice-based speech synthesis technique with model adaptation [4]. In this technique, average voice models are created from several speakers' speech data and are adapted with a small amount of speech data from a target speaker using model adaptation algorithms such as CSMAPLR [5]. Another successful method is based on cluster adaptive training (CAT) [6]. This model has multiple compact decision trees that are interpolated to produce a huge variety of possible contexts, and is trained using a multi-speaker speech corpus to improve the speech quality.

Motivated in a way similar to these previous studies in HMM-based speech synthesis, this work aims to improve the synthetic speech quality from a DNN by using a multi-speaker speech corpus. In speech recognition, one technique to model speaker variability in DNNs is to feed augmented speaker specific features like i-vectors [7] or speaker codes [8, 9] to the network in order to incorporate speaker-level information to the DNNs. This work is based on the assumption that such features can also be introduced to DNNs for speech synthesis.

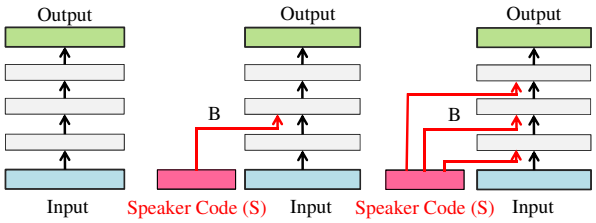

Figure 1: The architecture of DNNs. (left: the conventional model, middle: the proposed model using a single hidden layer, right: the proposed model using all hidden layers)

In a recent study [10] an experimental analysis was conducted using i-vector based feature augmentation. Although the evaluation in [10] focused on the speaker adaptation performance, the feature augmentation approach is also expected to improve the speech quality for each speaker used in the training process. Therefore, in the field of DNN-based speech synthesis, it still has not been revealed whether combining multiple speakers' speech corpora provides improved speech quality for the speakers in the corpora. Mainly to focus on this problem, this paper proposes to use augmented features based on speaker codes, which is a relatively simple method and has not yet been investigated precisely in the DNN-based speech synthesis field. Besides this main focus, the performance of the proposed model as an adaptation model to an unseen target speaker was also evaluated in a preliminary experiment.

2. Model Description

The baseline model is a DNN acoustic model similar to the one described in [1]. The baseline model is illustrated in the left side of Fig. 1. The DNN is used as a mapping function from the linguistic feature vectors to acoustic feature vectors. First, the input text is converted to the linguistic feature vector. The input features include binary answers to questions about linguistic contexts and numeric values. Then the linguistic feature vector is mapped to the output feature by forward propagation of DNN. The output features include spectral and excitation parameters and their time derivatives. The baseline model is trained using a single-speaker speech corpus.

As shown in the right side of Fig.1, in the proposed method, a speaker code $\mathcal{S}$ is fed to certain hidden layer(s) through an additional set of connection weights $\boldsymbol{B}$. Here the speaker code $\mathcal{S}$ represents the speaker information. These additional parameters of $\mathcal{S}$ and $\boldsymbol{B}$ in the proposed model are expected to represent the speaker characteristics in speech signals. The speaker codes can be fed to a certain hidden layer or all hidden layers as illustrated in the middle and the right side of Fig. 1, respectively. In this paper, the speaker code $\mathcal{S} = [s_1, \cdots, s_K]$ for speaker m is

http://dx.doi.org/10.21437/Interspeech.2016-589

set to the following fixed 1-of-K form for simplicity:

$$s_k = \begin{cases} 1 & (k = m) \\ 0 & (k \neq m) \end{cases} \quad (1)$$

where K is the dimension of $\mathcal{S}$ and equal to the number of speakers in the training data. The proposed method trains connection weights of the whole DNN using a multi-speaker speech corpus. When synthesizing a speech parameter sequence, a target speaker is chosen from the corpus and the speaker code corresponding to the selected target speaker is fed to the DNN to generate the speaker's voice. The proposed model is expected to generate more stable and natural speech because the networks from the linguistic feature to the acoustic feature are trained with a greater variety of contextual information by using a multi-speaker speech corpus. Furthermore, the speech from the proposed model is expected to have high similarity to the target speaker because the additional low dimensional networks $\boldsymbol{B}$ are expected to be trained effectively by using a smaller amount of speech data for each speaker.

3. Experiments

3.1. Experimental setup

In the experiments, we used speech data in Japanese from 35 speakers (17 male and 18 female speakers). Two speakers, one male and one female from the speech database, were used as target speakers. The training corpus includes 7260 utterances (about 1340 minutes) from 33 speakers apart from the two target speakers. We considered two training conditions: 5 utterances (about 1.2 minutes) and 300 utterances (about 63 minutes) for the training corpus for each target speaker. Twenty utterances were used as a testing set for each target speaker. The sampling rate of the corpus was 22.05 kHz. The STRAIGHT vocoder [11] was employed to extract 40 dimensional mel-cepstral coefficients, 5 band aperiodicities, and F0 in log-scale at 5 msec steps. We compared the performance of the following 4 acoustic models.

- SD: The speaker dependent DNN [1].
- HMM: The average voice model with adaptation.
- SPKCODE: The proposed model using speaker codes.
- SPKCODE_ADAPT: The proposed model with adaptation (The details are described in sec. 4.).

Prior to the following experiment, we compared the performance of the HMM-based method in two experimental conditions: the utterances by the target speaker were used (1) only in the adaptation process and (2) both in the average-voice-model training and the adaptation process. In the following experiments, we present the results of (2) because they showed slightly better performance than the results of (1).

The input vector of a DNN contained 506 dimensional linguistic features. Each observation vector consisted of 40 Mel-cepstral coefficients, log F0, 5 band aperiodicities, their delta and delta-delta features, and a voiced/unvoiced binary value. The input numeric features were normalized to the range of [0.01, 0.99], and the output features were normalized by speaker-dependent mean and variance. The DNN systems had 5 hidden layers and each hidden layer had 1024 units. A sigmoid function was used in the hidden layers followed by a linear activation at the output layer. For the training procedure, 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, using the Adam [12]-based back-propagation algorithm. The parameters for Adam algorithm were set as $\alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1\mathrm{e}-8$. Five percent of the utterances of the whole training data were used as a development set. The SD models were trained using only the target speaker utterances in the training corpus.

For the HMM, we used a five-state left-to-right hidden semi-Markov model with no skip topology. Each observation vector consisted of 138 features (40 Mel-cepstral coefficients, log F0, 5 band aperiodicities, and their delta and delta-delta features). The output distribution in each state was modeled as a single Gaussian density function, and the covariance matrices were assumed to be diagonal. We used the combined technique of CSMAPLR and MAP adaptation as the speaker adaptation algorithm [5]. The model size was determined automatically by the minimum description length (MDL) criterion [13], where the control parameter of the model size was set to $\alpha = 1.0$.

For all of the three methods evaluated in these experiments, segmentations (phoneme durations) from natural speech were used instead of predicting duration. We applied MLPG [14] to the output features for all of the three methods. For DNN-based methods, we used pre-computed variances from the training data for MLPG. We did not apply spectral enhancement techniques such as global variance [15] to reduce factors considered in the experiments.

3.2. Objective evaluation

We first investigated the relationship between the prediction performance and the architecture, i.e., the input hidden layer(s) for speaker codes (the 1st, 2nd, 3rd, 4th, 5th hidden layer and all hidden layers). Then, we compared the performance of the proposed method with that of the conventional methods.

3.2.1. Evaluation using 5 target speaker utterances

Figure 2 presents the mel-cepstral distortions (MCDs) and RMSEs of log F0 when using 5 target speaker utterances. Among the investigated model architectures, the models using all hidden layers for speaker codes gave lower MCDs and higher F0 RMSEs than most models using a single hidden layer. Among the models using a single layer for speaker codes, the models using the 2nd, 3rd or 4th hidden layer tended to give lower MCDs and F0 RMSEs than those obtained using the 1st or 5th hidden layer. One reason for this is that the connection weights to the first layer are difficult to train because of vanishing gradients. The other reason is that the model using the last layer could not represent the speaker characteristics precisely because augmented feature connections to the last hidden layer represent global transformation in the acoustic feature space, which can be partially substituted by speaker-wise feature space normalization. We decided that the models using the 4th layer were at optimal among the models trained using 5 target speaker utterances because both of the MCDs and F0 RMSEs are consistently low for each target speaker.

We then compared the performance of SPKCODE and HMM. We can see that the MCDs of SPKCODE are lower than HMM when the model architecture was at optimal. This is because the DNN-based methods has an advantage in modeling complex context dependencies over the tree-clustered HMM-based methods [1]. As for F0 RMSEs, the relation between SPKCODE and HMM differed for the two target speakers and showed no consistent tendency.

Figure 3 compares the performance between SPKCODE and SDs. The plots in this figure show the performance of SDs trained using different numbers of training data elements (10, 20, 50, 100, 200 or 300 utterances). The solid lines show the performance of SPKCODEs trained using 5 target speaker utterances. We can see that the SPKCODE gives MCDs equivalent to SDs trained using 50$\sim$100 utterances, and gives F0 RMSEs equivalent to SDs trained using 100$\sim$300 utterances. These results confirmed the prediction performance improvement of the proposed method. The proposed method showed greater performance improvement in F0 prediction than in mel-cepstral prediction. This is because accurate F0 prediction needs greater variety in contexts in training data than mel-cepstra prediction in order to model its complex dependency on prosodic information such as accent type and mora positions. The proposed model has an advantage in F0 prediction performance improvement because the model can be trained with a greater variety in contexts by using a multi-speaker speech corpus.

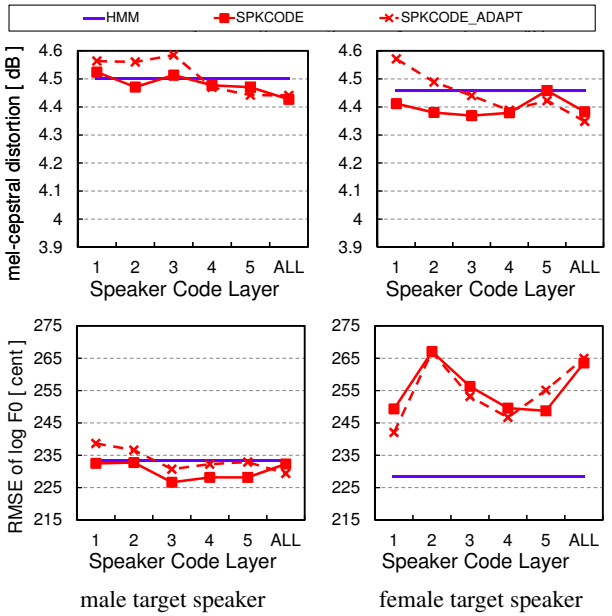

Figure 2: The objective evaluation results (The number of target speaker utterances: 5).

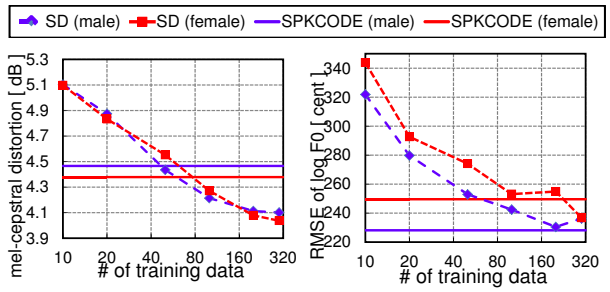

Figure 3: Performance of the proposed model and the conventional speaker dependent model (The number of target speaker utterances: 5) .

3.2.2. Evaluation using 300 target speaker utterances

Figure 4 presents the mel-cepstral distortions and RMSEs of log F0 when using 300 target speaker utterances. As with the evaluation using 5 target speaker utterances, MCDs by the model using all hidden layers were lower than most models using a single hidden layer while the relation between F0 RMSEs and the architecture differed for the target speakers. From these results, we decided that the models using all hidden layers were at optimal among models trained using 300 target speaker utterances because the MCDs and F0 RMSEs were consistently low for each of the target speakers.

We then compared the performance of SPKCODE, SD, and HMM. We can see the relations of SPKCODE $<$ SD $<$ HMM for MCDs and HMM $<$ SPKCODE $<$ SD for F0 RMSEs. For all experimental conditions, SPKCODE outperformed SD. These results confirmed the prediction performance improvement of the proposed method using a multi-speaker speech corpus. When SPKCODE is compared with HMM, it is found that MCDs of SPKCODE are lower than those for HMM while F0 RMSEs of SPKCODE are higher. The advantage of DNN over HMM was confirmed for both conditions using 5 and 300 target speaker utterances. We can see from Figs. 2 and 4 that F0 RMSEs of SPKCODE were equivalent to or higher than HMM when using either 5 or 300 target speaker utterances. These results revealed that the proposed method needs more accurate F0 prediction performance.

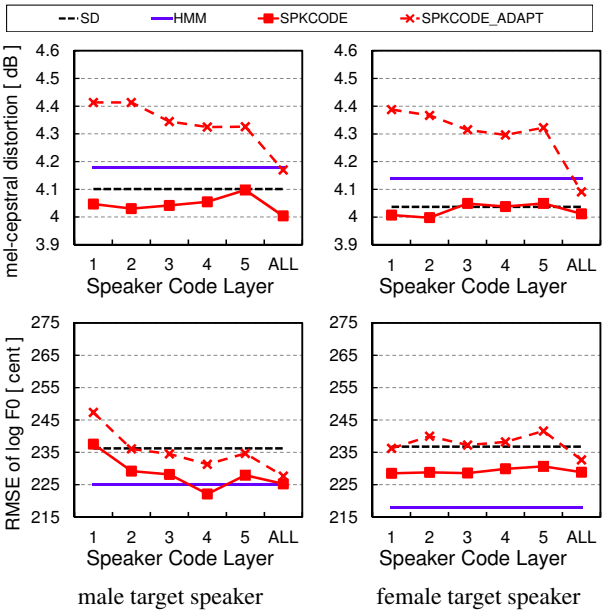

Figure 4: The objective evaluation results (The number of target speaker utterances: 300).

3.3. Subjective evaluation

We conducted subjective evaluations with respect to naturalness and similarity of the synthesized speech to confirm the effectiveness of the proposed method. We used the optimal architectures for the proposed method as discussed in the last section; the model using the 4th hidden layer when using 5 target speaker utterances and the model using all hidden layers when using 300 target speaker utterances. The number of listeners was 24 for a naturalness test and 22 for a similarity test. We conducted five-point MOS and DMOS tests. The scale for the MOS test was 5 for "very natural" and 1 for "very unnatural". The scale for the DMOS test was 5 for "very similar" and 1 for "very dissimilar".

Figures 5 and 6 show the naturalness and similarity scores obtained in the subjective evaluations with confidence intervals of 95%. We can see the relation of HMM $<$SPKCODE for both naturalness and similarity scores. Furthermore, the scores of SPKCODE using 5 target speaker utterances were equivalent to those of SD using 300 target speaker utterances. We can also see the relation of SD $<$ SPKCODE for both naturalness and similarity when using 300 target speakers utterances. These results confirmed that the proposed method can improve the synthetic speech quality by using a multi-speaker speech corpus in DNN-based speech synthesis. On the other hand, there were no significant differences between HMM and SPKCODE when using 300 target speaker utterances. The objective evaluation results suggest the need for further development of the proposed method for more accurate F0 prediction in order to give better performance than the HMM-based method. The degraded F0 prediction performance compared to HMMs has been already reported for the DNN architecture used in this experiment [1]. To address this problem, the recent research [16] has shown that different model architectures give F0 prediction performance improvement. Our future work will include evaluating F0 prediction by incorporating speaker codes into the model architecture in [16]

4. Preliminary study on speaker adaptation

4.1. Motivation and method

Although the experimental results in the last section confirmed the effectiveness of the proposed method, high computational

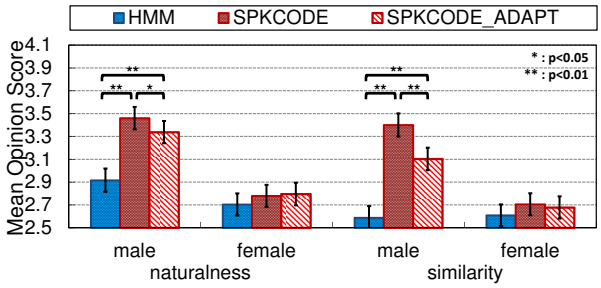

Figure 5: Naturalness and similarity test results with their 95% confidence interval. (The number of target speaker utterances: 5)

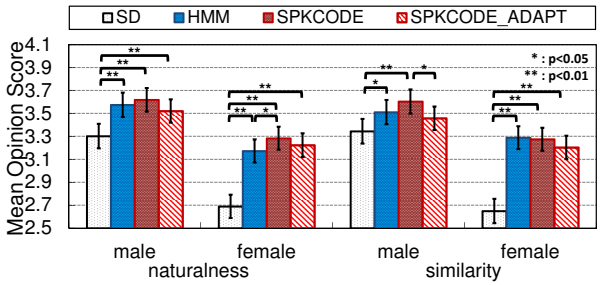

Figure 6: Naturalness and similarity test results with their 95% confidence interval. (The number of target speaker utterances: 300)

cost will be needed when we want to generate speech from a new target speaker. This is because the whole model needs to be retrained using a corpus including the new speaker's utterances. It is considered that model adaptation by reestimating a subset of model parameters using the target speaker utterances can address this problem. In addition to the previous researches of speaker adaptation for speech synthesis [10, 17, 18, 19], it is considered that the speaker code based DNN can also be used as a speaker adaptation method, since this approach has been shown to be effective in speech recognition [8, 9]. This section reports a preliminary study we conducted on speaker adaptation using speaker codes.

In this study, the DNN is adapted in the following procedure. First, the connection weights of the whole DNN are trained using a multi-speaker speech corpus. The speaker code $\mathcal{S}$ is set in a form similar to that given in sec. 2, but this time $\mathcal{S}$ is appended by an additional dimension to have $K+1$ dimensions in total. This additional dimension is used to represent an unseen target speaker information, and always set to 0 in the training procedure. Second, the model is adapted to a new target speaker using only the target speaker utterances as adaptation data. This time the additional dimension of $\mathcal{S}$ is set to 1 and other dimensions to 0, and only the connection weights $\boldsymbol{B}$ are reestimated to minimize the distance between the output features of the adaptation data and predicted values. When synthesizing, the speaker code $\mathcal{S}$ whose additional dimension is set to 1 is used.

This adaptation procedure is different from those in [8, 9]; those references estimate both $\mathcal{S}$ and $\boldsymbol{B}$ for training and reestimate $\mathcal{S}$ for adaptation, while the procedure in this study estimates and reestimates only $\boldsymbol{B}$ and uses a fixed $\mathcal{S}$ for both training and adaptation. Those references are for speaker adaptation for speech recognition and mainly focus on fast and robust adaptation with a limited amount of adaptation data. On the other hand, for speech synthesis, it is needed to model the speaker characteristics precisely to generate the target speaker's speech. The procedure in this study is expected to be flexible and more appropriate for speech synthesis, because it reestimates $\boldsymbol{B}$ whose dimension is generally larger than $\mathcal{S}$ for adaptation. From these expectations, as the first step of performance evaluation of speaker code based adaptation for speech synthesis, this study chose to conduct preliminary experiment in the procedure to reestimate $\boldsymbol{B}$.

4.2. Experimental results

In the following experiments, the target speaker's utterances were excluded from the training corpus and were used as adaptation data. The optimization parameters for adaptation procedure were set to the same values as the training procedure in sec. 3.1. The number of the development set was 1 for 5 adaptation utterances and 15 for 300 adaptation utterances. The other experimental conditions were set in a way similar to sec. 3.1. In order to make it reasonable to compare the scores with ones in sec. 3.3, the subjective evaluation was conducted simultaneously with ones in sec. 3.3 to eliminate the perceptual bias from the results.

The subjective evaluation results for SPKCODE_ADAPT are shown in Figs. 5 and 6. Although the scores of SPKCODE_ADAPT were slightly lower than SPKCODE under each condition, they were higher than those of HMM when using 5 target speaker utterances and comparable with HMM and higher than SD when using 300 target speaker utterances. These results confirmed that the adaptation based on speaker codes generates speech whose quality is comparable with or higher than conventional speaker dependent DNN and speaker adapted HMM.

We then analyzed the performance of SPKCODE_ADAPT from the objective evaluation results in Figs. 2 and 4. When using 300 target speaker utterances, the mel-cepstral prediction performance of SPKCODE_ADAPT was degraded greater from that of SPKCODE than the other experimental conditions. In addition, the prediction performance was even worse than those of SDs. From these results, there is concern that the adaptation models in this study could not represent the precise speaker characteristics by effectively using the large amount of adaptation data. Although this study chose to conduct adaptation with more free parameters than ones in [8, 9], the objective evaluation results for MCDs rather suggested that more flexible adaptation models were needed when given a large amount of adaptation data. To solve this problem, one promising approach would be a more elaborate speaker code based adaptation method whose number of parameters for adaptation can change concerning the amount of adaptation data, just as CSMAPLR [4] in HMM-based adaptation method.

5. Conclusion

In this paper, we proposed a DNN-based speech synthesis method using speaker codes to improve speech quality by using a multi-speaker speech corpus. Experimental results showed that the optimal architecture of the proposed model depends on the amount of target speaker utterances in the training data. The objective and subjective evaluation results showed that the proposed model can produce more natural speech than the conventional speaker dependent method. The experimental results also suggested that the proposed model still has room for F0 prediction performance improvement. Future works will include evaluating performance when speaker codes are incorporated into the model architecture in [16] to investigate this point. We then conducted a preliminary study to use a speaker code based DNN as a speaker adaptation model. The experimental results showed that the adaptation using speaker codes can generate speech with quality comparable to or better than the conventional methods, while it suggested the need for more elaborate adaptation technique which can change the number of parameters for adaptation concerning the amount of adaptation data.

6. References

[1] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[2] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3844–3848.

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

[4] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[5] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions. Audio, Speech, and Language Process.*, vol. 17, no. 1, pp. 66–83, 2009.

[6] V. Wan, J. Latorre, K. Chin, L. Chen, M. Gales, H. Zen, K. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," in *Proc. INTERSPEECH*, 2012, pp. 1134–1137.

[7] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, 2013, pp. 55–59.

[8] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7942–7946.

[9] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1713–1725, 2014.

[10] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "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, no. 3, pp. 187–207, 1999.

[12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EUROSPEECH*, 1997, pp. 99–102.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1315–1318.

[15] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[16] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai, "Modeling F0 trajectories in hierarchically structured deep neural networks," *Speech Communication*, vol. 76, pp. 82–92, 2016.

[17] B. Potard, P. Motlicek, and D. Imseng, "Preliminary work on speaker adaptation for dnn-based speech synthesis," Idiap, Tech. Rep., 2015.

[18] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for DNN-based TTS synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5135–5139.

[19] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5540–5544.