9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis

Shinji Takaki[1], *SangJin Kim*[2], *Junichi Yamagishi*[1,3]

[1]National Institute of Informatics (NII), Tokyo, 101-8430, Japan
[2]Naver Labs, Naver Corporation, Seongnam, 463-867, Korea
[3]University of Edinburgh, Edinburgh, EH8 9LW, United Kingdom
takaki@nii.ac.jp, sangjin.kim@navercorp.com, jyamagis@nii.ac.jp

Abstract

In this paper, we investigate the effectiveness of speaker adaptation for various essential components in deep neural network based speech synthesis, including acoustic models, acoustic feature extraction, and post-filters. In general, a speaker adaptation technique, e.g., maximum likelihood linear regression (MLLR) for HMMs or learning hidden unit contributions (LHUC) for DNNs, is applied to an acoustic modeling part to change voice characteristics or speaking styles. However, since we have proposed a multiple DNN-based speech synthesis system, in which several components are represented based on feed-forward DNNs, a speaker adaptation technique can be applied not only to the acoustic modeling part but also to other components represented by DNNs. In experiments using a small amount of adaptation data, we performed adaptation based on LHUC and simple additional fine tuning for DNN-based acoustic models, deep auto-encoder based feature extraction, and DNN-based post-filter models and compared them with HMM-based speech synthesis systems using MLLR.
Index Terms: Statistical Parametric Speech Synthesis, Deep Neural Network, Speaker Adaptation, Learning Hidden Unit Contributor

1. Introduction

Statistical speech synthesis research has been significantly advanced thanks to deep neural networks (DNNs) with many hidden layers. For example, DNNs have been applied for acoustic modeling. Zen et al. used a DNN to learn the relationship between input texts and extracted features instead of decision tree-based state tying [1]. Restricted Boltzmann machines or deep belief networks have been used to model output probabilities of hidden Markov model (HMM) states instead of GMMs [2]. Recurrent neural networks and long-short term memories have been used for prosody modeling [3] and acoustic trajectory modeling [4]. In addition, an auto-encoder neural network has also been used to extract low dimensional excitation parameters [5]. Furthermore a DNN-based probabilistic post-filter has also proposed [6], where a DNN is used to model the conditional probability of the spectral differences between natural and synthetic speech so that the fine spectral structure lost during modeling can be reconstructed at synthesis time.

In statistical parametric speech synthesis, changing characteristics or speaking styles using a small amount of training data is an important research topic. In HMM-based speech synthesis the well-established methods called MLLR [7, 8] or constrained MLLR [9, 10] are frequently used for speaker or speaking style adaptation. Moreover, various adaptation techniques such as vocal tract length adaptation [11], eigen voice [12] and clustering adaptive training [13] have also been proposed. In DNN-based recognition and synthesis fields, several adaptation techniques have been proposed, e.g., the use of speaker codes or speaker i-vectors as additional inputs of DNNs [14, 15, 16], training with regularization [17], multiple basis adaptation [18], matrix factorization [19], and adaptation of hidden units outputs with an additional small number of parameters [20, 21, 16].

In the meantime, since the DNN framework can be used not only for the acoustic modeling but also other modules, we have proposed a new speech synthesis system [22] where several standard steps of the statistical speech synthesis including the feature extraction from STRAIGHT spectral amplitudes [23], acoustic modeling, smooth trajectory generation and spectral post-filter are conducted using multiple DNNs. In [22], we have constructed three feed-forward DNNs for performing these standard steps in a data-driven way and confirmed that this system effectively provides higher quality of synthetic speech.

In this paper, we investigate the effectiveness of the speaker adaptation of several components of the above speech synthesis system, including acoustic feature extraction, acoustic modeling and post-filtering. Since, as mentioned above, these components are based on feed-forward DNNs, we can use the same adaptation approaches of feed-forward DNNs as the acoustic model adaptation for the adaptation of the feature extraction and post-filter models. Most of the adaptation approaches for DNNs reported in the past have been applied to an acoustic modeling part [16]. In contrast, this paper analyzes the performance of the speaker adaptation of different components and combined ones and analyzes whether such adaptation techniques are as effective as the adaptation of the acoustic models or not. For performing speaker adaption of the feed-forward DNNs, LHUC [21] and simple additional fine tuning are performed using a small amount of adaptation data, and we compare them with HMM-based speech synthesis systems using MLLR in objective and subjective experiments.

The rest of this paper is organized as follows. Section 2 shows text-to-speech synthesis based on multiple DNNs. In Section 3, the LHUC adaptation technique is described. The experimental conditions and results are shown in Section 4. Concluding remarks and future work are presented in Section 5.

2. Text-to-speech Synthesis based on Multiple DNNs

In this section, text-to-speech synthesis based on multiple DNNs [22], which can perform all standard steps of the statistical parametric speech synthesis from end to end, is briefly

http://dx.doi.org/10.21437/SSW.2016-25

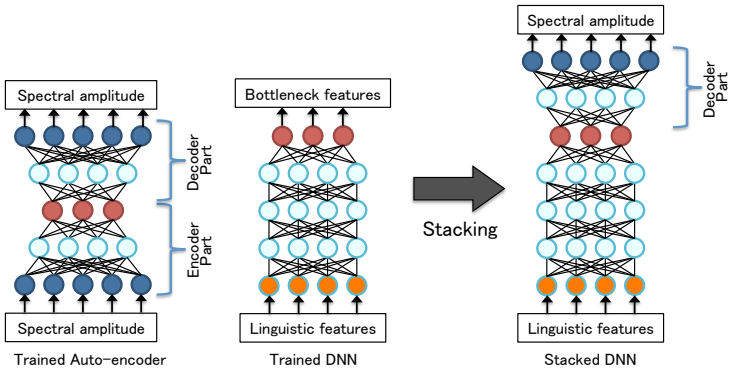


Figure 1: Procedure for constructing a DNN-based spectral model based on a deep auto-encoder and a DNN-based acoustic model.

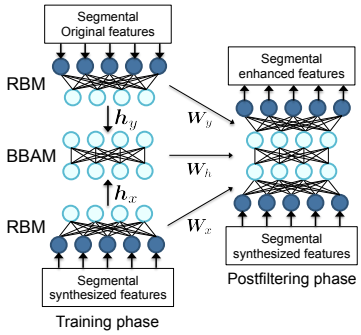


Figure 2: Structure of a DNN for the post-filter. Training and post-filtering procedures are also shown.

described. This system is based on three feed-forward DNNs, i.e., DNN-based acoustic feature extraction, DNN-based acoustic modeling, and DNN-based post-filtering.

2.1. DNN-based Acoustic Feature Extraction

An auto-encoder is an artificial neural network that is used generally for learning a compressed and distributed representation of a dataset. It consists of an encoder and a decoder. In the basic one-hidden-layer auto-encoder, the encoder maps an input vector $\mathbf{x}$ to a hidden representation $\mathbf{y}$ as follows:

$$\mathbf{y} = f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{1}$$

where $\theta = \{\mathbf{W}, \mathbf{b}\}$. $\mathbf{W}$ and $\mathbf{b}$ represent an $m \times n$ weight matrix and a bias vector of dimensionality m, respectively, where n is the dimension of $\mathbf{x}$. The function s is a non-linear transformation on the linear mapping $\mathbf{W}\mathbf{x} + \mathbf{b}$. The output of the encoder $\mathbf{y}$ is then mapped to the output of the decoder $\mathbf{z}$,. The mapping is performed by a linear mapping followed by an arbitrary function t that employs an $n \times m$ weight matrix $\mathbf{W}'$ and a bias vector of dimensionality n as follows:

$$\mathbf{z} = g_{\theta'}(\mathbf{y}) = t(\mathbf{W}'\mathbf{y} + \mathbf{b}'), \tag{2}$$

where $\theta' = \{\mathbf{W}', \mathbf{b}'\}$. An auto-encoder can be made deeper by stacking multiple layers of encoders and decoders to form a deep architecture.

A deep auto-encoder allows us to extract robust low-dimensional features automatically from high-dimensional spectral envelopes in a non-linear, data-driven and unsupervised way. In this paper, we apply the deep auto-encoder to log STRAIGHT spectral envelopes for extracting low-dimensional features based on the same way used in [22].

2.2. DNN-based Acoustic Model

In this work, we construct a DNN that directly synthesizes high-dimensional spectral amplitudes from linguistic features without using spectral envelope parameters such as mel-cepstrum [22]. In this technique, we stack two DNNs, an auto-encoder neural network for data-driven non-linear feature extraction from the spectral amplitudes and another network for acoustic modeling to train the DNN efficiently.

Fig. 1 shows the procedure for constructing the proposed DNN-based spectral model. The steps of the proposed technique are as follows.

Step 1. Train a deep auto-encoder using spectral amplitudes and extract bottleneck features.

Step 2. Train a DNN-based acoustic model using the bottleneck features extracted in Step 1.

Step 3. Stack the trained DNN-based acoustic model for bottleneck features and the decoder part of the trained deep auto-encoder as shown in Figure 1 and optimize the all networks.

A DNN that represents the relationship between linguistic features and spectra is constructed based on DNN-based spectral feature extraction and a DNN-based acoustic model using the bottleneck features. After this procedure, we can fine-tune the DNN to minimize the error over the entire dataset using pairs of linguistic features and spectral amplitudes in training data with stochastic gradient descend (SGD).

2.3. DNN-based Post-filter in the Spectral Domain

A feed-forward DNN for probabilistic modeling of the differences between spectra of the synthesized and natural speech has been proposed [6]. Fig. 2 shows the structure of the DNN based post-filter. In this paper, the DNN is trained layer-by-layer using two restricted Boltzmann machines (RBMs) [24] and a Bernoulli bi-directional associative memory (BBAM) [25] as shown in Figure 2. After this layer-wised pre-training, the DNN for the post-filter is fine-tuned using backpropagation. This model is directly applied to the high-dimensional spectral amplitudes.

3. Speaker Adaptation based on Learning Hidden Unit Contributions (LHUC)

The LHUC technique has been proposed for speaker adaptation in DNN-based speech recognition and synthesis [21]. LHUC is an effective speaker adaptation technique with a small number of adaptation parameters. In this technique, additional speaker adaptation parameters for each hidden unit, $\mathbf{r}$, are defined to modify hidden layer outputs as follows:

$$\mathbf{h}_m^l = a(\mathbf{r}_m^l) \circ s^l(\mathbf{W}^{l\top}\mathbf{h}_m^{l-1}), \tag{3}$$

where m, l, and $\circ$ represent a speaker index, a hidden layer index, and an element-wise multiplication, respectively. In this paper, the function a is defined as a sigmoid with amplitude 2, that is, $a(c) = 2/(1 + exp(-c))$ in a way similar to [21]. This technique regards $a(\cdot)$ as a scaling factor of a hidden unit output

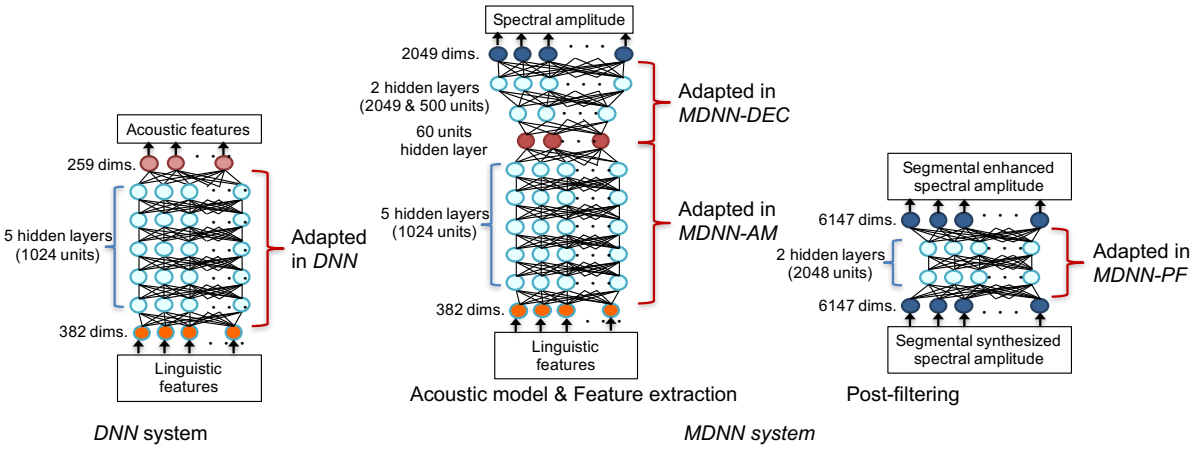


Figure 3: Network configurations in *DNN* and *MDNN* and adaptation components for each adapted system.

for a targeted speaker. Adaptation parameters can be updated with the same manner used in [21], although the mean square error (MSE) criterion is used as the loss function for our experiment. Speaker adaptation using a very small amount of adaptation data would be accomplished using LHUC nicely since the total number of adaptation parameters required is much lower compared with the number of all parameters included in the DNN.

One of the advantages of this technique is that LHUC can be applied to any trained feed-forward deep neural network. The multiple DNN-based speech synthesis system described above is based on three feed-forward neural networks, so LHUC can be applied not only to the acoustic modeling part but also to other components.

We compare the LHUC technique as well as the simple additional fine tuning for the adaptation of DNN-based acoustic models, deep auto-encoder based feature extraction, and DNN-based post-filter models in the following experiments.

4. Experiments

We evaluated the proposed systems in objective and subjective experiments using English databases. The database provided for the Blizzard Challenge 2011 [26], which contains approximately 17 hours of speech data, comprising 12K utterances uttered by a female speaker, was used for constructing base systems for speaker adaptation. Speaker adaptation was carried out based on the base systems with a small amount of adaptation data uttered by a different female speaker. A total of 116 utterances were used for test sets.

We constructed three base systems: *HMM* is a HMM-based speech synthesis system with a GV technique [27], *DNN* is a single DNN-based speech synthesis system with a signal processing-based post-filter for cepstrum vectors [28], and *MDNN* is a multiple DNN-based speech synthesis system [22].

In addition, we have constructed eleven speaker adapted systems: *HMM-AM-MLLR*, *DNN-AM-FT*, *DNN-AM-LHUC*, *MDNN-AM-FT*, *MDNN-AM-LHUC*, *MDNN-DEC-FT*, *MDNN-DEC-LHUC*, *MDNN-PF-FT*, *MDNN-PF-LHUC*, *MDNN-ALL-FT* and *MDNN-ALL-LHUC*. The three components of each system name refer to the base systems, adaptation parts and adaptation techniques used for constructing each system, respectively. In the system names, *AM*, *DEC*, *PF* and *ALL* represent adaptation parts, i.e., acoustic modeling, decoder, post-filtering parts and all networks, respectively. *MLLR*, *FT* and *LHUC* represent adaptation techniques, i.e., maximum likelihood linear regression, a simple additional fine tuning and lerning hidden unit contributions using the adaptation data, respectively. Three systems (*HMM-AM-MLLR*, *DNN-AM-FT* and *DNN-AM-LHUC*) were conventional systems. For adaptation of a post-filtering part, spectral features were synthesized through two DNNs for the acoustic model and feature extraction using linguistic features of adaptation data and then synthesized spectral features and natural features included in the adaptation data were used.

Figure 3 shows the network configurations used in the *DNN* and *MDNN* systems. This figure also shows network parts adapted by each system. In the systems *MDNN-AM-LHUC* and *MDNN-DEC-LHUC*, the hidden units of the bottleneck layer were both adapted. Three consecutive spectral amplitudes were used as the segmental input and output for the DNN-based post-filter. During the overlap-add operation using the segmental outputs of the post-filtering DNN, weighting coefficients were 0.25, 0.5, and 0.25 for previous, current and next frames respectively.

For each waveform, we extracted its frequency spectra with 2049 FFT points. For each system, 60 dimensional spectral features were extracted. Spectrum and cepstrum were both frequency-warped using the Bark scale. The feature vectors for *HMM* comprised 258 dimensions: 59 dimensional bark-cepstral coefficients (plus the 0th coefficient), log F0, 25 dimensional band aperiodicity measures, and their dynamic and acceleration coefficients. For constructing the system *DNN*, continuous log F0 interpolated linearly for unvoiced regions and voiced/unvoiced parameters were used as F0 parameters. Thus, 259 dimensional features were used as output features of the *DNN*. To construct the system *MDNN*, 2049-dim frequency-warped log spectra were used. The context-dependent labels were built using the pronunciation lexicon Combilex [29]. The linguistic features for DNN acoustic models comprised 382 dimensions. The linguistic features and spectral envelopes in the training data were pre-normalized for training DNNs. The input linguistic features were normalized to have zero-mean unit-variance, whereas the output spectral amplitudes were normalized to be within 0.0–1.0.

In this work, phoneme-level duration of test utterances was obtained using forced alignment based on HMMs because we

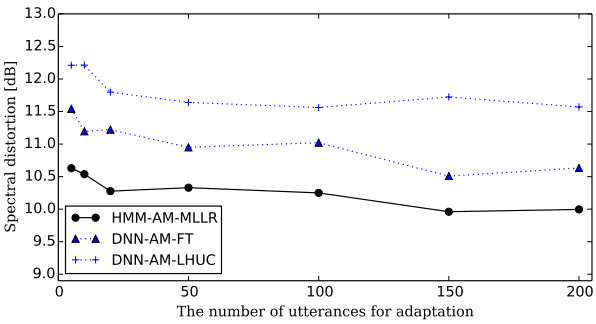


Figure 4: Spectral distortion calculated from log spectra synthesized by conventional systems (*HMM-AM-MLLR*, *DNN-AM-FT* and *DNN-AM-LHUC*). The results of *HMM* (13.472) and *DNN* (15.575) are excluded to make comparison easier.

want to focus on spectral adaptation. In the systems based on *HMM* and *DNN*, synthesized mel-cepstral vectors were converted into log spectra to calculate log spectral distortion and synthesize waveform using a STRAIGHT vocoder. Note that systems based on *MDNN* output only spectral information, so we used other features, F0 and aperiodicity measures, synthesized by systems based on *HMM* for utilizing the STRAIGHT vocoder.

For subjective evaluation, MUSHRA tests were conducted to evaluate the naturalness of synthesized speech. Natural speech was used as a hidden top anchor reference. Fifty native English speaking subjects participated in the experiments. Two sentences were randomly selected from the test set for each subject. The experiments were carried out using headphones in a soundproof room.

4.1. Experimental results

Figures 4 and 5 show objective results of each system. The results using 5, 10, 20, 50 100 150 and 200 utterances as adaptation data are shown in the figures. In all figures, *HMM-AM-MLLR* and *DNN-AM-FT* are included as conventional techniques for comparison with other systems.

4.1.1. Conventional adaptation methods of the acoustic models

It can be seen from Fig. 4, in which the results of the conventional techniques (*HMM-AM-MLLR*, *DNN-AM-FT* and *DNN-AM-LHUC*) are shown, that the result of *HMM-AM-MLLR* was better than the other techniques. We first see that the log spectral distortion of the base *DNN* model (15.575) was much larger than that of the base *HMM* (13.472). It seems that this large difference between base systems (*HMM* and *DNN*) caused a worse DNN adaptation performance. Also, the figure shows that the result of *DNN-AM-LHUC* was the worst within the three systems. This system had the smallest number of adaptation parameters, which seems to have limited its capability to transform high order spectral parameters compared with other ones.

4.1.2. Adaptation results of the multiple feed-forward DNNs

Next, Fig. 5 shows the results of the adapted systems based on multiple feed-forward DNNs. The results of *MDNN*, *HMM-AM-MLLR* and *DNN-AM-FT* are also included. First we can see from the results that the distortion of *MDNN* (12.780) was much smaller than the those of *HMM* (13.472) and *DNN* (15.575). This indicates that a more robust system was constructed based on multiple DNNs and a parameter generation with global variance or a signal processing based post-filtering used in *HMM* and *DNN* would cause the larger distortion.

Second, the effectiveness of adaptation for each part, that is, the acoustic modeling, the decoder and the post-filtering, can be seen in Fig. 5(a), 5(b) and 5(c), respectively. Compared to the results of the base system with *MDNN*, all adapted systems output closer log spectra to that of the target speaker. These results mean that speaker adaptation has been effectively performed for all the components although there were different tendencies among the results of each adaptation part.

Then, it can be seen from the Fig. 5(b) that the adaptation of the decoder part was less effective than the other parts. Moreover, using the simple fine-tuning with the larger amount of adaptation data tended to reduce the distortion as seen in the results of *MDNN-AM-FT* and *MDNN-PF-FT*, although such the improvement of the distortion was not observed in *MDNN-DEC-FT*.

Also, we can see that LHUC was effective for the speaker adaptation using the much smaller amount of adaptation data in *MDNN-AM-LHUC* and *MDNN-PF-LHUC* (excluding *MDNN-DEC-LHUC*) compared with the systems using simple additional fine-tuning. In the adapted systems based on LHUC using a larger amount of data (50, 100, 150 and 100 utterances), however, there were no improvements in any of the adaptation parts.

Third, as shown in Fig. 5(d), systems with adaptation for all parts (*MDNN-ALL-FT* and *MDNN-ALL-LHUC*) output the closest log spectra to the target speaker in all adapted systems. Similar to the results of *MDNN-AM-LHUC* and *MDNN-PF-LHUC*, *MDNN-ALL-LHUC* performed effectively for the case using the smaller amount of adaptation data.

4.1.3. Subjective evaluation results on the conventional adaptation methods of the acoustic models

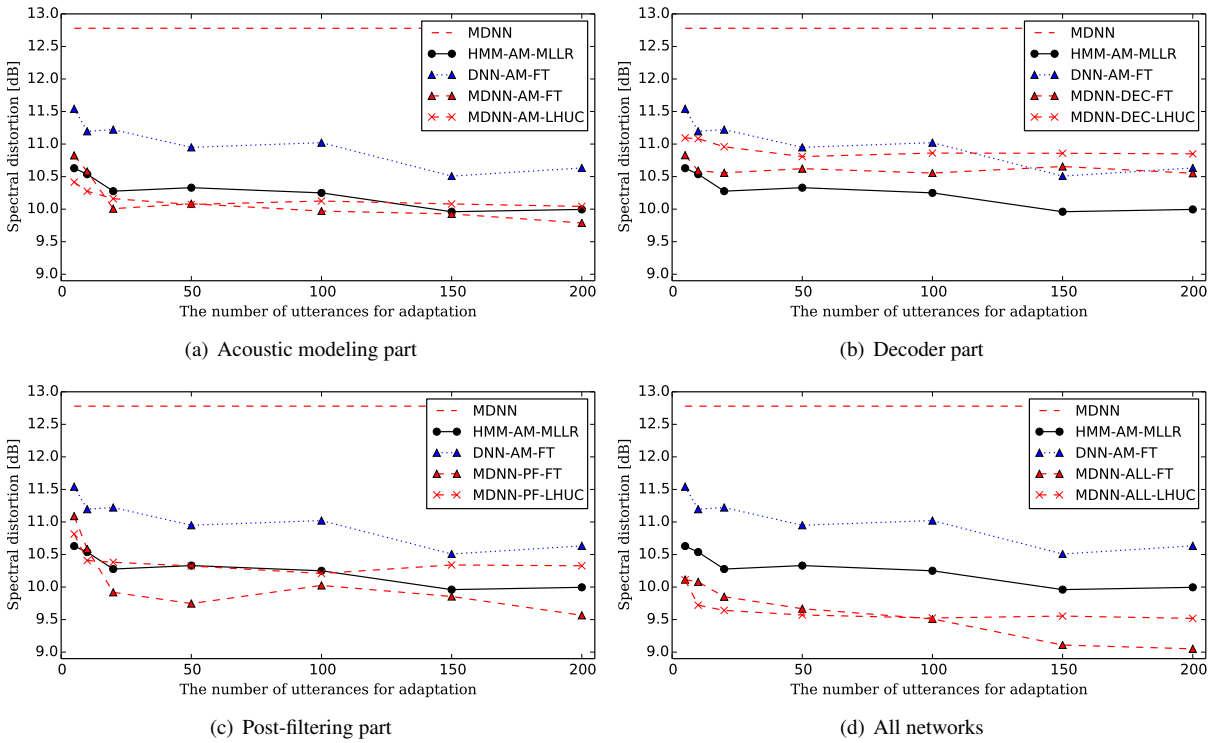
As described above, we have observed improved spectral distortion. However, from informal listening, we perceived quality degradation when we used some of the adapted systems. Therefore, we decided to carry out subjective evaluation on the naturalness of synthetic speech. As mentioned previously, we used the MUSHRA tests. For the listening tests, systems adapted with 10, 50 and 100 adaptation utterances were used.

In Fig. 6 the results on the conventional adaptation methods of the acoustic models (*HMM-AM-MLLR*, *DNN-AM-FT* and *DNN-AM-LHUC*) are shown. The systems without adaptation (*HMM*, *DNN*) are also included for references.

It can be seen from the figure that the adapted systems synthesize synthetic speech samples of a lower quality than those of systems without adaptation, unfortunately. The systems based on LHUC were rated the worst among the methods. In fact, we have found that the system adapted by LHUC output muffled voices compared to other methods. We can also see that the *DNN-AM-FT* systems outperformed *HMM-AM-MLLR* systems when the same number of adaptation utterances was used for the adaptation. Finally, as expected, we can see that the quality of synthetic speech samples gradually became better when we used a larger amount of adaptation data, excect for systems adapted by LHUC.

4.1.4. Subjective evaluation results on the adapted multiple feed-forward DNNs

Subjective evaluation results on the naturalness of the adapted multiple feed-forward DNN systems are presented in Fig. 7.



(a) Acoustic modeling part

(b) Decoder part

(c) Post-filtering part

(d) All networks

Figure 5: Spectral distortion calculated from log spectra synthesized by adapted systems based on *MDNN*. Results of the systems *MDNN*, *HMM-AM-MLLR* and *DNN-AM-FT* are also included in all figures.

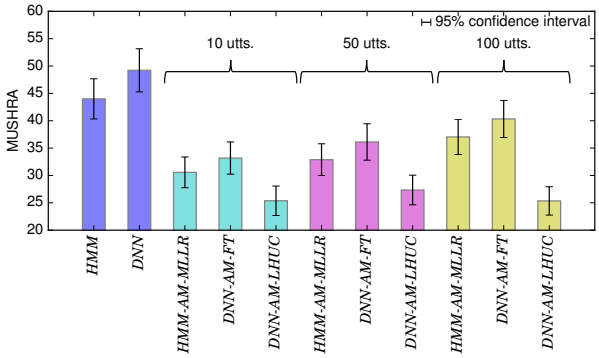


Figure 6: Subjective results of conventional systems (*HMM-AM-MLLR*, *DNN-AM-FT* and *DNN-AM-LHUC*). In the test, systems without adaptation (*HMM* and *DNN*) were also included.

The systems without adaptation (*HMM*, *MDNN*) and the HMM-based adapted system (*HMM-AM-MLLR*) are also included for references.

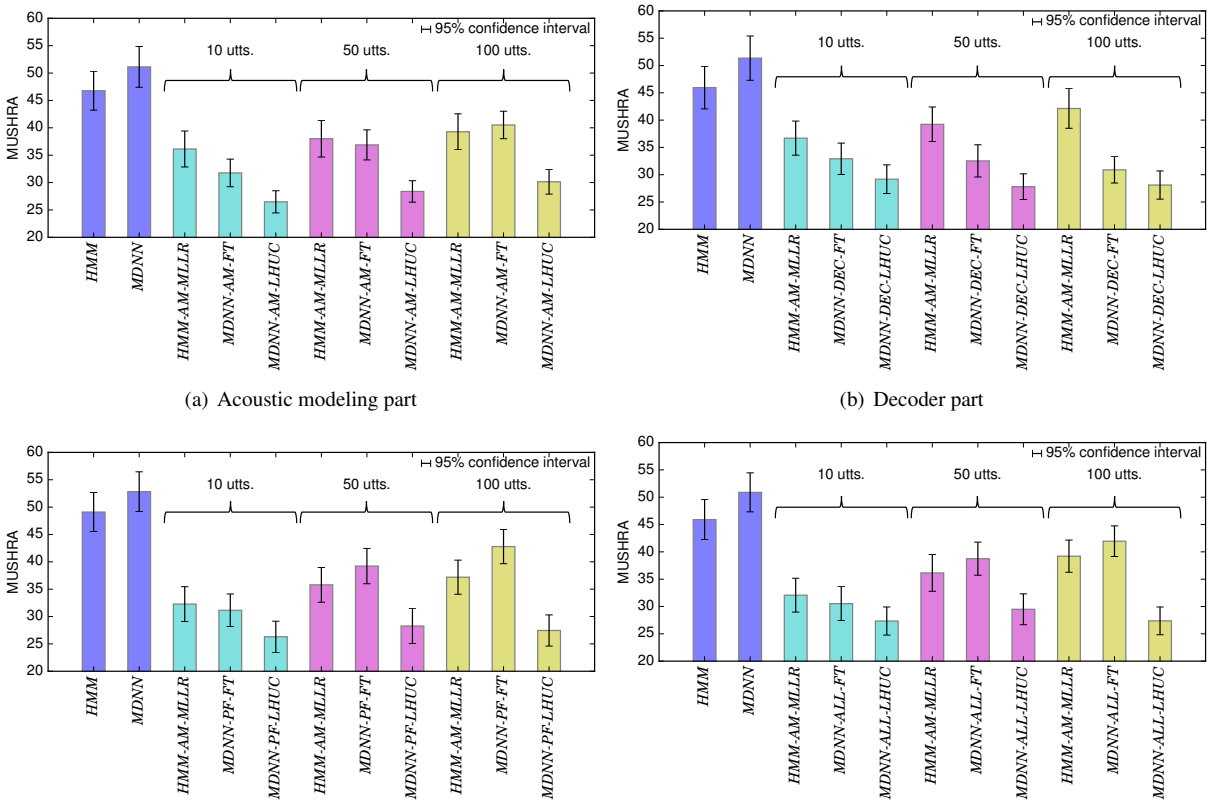
From the figure, we can first see that the adapted MDNN systems (*MDNN-AM-FT*, *MDNN-DEC-FT*, *MDNN-PF-FT* and *MDNN-ALL-FT*)) based on fine tuning were rated lower than *HMM-AM-MLLR* when the number of adaptation utterances was 10. This implies that the optimization of multiple neural networks using a very small amount of adaptation data is more difficult than that of single DNN.

However, when the number of adaptation utterances used was 50 or 100, the adaptation performance based on fine tuning depended on the modules. As seen in Fig. 7(a), the *MDNN-AM-FT* system outputs almost the same quality of synthetic speech as *HMM-AM-MLLR* for the cases using 50 or 100 utterances for the adaptation. As shown in Fig. 7(b), the performance of speaker adaptation for the decoder part was rated lower than *HMM-AM-MLLR*, similar to the objective results. In contrast, from Fig. 7(c), we can see that systems with adaptation of the post-filtering part outperformed *HMM-AM-MLLR* for the cases using the larger amount of adaptation data and the difference was statistically significant when 100 adaptation utterances were used.

Finally, Fig. 7(d) shows the results of speaker adaptation of all the networks. The results are very similar to the speaker adaptation of the post-filter models. However, the difference between the adapted multiple feed-forward DNN system *MDNN-ALL-FT* and *HMM-AM-MLLR* became statistically less significant. This simply means that the speaker adaptation of the decoder parts is less effective and therefore adapting all the networks canceled out improvements. These results lead us to conclude that it seems to be reasonable to adapt the acoustic modeling part and post filtering part for obtaining better speaker adaptation performance.

5. Conclusions

We have investigated the effectiveness of speaker adaptation for various essential components in DNN based parametric speech synthesis, including acoustic models, acoustic feature extraction, and post-filters. The objective results showed that the adaptation of each component can be effectively performed although there were different tendencies among adaptation parts.



(a) Acoustic modeling part

(b) Decoder part

(c) Post-filtering part

(d) All networks

Figure 7: Subjective results of adapted systems based on *MDNN*. In the test, systems without adaptation (*HMM* and *MDNN*) and the conventional system (*HMM-AM-MLLR*) were also included in all figures.

We also evaluated the naturalness of synthetic speech generated using the adapted multiple feed-forward DNN systems subjectively and found that speaker adaptation is effective for the post-filtering part and the system using the adapted post-filter outperformed HMM-based speech synthesis with MLLR. Further, we also found that LHUC degrades the naturalness of synthetic speech regardless of the modules used and the speaker adaptation of the decoder parts based on the fine tuning also resulted in lower quality of synthetic speech.

Our future work includes average voice model training and speaking style adaptation. Investigations as to why LHUC degrades the naturalness of synthetic speech is also our future work.

6. Acknowledgements

This work was partially supported by the Naver Corp., by EPSRC through Programme Grant EP/I031022/1 (NST) and EP/J002526/1 (CAF), by the Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency (JST) (uDialogue project), by MEXT KAKENHI Grant Numbers (26280066, 26540092, 15H01686, 15K12071, 16K16096) and by The Telecommunications Advancement Foundation Grant.

7. References

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

[2] 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.

[3] 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.

[4] 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.

[5] R. Vishnubhotla, S. Fernandez and B. Ramabhadran, "An autoencoder neural-network based low-dimensionality approach to excitation modeling for hmm-based text-to-speech," *Proceedings of ICASSP*, pp. 4614–4617, 2010.

[6] 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*, pp. 1954–1958, 2014.

[7] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," *Proceedings of ESCA/COCOSDA Third International Workshop on Speech Synthesis*, pp. 273–276, 1998.

[8] ——, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," *Proceedings of ICASSP 2001*, pp. 805–808, 2001.

[9] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[10] 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 on Audio, Speech, and Language Processing*, vol. 17, 2009.

[11] L. Saheer, J. Dines, and P. N. Garner, "Vocal tract length normalization for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2134–2148, 2012.

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

[13] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, 2000.

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

[15] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Speaker adaptation of deep neural network based on discriminant codes," *IEEE Trans. Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1713–1725, 2014.

[16] Z. Wu, P. Swietojanski, C. Veaux, R. Renals, and K. S., "A study of speaker adaptation for dnn-based speech synthesis," *Proceedings of Interspeech*, pp. 879–883, 2015.

[17] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," *Proceedings of ICASSP*, pp. 7893–7897, 2013.

[18] C. Wu and M. J. F. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," *Proceedings of ICASSP*, pp. 4315–4319, 2015.

[19] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," *Proceedings of ICASSP*, pp. 6359–6363, 2014.

[20] Y. Zhao, J. Li, J. Xue, and Y. Gong, "Investigating online low-footprint speaker adaptation using generalized linear regression and click-through data," *Proceedings of ICASSP*, pp. 4310–4314, 2015.

[21] P. Swietojanski and R. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," *Spoken Language Technology Workshop (SLT)*, pp. 171–176, 2014.

[22] S. Takaki, S.-J. Kim, J. Yamagishi, and J.-J. Kim, "Multiple feed-forward deep neural networks for statistical parametric speech synthesis," *Proceedings of Interspeech*, pp. 2242–2246, 2015.

[23] 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.

[24] P. Smolensky, "Information processing in dynamical systems: foundations of harmony theory," *Parallel distributed processing: explorations in the microstructure of cognition, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press*, vol. 1, pp. 194–281, 1986.

[25] B. Kosko, "Bidirectional associative memories," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 18, no. 1, pp. 49–60, 1988.

[26] Data and tools related to the Blizzard Challenge, http://www.cstr.ed.ac.uk/projects/blizzard/.

[27] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *Proceedings of Interspeech 2005*, pp. 2801–2804, 2005.

[28] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *IEICE*, vol. J87-D-II, no. 8, pp. 1565–1571, 2004.

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