USING PERSONALIZED SPEECH SYNTHESIS AND NEURAL LANGUAGE GENERATOR FOR RAPID SPEAKER ADAPTATION

Yan Huang, Lei He, Wenning Wei, William Gale, Jinyu Li, and Yifan Gong

Microsoft Corporation

ABSTRACT

We propose to use the personalized speech synthesis and the neural language generator to synthesize content relevant personalized speech for rapid speaker adaptation. It has two distinct aspects: First, it relieves the general data sparsity issue in rapid adaptation via making use of additional synthesized personalized speech; Second, it circumvents the obstacle of the explicit labeling error in unsupervised adaptation by converting it to pseudo-supervised adaptation. In this setup, the labeling error is implicitly rendered as less damaging speech distortion in the personalized synthesized speech. This results in significant performance breakthrough in the rapid unsupervised speaker adaptation. We apply the proposed methodology to a speaker adaptation task in a state-of-art speech transcription system. With 1 minute (min) adaptation data, our proposed approach yields 9.19 % or 5.98 % relative word error rate (WER) reduction for the supervised and the unsupervised adaptation, comparing to the negligible gain when adapting only with 1 min original speech. With 10 min adaptation data, it yields 12.53 % or 7.89 % relative WER reduction, doubling the gain of the baseline adaptation. The proposed approach is particularly suitable for unsupervised adaptation.

Index Terms— Acoustic model adaptation, speaker adaptation, speech synthesis, neural language generation

1. INTRODUCTION

The increasingly more sophisticated neural network acoustic model trained from tens of thousand hour speech is believed to be relatively robust to speaker variability. Rapid speaker adaptation for a well-trained large-scale neural network model is challenging due to the massive number of model parameters and the limited amount of adaptation data. Furthermore, the labeling error from the first-pass decoding result can lead to catastrophic gradient update. This makes effective rapid unsupervised speaker adaptation even more difficult.

There was abundant previous work in improving the robustness of neural network adaptation [1–12]. For example, the transformation [13, 14], the linear hidden unit contribution (LHUC) [15, 16], the singular value decomposition (SVD) [17], and the factorized sub-space [8] adaptation, constrain the adaptation in a reduced or highly compressed parameter space to address data sparsity. Alternatively, the Kullback-Leibler (KL) divergence regularized adaptation [3, 18] and the Bayesian adaptation [11, 12] make use of specially formulated objectives to prevent catastrophic forgetting and thus prevent overfitting. Furthermore, the i-vector [19, 20] and the speaking code [21] model adaptation utilize the speaker-level representation as the auxiliary input of a conditioning model.

In this paper, we propose to utilize the personalized speech synthesis [22–25] and the neural language generator [26, 27] to directly address the general data sparsity issue in the rapid model adaptation. We consider rapid adaptation with no more than 10 min speech in a system with millions of acoustic model parameters. Specifically, we first train a speaker embedding for the personalized speech synthesis; then use the neural language generator to generate relevant or even some random text to synthesize speech. The synthesized speech is added to the original data for adaptation. We adopt the KL-divergence regularized adaptation paradigm [3] and the subnet adaptation [28]. The configuration is designed to be consistent and rigorous. For example, in the unsupervised adaptation, only the first-pass decoding result of the original data is used in the speaker embedding training, neural language generation, and acoustic model adaptation; no human transcription is used.

There are several distinct aspects in the proposed approach. First, unlike data augmentation with noise or speaking rate perturbation [29], this approach can generate arbitrary personalized speech with no constraints on the content and data amount; therefore it fundamentally alleviates data sparsity in the rapid adaptation. Second, it implicitly converts an unsupervised adaptation to a pseudo-supervised one through the introduction of the personalized speech synthesis. The labeling error of the adaptation data is smoothed through the speaker embedding training. The rendered synthesized speech seldom exhibits perceptible mismatch with the text despite that our unsupervised adaptation is rigorously configured without using the human transcription at any step. Consequently, when consuming the synthesized speech for adaptation, catastrophic gradient update due to the explicit labeling error, a root-cause for the failure of unsupervised adaptation, is no longer a distinct obstacle.

We apply the proposed methodology to a speaker adaptation task in a state-of-art conversational speech transcription system. In the 1 min adaptation setup, our proposed approach yields 9.19 % or 5.98 % relative WER reduction for the supervised and unsupervised adaptation respectively, while adapting with the original 1 min speech only yields negligible gain. In the 10 min adaptation setup, the proposed approach roughly doubles the gain of adapting with the original 10 min speech. We further found that the content-relevant target text generated by the neural language generator outperforms the random text with small but consistent performance gain.

Our approach results in significant performance breakthrough in unsupervised adaptation, especially for rapid unsupervised adaptation. To the best of our knowledge, we are not aware of any previous work in using personalized speech synthesis and neural text generation for rapid hybrid acoustic model adaptation. In [30], speech synthesis was only used for speller training in an end-to-end system.

The rest of this paper is organized as: Section 2 introduces the methodology; Section 3 presents the experiments and results; Section 4 concludes the paper.

2. METHODOLOGY

We describe our proposed methodology of using personalized synthesis and neural language generator for rapid model adaptation.

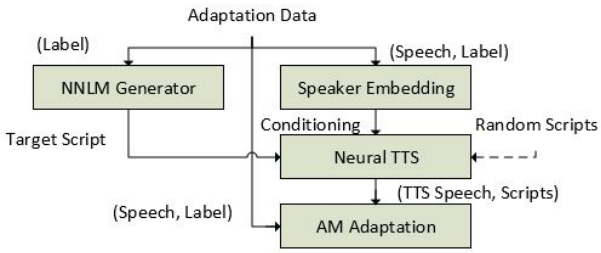

Fig. 1. System architecture of the proposed approach. Label can be human transcription or the first-pass decoding result, corresponding to the supervised and the unsupervised adaptation.

2.1. System Architecture

The proposed approach consists of personalized speech synthesis, neural language generator, and acoustic model adaption depicted in Figure 1. We first train a speaker embedding for the neural TTS to synthesize personalized speech using the adaptation data. Then we use a neural language generator to generate content relevant target text from the label of the adaptation data. Alternatively, random conversational speech text can be used. Lastly, the synthesized speech with the corresponding text is added to the original data for adaptation. In the supervised setup, we use the human transcription of the original adaptation speech to train the speaker embedding for personalized speech synthesis, to generate relevant text to be synthesized, and to supervise the model adaptation. Otherwise in the unsupervised setup, the first-pass decoding result is used. We discuss each component of the system in the rest of this section.

2.2. Acoustic Model Adaptation

We adopt the KL-divergence regularized model adaptation in this paper. The KL-divergence regularization is added to the adaptation criterion to prevent catastrophic forgetting and overfitting. The detailed formulation can be referred to in [3]. For the adaptation structure, we compared adapting different components of the original model or additional sub-space speaker-specific network components. We found that the linear projection layer sub-net adaptation yields competitive performance and thus adopt it throughout this work. The detail of the adaptation can be referred to in [28].

Data sparsity is a major barrier in rapid speaker adaptation. Adaptation with extremely small amount of data tends to result in overfitting. We studied data augmentation with noise and speaking rate perturbation and found that they are quite effective for rapid speaker adaptation [28]; nevertheless they cannot address the limited phonetic coverage. Utilizing personalized synthesized speech can flexibly generate arbitrary speech with rich phonetic coverage and even with desired relevant content.

Imperfect supervision is the fundamental challenge of unsupervised adaptation. When the first-pass decoding makes a mistake and is subsequently translated into the senone state error, this results in incorrect gradient update. The proposed methodology initiates a new perspective in how to consume the unlabeled speech for adaptation. Instead of directly using the unlabeled speech for adaptation, we use it to train a speaker embedding for personalized speech synthesis. The transcription error in the speaker embedding training is not expected to be directly translated into an explicit error in synthesized speech. Instead, it is more likely rendered as perceptible or imperceptible minor speech distortion.

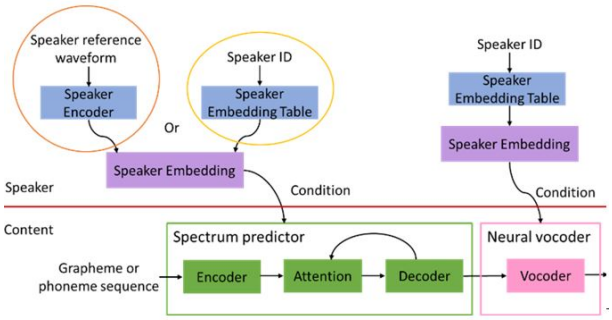

Fig. 2. Diagram of the personalized speech synthesis.

2.3. Personalized Speech Synthesis

We use a multi-speaker neural TTS system for personalized speech synthesis. As depicted in Figure 2, it consists a spectrum predictor and a neural vocoder. The spectrum predictor converts the input text into the Mel spectral. The neural vocoder generates the waveform conditioning on the Mel spectral. We use an encoder-decoder with attention model for spectrum prediction and WaveNet as the neural vocoder [22]. Speaker embedding is introduced to pool multi-speaker data during training so that we can efficiently create personalized speech for new speakers [23]. We use an in-house TTS corpus with around 30 professional en-US speakers and more than 200 hours phonetic-rich recordings for model training. The spectrum predictor is adapted to each target speaker as the speaker embedding given some registration data. The vocoder is a universal WaveNet trained with the same corpus without adaptation. The detail of personalized speech synthesis can be referred to in [23].

One particular challenge is that, in unsupervised adaptation, only the first-pass decoding result is available. The imperfect transcription may affect the speaker embedding training. Furthermore, rapid adaptation with as little as 1 min speech also makes robust speaker embedding estimation more difficult. The quality of the synthesized speech, according to our listening test, is quite robust even with only 1 min data and imperfect transcription. Lastly, although some data selection based on the quality of the synthesized speech may further help improve the adaptation performance, we didn't apply any data filtering throughout this paper.

2.4. Neural Language Generator

We use an LSTM language model [26] with a beam search algorithm [27] to generate content relevant target text. Specifically, each sentence is provided as a prompt to the neural language generator to generate various continuations of the prompts. Similar to [27], we impose diversity constraints during the beam search, namely by penalizing repeated tokens, restricting the number of beams that end with the same bigram, and preventing n-gram repetitions within a beam. The language model has a vocabulary size of 59K byte pair encoding (BPE) tokens [31] and three LSTM layers, with a total of 220M parameters. We trained the language model to convergence on 3B words of paragraph level web-crawled data.

3. EXPERIMENTS AND RESULTS

In this section, we present results on using personalized synthesized speech for model adaptation in a presentation transcription system. All our experiments were conducted on anonymized data with personally identifiable information removed.

3.1. Experimental Setup

The baseline is a bi-directional LSTM trained from tens of thousand hour speech. It has six bi-directional LSTM layers followed by a fully connected top layer. Each layer has 2048 hidden units. The input is 80-dim log-filter bank feature. The output layer has 10K senone states. The speaker adaptation task consists of six speakers (three native and three non-native speakers), each with 10 min for training and 20 min for testing.

The quality of synthesized personalized speech depends upon the original data amount and the label quality. The relevancy of the text script used for speech synthesis may as well affect the adaptation performance. We therefore configured eight setups, specified by the amount of the original adaptation data (e.g. 1 or 10 min), the label type (e.g. human transcription or the first-pass decoding result), and the text script type (e.g. random text or target text), summarized in Table 1. The acoustic model adaptation uses an identical setup as the acoustic model adaptation. This consistent setup allows rigorous study on the impact of adaptation with synthesized speech.

Table 1. Configuration of the personalized speech synthesis, specified by the original data amount, the label type, and the text type.

	1 min	10 min
SUP	[1, Human, Random]	[10 Human, Random]
SUP	[1, Human, Target]	[10, Human, Target]
UNSUP	[1, ASR, Random]	[10, ASR, Randrom]
UNSUP	[1, ASR, Target]	[10, ASR, Target]

3.2. Baseline Adaptation Result

We adopt the KL-regularized sub-net adaptation. The baseline adaptation performance is summarized in Table 2. We use the relative WER reduction (WER.R) to measure the adaptation performance throughout this paper. With 1 min adaption data, neither supervised nor unsupervised adaptation yields noticeable gain due to insufficient data. As the adaptation data amount increases to 10 min, the supervised and the unsupervised adaptation yield 7.60 % and 3.61 % relative WER reduction respectively. The unsupervised adaptation only achieves half of the gain of the supervised adaptation due to the imperfect supervision.

Table 2. Baseline sup/unsup adaptation performance with 1 min or 10 min data. WER.R referes to the relative WER reduction.

Model	1 min	**WER.R**	10 min	**WER.R**
Baseline	14.65	**NA**	14.65	**NA**
SUP	14.53	**0.82**	13.54	**7.60**
UNSUP	14.52	**0.88**	14.13	**3.61**

3.3. Adaptation with Synthesized Speech

Table 3 presents the adaptation performance of combining the synthesized personalized speech with the original speech. We will focus on setups with target text (tar) here, while leaving its comparison with the random text to Section 3.6.

In the 1 min adaption setup, the supervised and the unsupervised adaptation with additional 100 min synthesized speech yield 6.72 % and 5.70 % relative WER reduction respectively. Both significantly outperform the baseline adaptation with only the original data. With the personalized synthesized speech, adaptation with 1 min speech becomes beneficial even in this well-trained large-scale system. For the 10 min adaptation, the baseline sup/unsup adaption yield 7.60 % and 3.61 % relative WER reduction; after adding the synthesized speech, the corresponding gain increases to 9.35 % and 6.82 %.

The benefit of using the synthesized speech for unsupervised adaptation is clear. For example, in the 1 min setup, when the additional synthesized speech is used, the unsupervised adaptation achieves comparable performance as the supervised adaptation. In the 10 min setup, the gap of supervised and unsupervised adaptation shrinks after adding synthesized speech. This confirms our hypothesis that the proposed approach can implicitly smooth and disperse the impact of labeling error, which essentially converts an unsupervised adaptation to a pseudo-supervised one. With this property, the 1 min unsupervised adaption, previously exhibiting as an extremely challenging problem, becomes feasible and practically effective.

We further apply some weighting to the original speech to balance the relative amount of the original speech and the synthesized speech. The original speech is weighted by 10 or 5 for the 1 min and 10 min setup respectively. The relative WER reduction increases from 6.72 % to 9.19 % and 9.35 % to 12.53 % for 1 min and 10 min supervised adaptation, respectively. This suggests that it is important to maintain reasonable representatives of the original speech, especially given we can theoretically generate unlimited synthesized speech. Moreover, it is interesting to observe that the data weighting is helpful but with smaller impact on the unsupervised adaptation side. The original data with imperfect label in unsupervised adaptation is not as valuable as the human transcription in supervised adaptation. This, on the other hand, again confirms the favorable utilization of synthesized speech in unsupervised adaptation.

Table 3. Performance of the sup/unsup speaker adaptation with 100 min synthesized speech. When annotated with (w), the original data is weighted by 10 or 5 for the 1 min and the 10 min adaptation.

Model	1min	**WER.R**	10min	**WER.R**
Baseline	14.65	**NA**	14.65	**NA**
SUP	14.53	**0.82**	13.54	**7.60**
SUP$_{+tar(100)}$	13.67	**6.72**	13.28	**9.35**
SUP$_{+tar(100)}(w)$	13.31	**9.19**	12.82	**12.53**
SUP$_{+rand(100)}$	13.91	**5.10**	13.34	**8.94**
SUP$_{+rand(100)}(w)$	13.58	**7.32**	12.91	**11.87**
UNSUP	14.52	**0.88**	14.13	**3.61**
UNSUP$_{+tar(100)}$	13.82	**5.70**	13.65	**6.82**
UNSUP$_{+tar(100)}(w)$	13.82	**5.98**	13.50	**7.89**
UNSUP$_{+rand(100)}$	13.92	**5.04**	13.68	**6.63**
UNSUP$_{+rand(100)}(w)$	13.81	**5.79**	13.53	**7.70**

3.4. Random Text versus Target Text

It is generally believed that it is beneficial to use relevant speech for adaption in an end-to-end ASR system. In the hybrid model, the advantage of relevant content is not obvious. It may have a better matched senone coverage and thus can possibly help the adaptation. As speech synthesis has the flexibility to render an arbitrary text, we would like to find out whether content relevancy is important for the hybrid model adaptation. We experiment with adaptation using synthesized random conversational speech text. The pair-wise comparison with the content relevant target text counterpart is presented in Table 3. We can see that adaptation with the target text consistently outperforms the random text with around 1 % relative WER reduction. Despite of the relatively small additional performance gain, the consistent gain across all setups suggests that in the hybrid system, adaptation with content relevant data is also beneficial.

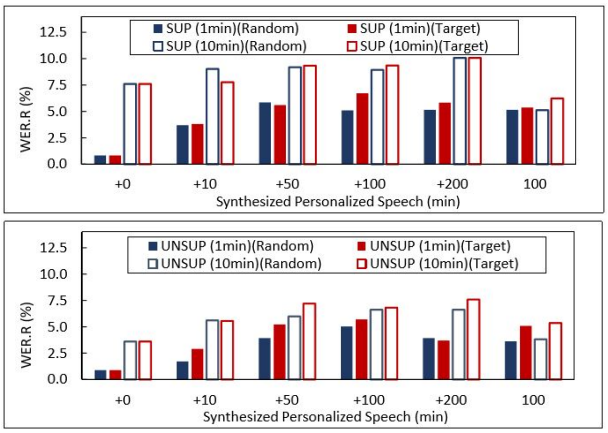

Fig. 3. Adaptation performance with added 10, 50, 100, 200 min synthesized speech (+) or only with 100 min synthesized speech.

3.5. Adaptation with Different Data Amount

Figure 3 illustrates the adaptation performance with different amount of added synthesized speech or synthesis speech alone. No weighting for the original data is applied, which would otherwise yield larger gain when large amount of synthesized speech is added as discussed earlier. Both supervised and unsupervised adaptation improves with more synthesized data. The performance gain has not plateaued even with 200 min synthesized speech. The improvement slope is steeper for the 1 min setup as the proposed approach is more beneficial for very rapid adaptation. Unsupervised adaptation also exhibits steeper performance gain as the proposed approach circumvents the direct labeling error in unsupervised adaption.

We further discard the original speech and only use the 100 min synthesized speech for adaptation. This reflects how the personalized speech synthesis can distill the speaker trait from small amount of original speech and render it robustly. In the 1 min setup, speaker adaptation with only synthesized speech performs significantly better than adaptation with the original data. In the 10 min unsupervised setup, adaptation with synthesized speech also outperforms or performs as well as adaptation with the original data. The only exception is the 10 min supervised adaptation setup, where adaptation with the original speech wins with a small margin. This suggests that the proposed approach can effectively distill speaker trait and render it with additional phonetic and phonological knowledge embedded in the synthesis model and language knowledge embedded in the neural language generator. In all cases, adding original speech is still helpful, especially in the 10 min setup and in the supervised case. This indicates that the original speech is still valuable, especially when with reasonable amount and with the human label.

3.6. Analysis of Speaker Embedding Robustness

Acoustic model adaptation is highly sensitive to speech data amount and the label quality. This is clearly illustrated in Section 3.2. How would these affect the speaker embedding training and the consequent speaker adaptation performance which consumes the synthesized speech? In adaptations with synthesized speech only, as the original speech is not used in adaptation, the sup/unsup adaptation only differ in how the speaker embedding is trained. Consequently, we can measure the impact of the label quality and the data amount on the speaker embedding by comparing speaker adaptation performance with synthesized speech on different setups. To facilitate this comparison, we re-organize the relevant results in Table 4.

Adaptation with synthesized speech based on speaker embedding trained from 10 min consistently outperforms from 1 min and the performance gap is small. This suggests that the speaker embedding is robust even when small amount of data is available. Moreover, as the original speech is not used in adaptation, the supervised and unsupervised adaptation here only differ in whether speaker embedding is trained from human transcription or the first-pass decoding result. For the random text setup, the supervised adaptation is noticeably better than the unsupervised counterpart both in the 1 min and the 10 min setup. This suggests that the imperfect transcription affect the speaker embedding training and consequently degrades the synthesized speech quality. We listened to the synthesized speech in different setups carefully and found that the transcription error is rendered as perceptible or imperceptible minor sound distortion, instead of being translated into distinct errors in synthesized speech.

Table 4. Performance of sup/unsup adaptation with 100 min synthesized speech. The synthesized speech is generated from the speaker embedding trained from 1 or 10 min in the sup/unsup setup.

Model	1min	**WER.R**	10min	**WER.R**
Baseline	14.65	**NA**	14.65	**NA**
SUP	14.53	**0.82**	13.54	**7.60**
UNSUP	14.52	**0.88**	14.13	**3.61**
$\text{SUP}_{tar(100)}$	13.87	**5.36**	13.74	**6.23**
$\text{UNSUP}_{tar(100)}$	13.90	**5.10**	13.87	**5.36**
$\text{SUP}_{rand(100)}$	13.90	**5.15**	13.91	**5.11**
$\text{UNSUP}_{rand(100)}$	14.12	**3.63**	14.09	**3.82**

3.7. Others

We compared the proposed methodology with noise and speaking rate perturbation based data augmentation. All three are beneficial, but the proposed methodology performs notably better. They are also found to be complimentary. We analyzed the performance for the native and the non-native speakers. The gain for non-native speakers is more significant. We think this is because the baseline is a strong native model.

4. CONCLUSION

In summary, we presented using the personalized speech synthesis and the neural language generator to synthesize personalized speech for rapid speaker adaptation. This approach not only alleviates the data sparsity in rapid speaker adaptation, but also circumvents the obstacle of the explicit labeling error in unsupervised adaptation. This makes it particularly suitable for unsupervised adaptation. The proposed approach yields 9.19 % or 5.98 % relative WER reduction for sup/unsup adaptation while the baseline 1 min adaptation only yields negligible gain. In the 10 min adaptation setup, it roughly doubles the gain of the baseline adaptation.

Future work includes applying this approach to the far-field speech recognition applications which would require the personalized speech synthesis to be robust to channel and environmental noise. We are also considering ways to improve the run-time efficiency of synthesis. Lastly, we plan to apply it to an end-to-end ASR system and expect similarly interesting results with some nuisance.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge Dr. Liping Chen for the helpful discussion.

6. REFERENCES

[1] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proceedings of SLT*, 2012.

[2] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proceedings of ICASSP*, 2013.

[3] D. Yu, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proceedings of ICASSP*, 2013.

[4] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proceedings of Interspeech*, 2014.

[5] P. Swietojanski and S. Renals, "Differentiable pooling for unsupervised speaker adaptation," in *Proceedings of ICASSP*, 2015.

[6] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," in *Proceedings of ICASSP*, 2016.

[7] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, p. 459–468, 2016.

[8] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," in *Proceedings of ICASSP*, 2014.

[9] Y. Q. Wang, C. Zhang, M.J.F. Gales, and P.C. Woodland, "Speaker adaptation and adaptive training for jointly optimized tandem system," in *Proceedings of Interspeech*, 2018.

[10] M. Delcroix, S. Watanabe, A. Ogawa, S. Karita, and T. Nakatani, "Auxiliary feature based adaptation of end-to-end ASR systems," in *Proceedings of Interspeech*, 2018.

[11] Z. Huang, S. M. Siniscalchi, I. F. Chen, J. Li, J. Wu, and C. H. Lee, "Maximum a posteriori adaptation of network parameters in deep models," in *Proceedings of Interspeech*, 2015.

[12] X. Xie, X. Liu, T. Lee, S. Hu, and L. Wang, "BLHUC: Baysian learning of hidden unit contributions for deep neural network adaptation," in *Proceedings of ICASSP*, 2019.

[13] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker adaptation for hybrid HMM/ANN continuous speech recognition system," in *Proceedings of EUROSPEECH*, 1995.

[14] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Communication*, pp. 827–835, 2007.

[15] P. Swietojanski, J. Li, and S. Renal, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transaction on Audio Speech Language Processing*, pp. 1450–1463, 2016.

[16] S. M. Siniscalchi, J. Li, and C. H. Lee, "Hermitian based hidden activation functions for adaptation of hybrid HMM/ANN models," in *Proceedings of Interspeech*, 2012.

[17] 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," in *Proceedings of ICASSP*, 2014.

[18] Y. Huang and Y. Gong, "Regularized sequence-level deep neural network model adaptation," in *Proceedings of Interspeech*, 2015.

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

[20] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proceedings of ICASSP*, 2014.

[21] 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 *Proceedings of ICASSP*, 2013.

[22] J. Shen, P. Pang, R. J. Weiss, M. Schuster, M. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *Proceedings of ICASSP*, 2018.

[23] Y. Deng, L. He, and F. K. Song, "Modeling multi-speaker latent space to improve neural TTS: quick enrolling new speaker and enhancing premium voice," in *arXiv preprint arXiv:1812.05253*, 2016.

[24] A. Van Den Oord, D. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *arXiv preprint arXiv:1609.03499h*, 2016.

[25] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Y. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: towards end-to-end speech synthesis," in *Proceedings of Interspeech*, 2018.

[26] M. Sundermeyer, R. Schluter, and H. Ney, "LSTM neural networks for language modeling," in *Proceedings of Interspeech*, 2010.

[27] K. Vijayakumar, A, M. Cogswell, R. S. Ramprasath, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beamsearch: Decoding diverse solutions from neural sequence models," in *arXiv preprint arXiv:1610.02424*, 2016.

[28] Y. Huang and Y. Gong, "Rapid speaker adaptation for meetings," in *Proceedings of ICASSP*, 2020.

[29] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transaction on Audio Speech Language Processing*, pp. 1469 – 1477, 2015.

[30] J. Guo, T. N. Sainath, and R. J. Weiss, "A spelling correction model for end-to-end speech recognition," in *Proceedings of ICASSP*, 2019.

[31] R. Sennrich, B. Haddow, and B. Alexandra, "Neural machine translation of rare words with subword units," in *arXiv preprint arXiv:1508.07909*, 2015.