PHONEME DEPENDENT SPEAKER EMBEDDING AND MODEL FACTORIZATION FOR MULTI-SPEAKER SPEECH SYNTHESIS AND ADAPTATION

Ruibo Fu [1, 2], Jianhua Tao [1,2,3], Zhengqi Wen [1], Yibin Zheng [1, 2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{ruibo.fu, jhtao, zqwen, yibin.zheng}@nlpr.ia.ac.cn

ABSTRACT

This paper presents an architecture to perform speaker adaption in long short-term memory (LSTM) based Mandarin statistical parametric speech synthesis system. Compared with the conventional methods that focused on using fixed global speaker representations in utterance level for speaker recognition task, the proposed method extracts speaker representations in utterance and phoneme level, which can describe more pronunciation characteristics in phoneme level. And an attention mechanism is deployed to combine each level representations dynamically to train a task-specific phoneme dependent speaker embedding. To handle the unbalanced database and avoid over-fitting, the model is factored into an average model and an adaptation model and combined by an attention mechanism. We investigate the performance of speaker representations extracted by different methods. Experimental results confirm the adaptability of our proposed speaker embedding and model factorization structure. And listening tests demonstrate that our proposed method can achieve better adaptation performance than baselines in terms of naturalness and speaker similarity.

Index Terms— speech synthesis, speaker adaptation, speaker embedding, phoneme representation

1. INTRODUCTION

Statistical parametric speech synthesis (SPSS) [1] is more robust and apt to control the speaking style of generated speech using a small database at current stage, compared with unit selection methods [2] and end-to-end methods [3,4]. A lot of improvements in SPSS researches were based on a large mono speaker recording corpus [5-10]. However, training a high-quality acoustic model with limited database and synthesizing a new voice of speaker is still a hot topic.

Generally, speaker adaptive training methods boil down to two aspects. One aspect is the speaker representations. In [11,12], i-vectors had shown improvements in the quality of synthetic speech and controllability of speaker style. Learning hidden unit contributions [13] and conversion of predicted features using Gaussian mixture models as different speaker representations were compared in [11]. In [14], d-vectors was applied for adaptive speech synthesis. It reported that d-vectors based approach achieved better speech quality than i-vectors based approach. The above methods used fixed global speaker representations for speaker recognition task. It is not optimal for the multi-speaker speech synthesis task. On this count, the speaker representations were directly extracted from raw waveform in [15]. The speaker representations, that were trained for speech synthesis task, improved the similarity of synthetic speech. The speaker representations used in the speaker recognition task only emphasize the differences between two persons. While in the speech synthesis task, we need more subtle control on each phoneme of speech and a speaker embedding that is optimal for the task.

Another aspect for speaker adaptive training is acoustic model structure. Mono speaker output layers were used in some researches [11,12,14]. The performance of speaker representations forwarded in different positions of network were investigated in [11]. These mono output layers structure only relied on speaker representations to guide various style generations and tended to generate over-average speech. To achieve adaptation in various network for each style of speaker, most of researches applied speaker dependent layers [16,17] and speaker and language factorization methods [18-20]. But it may occur over-fitting phenomenon when the adaptation database is small. Therefore, the network structure for multi-speaker acoustic model need to have adaptation function in separate physical areas of network and be able to handle the unbalanced database.

In this paper, we propose a speaker and phoneme dependent representations extraction procedure for subtle control of synthetic speech. Experiments are conducted to investigate the effects of the speaker representations extracted by different methods in aspects of corpus and algorithm. And the phoneme dependent speaker embedding combined the utterance and phoneme level speaker representations by an attention mechanism and is trained to optimum with acoustic model jointly. In the aspect of model structure, we design a network factorization structure with each designed function in separate areas to avoid over-average and over-fitting. The average model and adaption model have its own physical network and are combined by an attention mechanism to handle the unbalanced data problem.

The rest of the paper is organized as follows. Section 2 describes our proposed system architecture. Section 3 presents the experiments. And the results and analysis are presented in Section 4. The conclusions and future work are discussed in Section 5.

2. METHOD

Fig. 1 shows the architecture of acoustic model for LSTM based Mandarin SPSS system. The proposed architecture has four components including mean column, adaptation column, shared hidden layer and speaker dependent output layer.

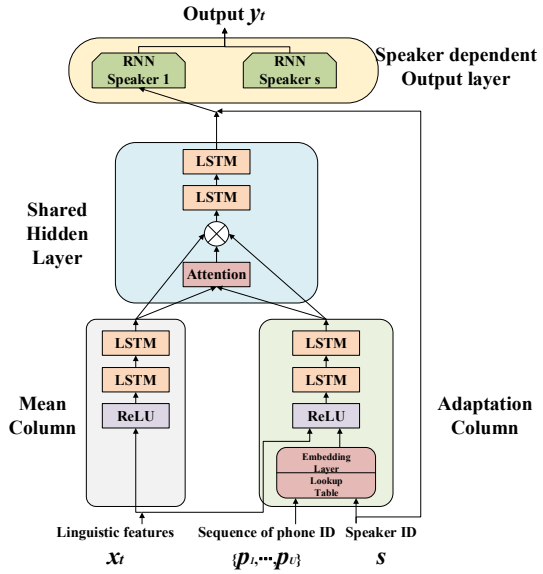

Fig. 1 Architecture for the LSTM-based multi-speaker SPSS system. It consists of a mean column, an adaptation column, a shared hidden layer and a speaker dependent RNN output layers.

2.1. Framework

There are three types of input features: speaker ID s, sequence of frame-level linguistic feature vectors $\{x_1, \dots, x_T\}$ and sequence of phone id $\{p_1, \dots, p_U\}$ for each utterance. All of them are sent into the adaptation column. The adaptation column is designed for adaptative training, which contains the embedding layer. There is a look up table procedure in the embedding layer. The dictionaries for look up table procedure is the global utterance level and local phoneme level speaker representations that we extracted respectively before the training procedure. In the training procedure, utterance and phoneme level speaker representations are combined by an attention mechanism to generate phoneme dependent speaker embedding. This structure allows the model to decide the portion of usage from each level of representations. The joint training procedure ensure the speaker embedding is optimal. Besides, linguistic features are also sent into mean column which captures the shared knowledge across different speakers. The outputs of mean column Ω^{MEAN} and adaptation column Ω^{ADA} are sent into the hidden shared layer and combined by an attention mechanism. Finally, the output of hidden shared layer Ω^{HSL} are sent into speaker dependent RNN output layer for each training speaker $s \in \{1, \dots, S\}$, which is then sent to a vocoder [21] to synthesize speech.

2.2. Speaker representations extraction

The speaker representations are extracted before the training and used as the dictionaries for the embedding layer in the adaptation column. We investigate the speaker representations in two aspects. One aspect is the corpus, which can be divided into the original utterances, concatenated utterances with same phoneme and concatenated utterances with only voiced frames. Another aspect is the algorithm, which include the d-vector method [22] and the i-vector method [23].

The common unit for speaker representation extraction is an utterance of speech containing different phonemes. It is a global utterance level speaker representation. To obtain a more delicate phoneme level speaker representation, we seg the wave based on the phoneme forced alignment information by HTS [24]. And we concatenated the segs of wave that contain the same phoneme of specific speaker together and extract the speaker and phoneme dependent speaker representations. Besides, there are two types of phoneme (vowel and consonant) for Mandarin. Further corpus processing is done to reserve only voiced frames in the concatenated utterances with same vowel, which is called concatenated utterances with only voiced frames.

2.3. Phoneme dependent speaker embedding

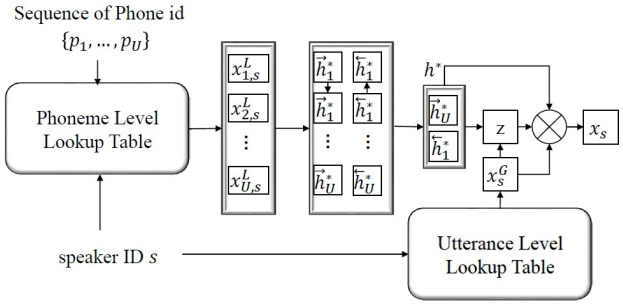

Fig. 2 Combining phoneme level speaker representations with utterance level speaker representations using an attention mechanism.

The previous methods [11,12,14] directly took speaker representations (i-vector, d-vector) as speaker embedding, which wouldn't be updated during acoustic model training. In this paper, an embedding layer is added in the adaptation column to induce task-specific speaker representation. As shown in the Fig.2, an input sequence of phone id $\{p_1, \dots, p_U\}$ of U phones and speaker ID s are transformed by the first embedding layer into a sequence of local phoneme level speaker representations $\{x^L_{1,s}, \dots, x^L_{U,s}\}$ and global utterance level speaker representations x^G_s, by applying the lookup table operation. Then $\{x^L_{1,s}, \dots, x^L_{U,s}\}$ are encoded by a BLSTM. The last hidden state from both directions are then concatenated to form an alternative representation h^*.The attention mechanism can adaptively control the balance between utterance level x^G_s and phoneme level h^* speaker representations. x^G_s and h^* are added together using a weighted sum, where the weights are predicted by a two-layer network:

$$\mathrm{z} = \sigma\left(W_z^{(3)} \tanh\left(W_z^{(1)} x_s^G + W_z^{(2)} h^*\right)\right) \quad (1)$$

$$x_s = \mathrm{z} \cdot x_s^G + (1 - z) \cdot h^* \quad (2)$$

where $W_z^{(1)}$, $W_z^{(2)}$ and $W_z^{(3)}$ are weight matrices for calculating z, $\sigma()$ is the logistic function, and x_s is the phoneme dependent speaker embedding.

2.4. Shared hidden & speaker dependent output layers

As shown in the Fig.1, instead of simply concatenating the outputs of mean column Ω^{MEAN} and adaptation column Ω^{ADA} , we found that using an attention mechanism that weights the portion of two columns can improve the system performance more significantly when the adaptation data is small. Therefore, the outputs of mean column Ω^{MEAN} and adaptation column Ω^{ADA} are combined by a dynamic weighting mechanism that has been introduced in section 2.3. Then the combining vector are forwarded into two layers of LSTM. Each speaker has its own output layer for adaptation. The vocoder parameters generated for speaker s, y_t^s, can be derived as follows:

$$h_t = \Omega^{HSL}\left(Attention(\Omega^{MEAN}, \Omega^{ADA})\right) \quad (3)$$

$$y_t^s = \Omega_s^{SDOR}(h_t, y_{t-1}^s) \quad (4)$$

3. EXPERIMENTAL SETUP

The composition of Mandarin database is shown in the Table 1. All the wav files are sampled at 44.1KHz. The 60-dim line spectral pairs (LSP) features, 1-dim band aperiodicity (BAP) feature, 1-dim logarithmic fundamental frequency (log F0) together with their delta and delta-delta deviation, and voiced/unvoiced (V/UV) flag are extracted with frame shift 5-ms, and frame length 25-ms using WORLD [21]. The input features are the encoded 379 dimensional one-hot and numerical linguistic features.

The d-vector neural network had 3 hidden layers. The bottle-neck layer had 36 neurons while the hidden layers had 1024 neurons each. The soft-max output layer had 8 neurons corresponding to the number of training speakers. The network was trained using cross-entropy training criterion and convergence was achieved after 25 epochs. The i-vector system was trained using MFCC features, where the UBM had 512 Gaussian mixtures. Both d-vector neural network and i-vector systems were trained using the KALDI toolkit [25] and had 100 dimensions representations for each speaker. All the BLSTM-related architectures have two hidden layers; each layer contains 160 memory blocks in each direction. Parameters of the proposed models are optimized using AdaDelta [26] with learning rate 0.001. Our implementation is in TensorFlow [27] and the dropout rate [28] is set at 0.5.

Table 1: Composition of Mandarin database (M=Male; F= Female)

Sentence number		Training	Validation	Test	Speaker
Training	Large set	9,000×3	500×3	500×3	1 M 2 F
	Small set	900×3	50×3	50×3	1 M 2 F
	Total	29,700	1650	1650	2 M 4 F
Adaptation		200×2 50×2	10×2	20×2	1 M 1 F

4. EVALUATION AND DISCUSSION

Objective measures used in this paper are Mel-cepstral Distortion (MCD) [29], F0 distortion in the root mean squared error (RMSE), BAP distortion and voiced/unvoiced (V/UV) swapping errors. As for subjective evaluation, we conducted MOS tests and AB preference tests to evaluate the naturalness and similarity. 20 listeners participated the evaluation. In each experimental group, 20 parallel sentences are selected randomly from testing sets of each system.

In this section, OS stands for individual modeling is trained with only one target speaker data. SI represents the speaker independent model which is trained by multiple speakers data but without speaker identity information. Two types of systems described in [12] and [14] are built and marked as I-Base and D-Base, which use i-vector and d-vector as speaker representations input respectively. These two systems both contain four LSTM layers with 256 memory blocks. The above systems are four baselines. P means our proposed method, where G and L stand for the global utterance level speaker representations and local phoneme level speaker representations respectively. And I and D mean i-vector and d-vector methods. “all” and “voiced” repents the concatenated utterances with all the frames and only voiced frames as described in section 2.2.

4.1. Evaluations for multi-speaker speech synthesis

Table 2 shows the average objective evaluation results over female and male speakers respectively. Firstly, by observing the results of baselines, the speaker representations can improve the accuracy of the acoustic model in the objective measures and d-vector based approaches perform better than i-vector based approaches. Secondly, compared the proposed methods with baselines, there are about 20% accuracy improvements in all the objective measures.

Thirdly, by comparing the P-G* with P-G*-L*-*, the P-GD-LI-voiced system achieve the best objective evaluations results. We can draw the conclusion that adding the local phoneme level speaker representations can improve the accuracy. This demonstrate the effectiveness of proposed speaking embedding that combine the global utterance level and local phoneme level speaker representations. And the global speaker representations extracted by d-vector methods is better, while the local phoneme speaker representations extracted by the i-vector methods is better. A possible explanation is that the data size considering different phonemes of speakers may be small for DNN training in the d-vector extraction processing because there are 47 types of phonemes in Mandarin. Besides, using corpus with only voiced frames can further improve the performance of speaker phoneme representations by the i-vector method. We suppose that deleting the unvoiced frames makes the variance of parameters become small in the i-vector extraction method. This procedure reduces the noise. Fourthly, the female corpus achieves overall better performance the male corpus because there is larger corpus.

Subjective evaluation results are showed in Fig.3. Among all the system, P-GD-LI-voiced achieves the best preference. We can observe that the introduction of speaker and phoneme dependent representations can improve the similarity performance more than the naturalness.

Table 2: Objective evaluations for multi-speaker speech synthesis.

Speaker	Male				Female			
Systems	MCD (dB)	F_0 RMSE(Hz)	BAP (dB)	V/UV Err (%)	MCD (dB)	F_0 RMSE (Hz)	BAP (dB)	V/UV Err (%)
OS	6.84	30.74	2.38	5.35	6.65	27.32	2.34	4.82
SI	7.14	33.65	2.13	5.27	6.94	31.14	2.16	4.87
I-Base	6.74	25.83	2.11	4.93	6.53	24.62	2.08	4.58
D-Base	6.57	24.64	2.07	4.86	6.35	24.15	2.03	4.53
P-GI	6.38	21.93	2.03	4.61	6.41	20.41	2.06	4.49
P-GD	6.35	21.67	2.26	4.52	6.20	20.37	2.14	4.46
P-GI-LI-all	5.98	20.34	2.16	4.38	5.78	17.58	1.94	4.25
P-GI-LI-voiced	5.94	20.36	2.03	4.29	5.73	17.48	1.99	4.22
P-GI-LD-voiced	5.83	20.51	1.97	4.23	5.83	17.92	1.93	4.15
P-GD-LD-voiced	5.82	20.32	1.99	4.23	5.74	17.73	1.95	4.15
P-GD-LI-all	5.77	19.99	**1.88**	4.21	5.64	16.93	1.87	4.16
P-GD-LI-voiced	**5.53**	**19.85**	1.89	**4.16**	**5.49**	**16.85**	**1.87**	**4.13**

Table 3: Objective evaluations for new speaker adaptation. No attention is trained without attention mechanism in the shared hidden layer.

Sentence number	200 sentences				50 sentences			
Systems	MCD (dB)	F_0 RMSE(Hz)	BAP (dB)	V/UV Err (%)	MCD (dB)	F_0 RMSE (Hz)	BAP (dB)	V/UV Err (%)
OS	8.36	30.63	2.57	16.75	9.18	32.56	2.89	23.82
SI	7.57	28.87	2.32	8.75	8.94	30.25	2.67	10.54
I-Base	6.93	27.94	2.24	5.35	7.26	28.59	2.39	5.86
D-Base	6.68	25.38	2.13	4.97	6.97	27.78	2.46	5.76
P-GD-LI-voiced	**5.68**	**20.37**	**1.96**	**4.35**	**5.96**	**21.35**	**2.15**	**4.68**
P-GD-LI-voiced(no attention)	5.85	20.86	1.98	4.58	6.76	24.65	2.35	4.96

4.2. Evaluations for speaker adaptation

The average objective evaluation results of the new speaker adaptation are presented in Table 3. The OS systems are trained with the same adaptation data of the new speaker with random initialization. According to Table 3, distortions of adapted speeches are much lower than the distortion of OS, which suggests the importance of model initialization. For speaker adaptation, the neural network is updated based on a well-trained multi-speaker acoustic model, but for individual synthesis it is initialized randomly.

To investigate of the effectiveness of attention mechanism in the shared hidden layer, we use directly concatenating the two outputs of Ω^{MEAN} and Ω^{ADA} to replace the proposed attention mechanism, which is marked as "P-GD-LI-voiced (no attention)". It can be observed that the attention mechanism in the shared hidden layer can improve the accuracy of acoustic model, especially in the low resource situation. A possible explanation is that the attention mechanism dynamically adjusts the portion of Ω^{MEAN} and Ω^{ADA} to reach an optimum. By observing the preference scores of subjective evaluations in the Table 4, it is worth noticing that the attention mechanism in the shared hidden layer can improve the naturalness of synthetic speech more than the similarity. The proposed method achieves more preference in both naturalness and similarity.

5. CONCLUSION

In this paper, we mainly investigate phoneme dependent speaker embedding by combining global utterance level speaker representations and local phoneme level speaker representations using attention mechanism. The speaker and phoneme dependent speaker representations give more details about each style of speaker and can generate a specific vector for guiding different texts. The attention mechanism makes the model learn the average model and variation among each speaker simultaneously in separate structures. Experimental results showed that speaker embedding we proposed can control speaker identity effectively during speaker adaptive training. Further, we will explore to perform speaker identity control by some end-to-end methods.

6. ACKNOWLEDGEMENTS

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002801) and the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61771472, No.61603390), and Inria-CAS Joint Research Project (No. 173211KYSB20170061).

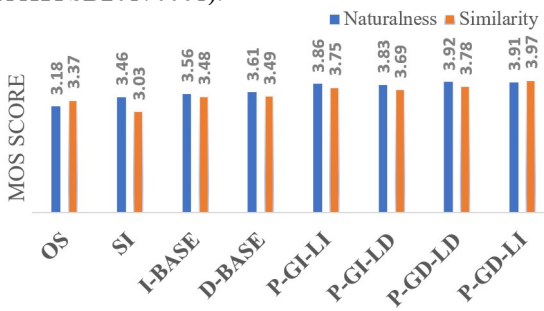

Fig.3 MOS results for multi-speaker speech synthesis. Each system evaluates the naturalness (left) and similarity (right) of synthetic speech. Corpus with only voiced frames is used to extract local phoneme level speaker representations.

Table 4: Preference scores subjective evaluations

System A	Scores A (%)	Scores B (%)	System B
P-GD-LI-voiced	53.6/50.86	46.4/49.14	P-GD-LI-voiced (no attention)
	62.5/68.45	37.5/31.55	I-Base
	64.3/69.63	35.7/30.37	D-Base
	70.7/80.36	29.3/19.64	SI
	80.6/78.65	19.4/11.35	OS

REFERENCES

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

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

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, et al, "Tacotron: Towards End-to-End Speech Synthesis," in Proceedings INTERSPEECH. ISCA,2017,4006-4010.

[4] J Shen, R Pang, R J Weiss, et al, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in Proceedings ICASSP . IEEE, 2018, pp. 373–376.

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

[6] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in Proceedings ICASSP. IEEE, 2015, pp. 4470–4474.

[7] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in Proceedings INTERSPEECH. ISCA,2014, pp. 1964–1968.

[8] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in Proceedings ICASSP. IEEE, 2014, pp. 3829–3833.

[9] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in Proceedings ICASSP. IEEE, 2015, pp. 4470–4474.

[10] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural network employing multi-task learning and stacked bottleneck features for speech synthesis," in Proceedings ICASSP. IEEE, 2015, pp. 4460–4464.

[11] Z Wu, P Swiqetojanski, C Veaux, et al. "A study of speaker adaptation for DNN-based speech synthesis," in Proceedings INTERSPEECH. ISCA,2015.

[12] Y Zhao, D Saito, N Minematsu, "Speaker Representations for Speaker Adaptation in Multiple Speakers' BLSTM-RNN-Based Speech Synthesi," in Proceedings INTERSPEECH. ISCA, 2016:2268-2272.

[13] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in Proceedings IEEE Spoken Language Technology Workshop. IEEE, 2014.

[14] R Doddipatla, N Braunschweiler, R Maia, "Speaker Adaptation in DNN-Based Speech Synthesis Using d-Vectors," in Proceedings INTERSPEECH. ISCA, 2017:3404-3408.

[15] M Wan, G Degottex, M Gales, "Waveform-Based Speaker Representations for Speech Synthesis," in Proceedings INTERSPEECH. ISCA,2018.

[16] Y Fan, Y Qian, F K Soong, et al. "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in Proceedings ICASSP. IEEE, 2015:4475-4479.

[17] Q. Yu, P. Liu and L. Cai, "Learning cross-lingual information with multilingual BLTSM for speech synthesis of low-resource language," in Proceedings ICASSP. IEEE, March 2016, pp. 1233–1236.

[18] H Zen, N Braunschweiler, S Buchholz, et al. Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization. IEEE Transactions on Audio Speech & Language Processing. IEEE, 2012, 20(6):1713-1724.

[19] Y Fan, Y Qian, F K Soong, et al. "Speaker and language factorization in DNN-based TTS synthesis," in Proceedings ICASSP. IEEE, 2016:5540-5544.

[20] B Li, H Zen, "Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis," in Proceedings INTERSPEECH. ISCA, 2016:2468-2472.

[21] M. Morise, F. Yokomori, K. Ozawa, "WORLD, A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," Ieice Transactions on Information & Systems,2016, 99(7):1877-1884.

[22] E Variani, L Xie, E Mcdermott, et al. "Deep neural networks for small footprint text-dependent speaker verification," in Proceedings ICASSP. IEEE, 2014:4052-4056.

[23] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in Proceedings INTERSPEECH. ISCA, 2014, pp. 2189–2193.

[24] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMMBased Speech Synthesis," in Proceedings EUROSPEECH, vol.5, pp.2347– 2350, 1999.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE, Dec. 2011.

[26] M. D. Zeiler, "Adadelta: An adaptive learning rate method," Computer Science, 2012.

[27] M. Abadi, A. Agarwal, P. Barham, et al, "Tensorflow: large-scale machine learning on heterogeneous distributed systems", 2016.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research,vol. 15, no. 1, pp. 1929–1958, 2014.

[29] R. F. Kubichek, "Mel-cepstral distance measure for objective peech quality assessment," in Communications, Computers and Signal Processing, vol. 1. IEEE, 1993, pp. 125-128.