A Comparison of Expressive Speech Synthesis Approaches based on Neural Network

Liumeng Xue
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing School of Computer Science
Northwestern Polytechnical University
Xi'an, China
lmxue@nwpu-aslp.org

Xiaolian Zhu
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing School of Computer Science
Northwestern Polytechnical University
Xi'an, China
Hebei University of Economics and Business
Shijiazhuang, China
xlzhu@nwpu-aslp.org

Xiaochun An
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing School of Computer Science
Northwestern Polytechnical University
Xi'an, China
xiaochunan@npu-aslp.org

Lei Xie*
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing School of Computer Science
Northwestern Polytechnical University
Xi'an, China
lxie@nwpu-aslp.org

ABSTRACT

Adaptability and controllability in changing speaking styles and speaker characteristics are the advantages of deep neural networks (DNNs) based statistical parametric speech synthesis (SPSS). This paper presents a comprehensive study on the use of DNNs for expressive speech synthesis with a small set of emotional speech data. Specifically, we study three typical model adaptation approaches: (1) retraining a neural model by emotion-specific data (retrain), (2) augmenting the network input using emotion-specific codes (code) and (3) using emotion-dependent output layers with shared hidden layers (multi-head). Long-short term memory (LSTM) networks are used as the acoustic models. Objective and subjective evaluations have demonstrated that the multi-head approach consistently outperforms the other two approaches with more natural emotion delivered in the synthesized speech.

CCS CONCEPTS

• **Information systems → Speech / audio search**; • **Computing methodologies → Neural networks**;

KEYWORDS

Statistical parametric speech synthesis; expressive speech synthesis; text-to-speech; neural networks; retrain; code; multi-head network

*Lei Xie is the corresponding author.

ASMMC-MMAC'18, October 26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5985-6/18/10…$15.00
https://doi.org/10.1145/3267935.3267947

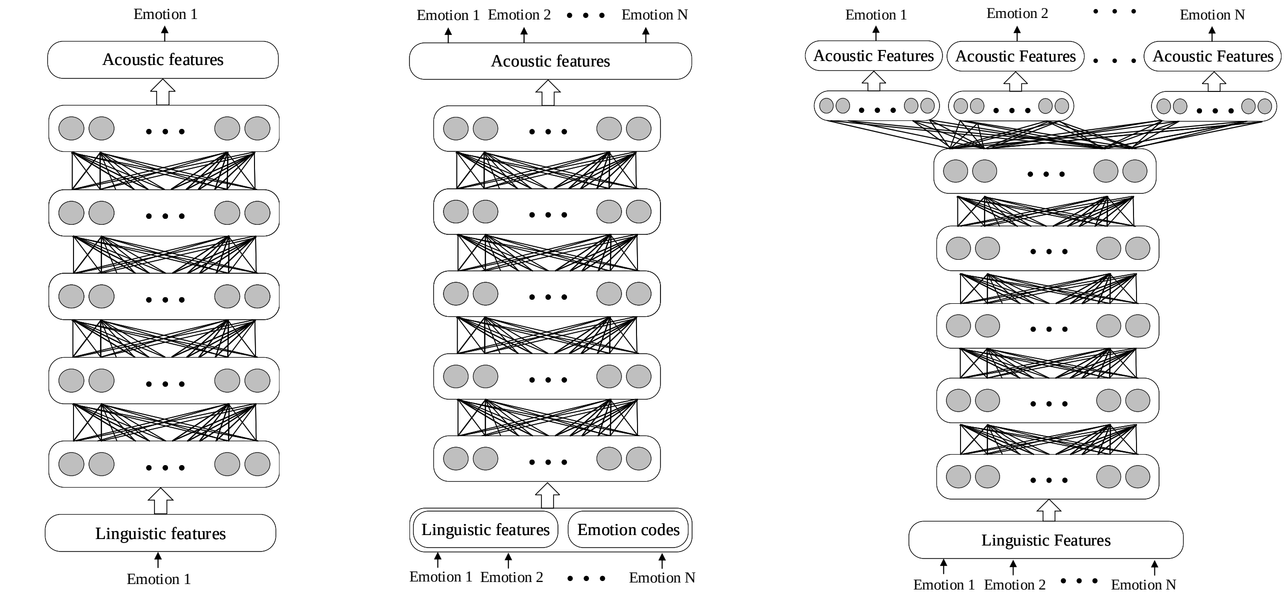
ACM Reference Format:
Liumeng Xue, Xiaolian Zhu, Xiaochun An, and Lei Xie. 2018. A Comparison of Expressive Speech Synthesis Approaches based on Neural Network. In *The Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data Workshop (ASMMC-MMAC'18), October 26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3267935.3267947

1 INTRODUCTION

Deep neural networks (DNNs) have become the main stream in a variety of speech processing tasks, including speech synthesis. Many researches have indicated that DNNs considerably outperform hidden Markov models (HMMs) in statistical parametric speech synthesis (SPSS) [1–4]. Besides, recent end-to-end approaches, together with the WaveNet vocoder [5], have shown exciting improvements in generating natural human sounding [6–9].

The quality and naturalness of the synthesized speech have been significantly improved since the use of DNNs, but the expressiveness still lags far behind. The demand for expressive speech synthesis is increasing greatly for a number of applications, such as audiobook narration, news readers and conversational assistants. DNNs are well known for their adaptability and controllability with a small amount of target training data at hand. Given a base model, e.g., a neural network acoustic model trained using a large set of neutral speech, the target model can be achieved by adapting the base model using a small set of data from the target (e.g., a new speaker or an emotional style).

During the past years, significant efforts have been made to control the speaker variability. Wu *et al.* [10] conducted a systematic study of speaker adaptation techniques, including augmenting a low-dimensional speaker identity vector with linguistic features, performing model adaptation to scale the hidden activation weights, and conducting a feature space transformation. Zhao *et al.* [11] presented a comprehensive study on the use of speaker identity vectors, namely, i-vectors and speaker codes. Specifically, Hojo *et al.* [12] compared the performance of feeding speaker codes into different



(a) Emotion-dependent network.

(b) Emotion-code network.

(c) Multi-head emotion network.

Figure 1: The network structures of three models.

hidden layers. Fan *et al.* [13] proposed a multi-speaker DNN and transferred the hidden layers of the DNN to a new speaker. Later, Li *et al.* [14] presented a multi-language multi-speaker acoustic model using LSTM-RNN.

It is straightforward to migrate from controlling speaker variability to emotional variability. The same approaches used in speaker adaption may be directly applied for emotion adaption. In [15], An *et al.* studied two emotional SPSS approaches: retraining a neutral neural network model and adding emotion codes to each layer of the model. Inoue *et al.* [16] investigated how to control speaker variability and emotional variability at the same time. Specifically, three approaches were studied, including a parallel model with an output layer consisting of both speaker-dependent layers and emotion-dependent layers, a serial model with an output layer consisting of emotion-dependent layers preceded by speaker-dependent layers and an auxiliary input model with an input layer consisting of emotion and speaker codes. More recently, Wang *et al.*[17] proposed an unsupervised style modeling approach to control and transfer speaking styles under the end-to-end speech synthesis framework. Although similar to speaker adaptation, emotion adaptation is not trivial because emotional speech has strong prosody variations that are difficult to model [18–20].

In this paper, we present a comprehensive study on the neural network based expressive SPSS. Specifically, we study three typical approaches: retraining a neural model by emotion-specific data (retrain), augmenting the network input using emotion codes (code) and using emotion-dependent output layers with shared hidden layers (multi-head). Long-short term memory (LSTM) [21] networks are used as the acoustic models that map linguistic features to acoustic features. Different from previous approaches, we provide a study with a variety of emotion styles and test with different size of emotional data. Six typical emotions, i.e., surprise, happiness, sadness, angry, disgust and fear, are investigated and the adaption data size ranges from 10 to 500 utterances. Both objective and subjective evaluations are provided. We have discovered that the multi-head approach consistently outperforms the other two approaches.

The rest of this paper is organized as follows. In Section 2, we introduce the three approaches. Next, we describe a series of experiments and report the results in Section 3. Finally, some conclusions are drawn in Section 4.

2 MODEL DESCRIPTION

2.1 Emotion-dependent Model Retraining

An apparent idea is to retrain a neutral model using emotion-dependent data [22]. The neutral model is usually trained with a large set of data, which provides a good coverage on pronunciations. The small set of emotion-dependent data is used to *fine-tune* the parameters of the neutral model. Figure 1 (a) illustrates such kind of approach, which finally results in an emotion-dependent network for each emotion type.

2.2 Emotion-code Modeling

Augmenting linguistic features with emotion codes as neural network input is another straightforward approach for emotional SPSS. Other approaches, e.g., adding emotion codes to each network layer [15] and using emotion IDs to switch layers for training different emotions [23], can also be considered. But in our model, we simply use a one-hot vector to represent emotion codes. Specifically, we use a 7-dimensional one-hot vector, representing six emotion types and the neutral emotion, respectively. The hidden layers

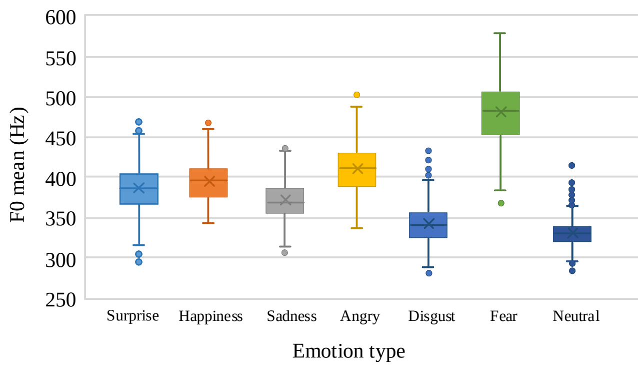


Figure 2: F0 mean distribution for different emotion types.

are shared among all emotion types. The network architecture is illustrated in Figure 1 (b). Figure 2 shows that the fundamental frequency (F0) value varies across different emotions. In [24], Yang *et al.* showed that speaker-dependent normalization could achieve relatively higher speaker similarity scores than the global normalization. Inspired by this, we perform data normalization and de-normalization for each emotion type respectively to take emotion variations into account. In the model training stage, all emotion data are shuffled, and each mini-batch contains data from different emotion categories. In addition, a pre-trained neutral emotion network is used to initialize the parameters.

2.3 Multi-head Emotion Modeling

Another approach is called multi-head emotion modeling, illustrated in Figure 1 (c). This method is motivated by a multi-speaker DNN [13] and the emotion additive model [25]. The input of this network is linguistic features without any auxiliary features. The hidden layers can be regarded as global feature transformation shared across multiple emotions, while the output layers are emotion-dependent, each 'head' representing the specific emotion characteristics. Similar with the emotion-code modeling introduced in Section 2.2, we perform data normalization and de-normalization respectively for different emotions.

In the practice, we firstly pre-train a standard neutral-emotion model with single-head output and then expand the model to multiple heads. Specifically, the parameters of the hidden layers of the original single-head network are used to initialize the hidden layers of the new multi-head emotion model. The last-layer parameters of the single-head model are simply copied to each head of the multi-head emotion model as initialization.

Different emotion heads are trained simultaneously under a multi-task learning framework. The training strategy is as follows. For each training epoch, the training data of each emotion is used to calculate an emotion-dependent loss. As for six emotions considered in this study, six losses are obtained in each epoch. Then we calculate the average loss of all the six emotions, which is back propagated to update the parameters of the hidden layers. The emotion-dependent loss is back propagated to each head to update each emotion-dependent output layer. Note that each mini-batch contains data from one emotion type only.

3 EXPERIMENTS

3.1 Experimental Setup

Speech data in Mandarin Chinese from a female professional speaker is used for the experiments. The dataset contains about 10,000 utterances with neutral emotion (standard reading) and 3498 utterances for six emotions (583 utterances each). The emotion types are surprise, happiness, sadness, angry, disgust and fear. We randomly select 500 utterances as the training set, 58 utterances as the validation set, and 25 utterances as the testing set. In the experiments, we vary the size of emotion training data from 10 to 500 to observe the performances. The sampling rate of all waveform files is 16 kHz. The REAPER tool is used to extract F0 in log-scale at a 5-ms step and the STRAIGHT vocoder [26] is used to extract 41-dimensional line spectral pairs (LSP). The final acoustic features are 51-dimensions including 41-dimensional LSP, one extra binary voiced/unvoiced flag and 9-dimensions F0 scores (previous 4 frames, current frame and proceeding 4 frames). Our empirical test shows that modeling F0 context is essential for a stable intonational performance.

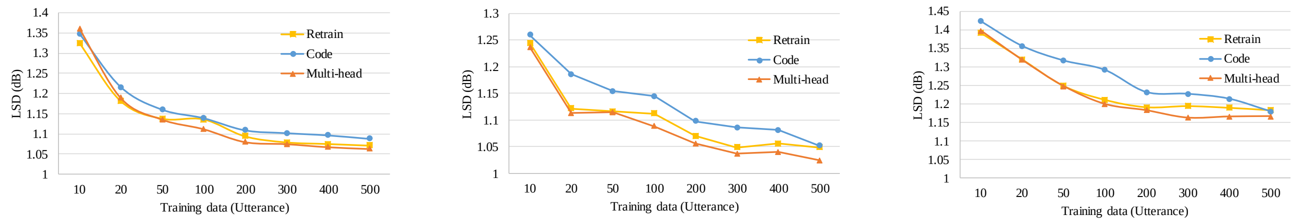
As for text analysis, we extract a rich set of textual features including phoneme information, prosodic boundary, state information and the corresponding position index, represented by a 297-dimensional vector with binary and/or numerical features. The state information is obtained by forced alignment using the Hidden Markov Model Toolkit (HTK) [27]. For the neural network input, the only difference between the emotion-code approach and the other two approaches is that the linguistic features are supplemented by the one-hot emotion codes.

For all the tested models, we use bidirectional long short-term memory (BLSTM) based acoustic models. The network structure is three feed-forward layers with 512 nodes, followed by two BLSTM layers with 512 cells and a liner output layer. ReLU is used as the activation function. All systems are optimized using Adam optimizer [28]. The networks are trained using an initial learning rate of 0.0001. All the experiments are carried out using TensorFlow [29].

3.2 Objective Evaluation

Our aim is to use a small set of emotion data to realize decent emotional speech synthesis. Hence we vary the training data size to evaluate the performances of the three approaches. Specifically, LSP distortion (LSD) is measured to evaluate the spectrum distortion, and root mean squared error (RMSE) is used to calculate the F0 prediction error.

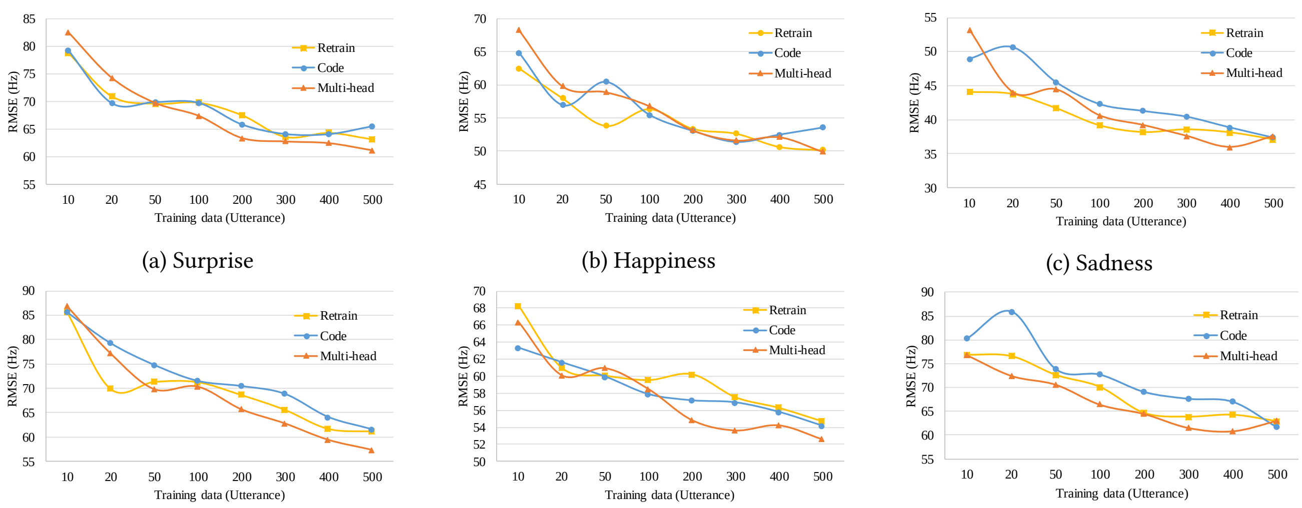
LSD and F0 RMSE trajectories are shown in Figure 3 and Figure 4, respectively. From the results, we can see that the multi-head emotion modeling approach achieves lower prediction errors as compared with the other two approaches. The emotion-code modeling approach results in highest LSD and F0 RMSE in most cases. We can also observe that, the changes of LSD curve with different training utterances are relatively steady for three methods. As for F0 RMSE, the emotion-dependent model retraining approach is not stable. For example, the values of F0 RMSE suddenly increase with 20 training utterances for the sadness emotion and the fear emotion. On the contrast, the F0 RMSE of the code approach and the multi-head approach are relatively stable. It indicates that sharing data from different emotion categories is quite helpful.



(a) Surprise (b) Happiness (c) Sadness

(d) Angry (e) Disgust (f) Fear

Figure 3: Objective evaluation results of LSD with different amount of training data using three methods for emotions: (a) surprise, (b) happiness, (c) sadness, (d) angry, (e) disgust and (f) fear.



(a) Surprise (b) Happiness (c) Sadness

(d) Angry (e) Disgust (f) Fear

Figure 4: Objective evaluation results of F0 RMSE with different amount of training data using three methods for emotions: (a) surprise, (b) happiness, (c) sadness, (d) angry, (e) disgust and (f) fear.

Obviously, the LSD and F0 RMSE values are decreased as the increase of the training data. But the trend becomes flat when the number of training utterances exceeds 100. The values achieved by the multi-head approach trained using 200 or 300 utterances are similar to those achieved by the code and the retrain approaches trained using 500 utterances. We can notice that LSD and F0 RMSE vary dramatically across different emotions. This is reasonable because different emotions have different acoustic characteristics. For example, as shown in Figure 2 earlier, the F0 values have quite different ranges for different emotion types. This poses significant challenges to expressive speech synthesis.

3.3 Subjective Evaluation

We conduct subjective evaluations using mean opinion score (MOS) tests. The synthesized speech samples are chosen from the three approaches trained using 500 sentences. We randomly select 5 tested sentences for each emotion respectively, resulting in a total of 30 sentences for subjective listening. All the listening samples are presented in a shuffled order. There are 25 native listeners with normal hearing participate in the test. The listeners are asked to rate the overall impression and the expressiveness of the testing samples generated by the three approaches as well as the original recordings.

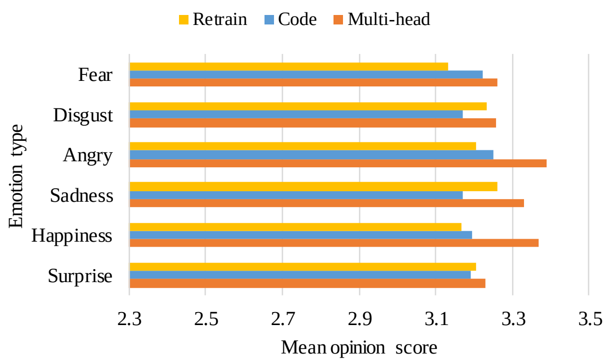


Figure 5: Overall impression results for six emotions.

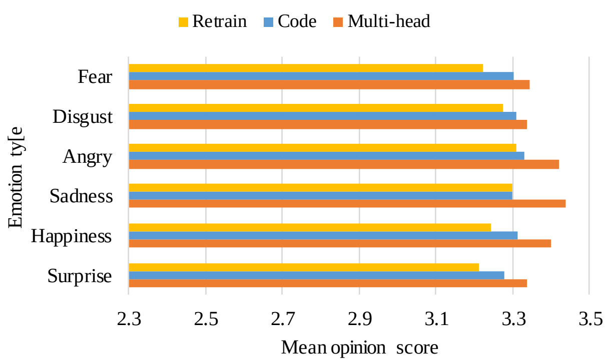


Figure 6: Expressiveness results for six emotions.

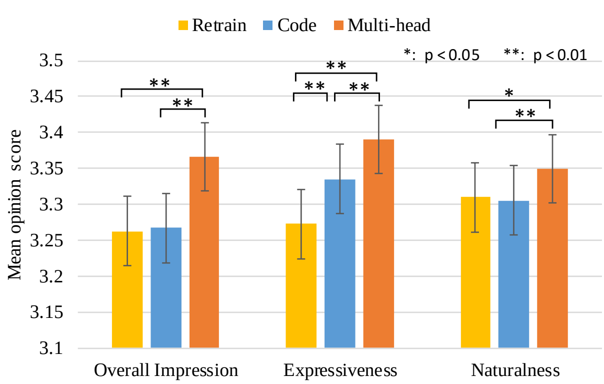


Figure 7: Subjective evaluation results of three methods with 95% confidence interval.

The MOS results for overall impression and the expressiveness for six emotions are summarized in Figure 5 and Figure 6, respectively. From the results, we can clearly see that the multi-head approach performs the best among the three approaches, in terms of both overall impression and expressiveness. The results without distinguishing emotion categories are further summarized in Figure 7. According to both overall impression and expressiveness, we can conclude that the differences between the multi-head approach and the other two approaches are significant. The performance achieved by the multi-head approach is obviously superior. Listeners have pointed out that, although they can hear clear emotions, the intonation presence of the retrain and emotion-code approaches is not stable and sometimes even unnatural. On the contrast, the multi-head approach always provides stable intonation and more natural speech. The difference among emotions mostly lies in the changes of acoustic features. The multi-head network can effectively make use of the data from all emotions to train the hidden layers while the emotion-dependent output layers are used to represent the acoustic difference among emotions. Thus this approach uses both increased data volume and emotional discrimination.

4 CONCLUSIONS

In this work, we have compared the performances of three emotion adaptation approaches, namely emotion-dependent model retraining, emotion-code modeling and multi-head emotion modeling. Both objective and subjective experiments on six typical emotion types have confirmed that the multi-head emotion modeling approach obtains superior performance. In the future, we will study emotion adaptation under the end-to-end text-to-speech framework [17] [30, 31].

ACKNOWLEDGMENTS

The research work is supported by the National Key Research and Development Program of China under Grant No.: 2017YFB1002102 and the National Natural Science Foundation of China under Grant No.: 61571363.

REFERENCES

[1] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7962–7966. IEEE, 2013.

[2] Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong. On the training aspects of deep neural network (DNN) for parametric tts synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3829–3833. IEEE, 2014.

[3] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4460–4464. IEEE, 2015.

[4] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen Meng, and Li Deng. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, 32(3):35–52, 2015.

[5] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, page 125, 2016.

[6] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[7] Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*, 2017.

[8] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *arXiv preprint arXiv:1712.05884*, 2017.

[9] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint*, 2017.

[10] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King. A study of speaker adaptation for DNN-based speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Yi Zhao, Daisuke Saito, and Nobuaki Minematsu. Speaker representations for speaker adaptation in multiple speakers BLSTM-RNN-based speech synthesis. *space*, 5(6):7, 2016.

[12] Nobukatsu Hojo, Yusuke Ijima, and Hideyuki Mizuno. An investigation of DNN-based speech synthesis using speaker codes. In *INTERSPEECH*, pages 2278–2282, 2016.

[13] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He. Multi-speaker modeling and speaker adaptation for DNN-based tts synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4475–4479. IEEE, 2015.

[14] Bo Li and Heiga Zen. Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis. In *INTERSPEECH*, pages 2468–2472, 2016.

[15] Shumin An, Zhenhua Ling, and Lirong Dai. Emotional statistical parametric speech synthesis using LSTM-RNNs. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*, pages 1613–1616. IEEE, 2017.

[16] Katsuki Inoue, Sunao Hara, Masanobu Abe, Nobukatsu Hojo, and Yusuke Ijima. An investigation to transplant emotional expressions in DNN-based tts synthesis. In *Asia-Pacific Signal and Information Processing Association Summit and Conference*, pages 1253–1258, 2017.

[17] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018.

[18] Mireille Besson, Cyrille Magne, and Daniele Schön. Emotional prosody: sex differences in sensitivity to speech melody. *Trends in cognitive sciences*, 6(10):405–407, 2002.

[19] A Paeschke. Global trend of fundamental frequency in emotional speech. *Speech Prosody*, 2004.

[20] Yasuki Hashizawa, Shoichi Takeda, Muhd Dzulkhiflee Hamzah, and Ghen Ohyama. On the differences in prosodic features of emotional expressions in Japanese speech according to the degree of the emotion. In *Speech Prosody 2004, International Conference*, 2004.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[22] Dimitrios Kollias, Athanasios Tagaris, and Andreas Stafylopatis. On line emotion detection using retrainable deep neural networks. In *Computational Intelligence*, pages 1–8, 2017.

[23] Katsuki Inoue, Sunao Hara, Masanobu Abe, Nobukatsu Hojo, and Yusuke Ijima. An investigation to transplant emotional expressions in DNN-based tts synthesis. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*, pages 1253–1258. IEEE, 2017.

[24] Shan Yang, Zhizheng Wu, and Lei Xie. On the training of DNN-based average voice model for speech synthesis. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6, 2016.

[25] Yamato Ohtani, Yu Nasu, Masahiro Morita, and Masami Akamine. Emotional transplant in statistical speech synthesis based on emotion additive model. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[26] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De 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 sounds1. *Speech communication*, 27(3-4):187–207, 1999.

[27] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3:175, 2002.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[29] MartÃŋn Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. Tensorflow: a system for large-scale machine learning. 2016.

[30] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *arXiv preprint arXiv:1803.09047*, 2018.

[31] Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A Saurous. Uncovering latent style factors for expressive speech synthesis. *arXiv preprint arXiv:1711.00520*, 2017.