A DNN-based emotional speech synthesis by speaker adaptation

Hongwu YANG *†‡ and Weizhao ZHANG*† and Pengpeng ZHI*
* College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China
† Engineering Research Center of Gansu Province for Intelligent Information Technology and Application, Lanzhou 730070, China
‡ National and Provincial Joint Engineering Laboratory of Learning Analysis Technology in Online Education, Lanzhou 730070, China
E-mail: yanghw@nwnu.edu.cn

***Abstract*—The paper proposes a deep neural network (DNN)-based emotional speech synthesis method to improve the quality of synthesized emotional speech by speaker adaptation with a multi-speaker and multi-emotion speech corpus. Firstly, a text analyzer is employed to obtain the contextual labels from sentences while the WORLD vocoder is used to extract the acoustic features from corresponding speeches. Then a set of speaker-independent DNN average voice models are trained with the contextual labels and acoustic features of multi-emotion speech corpus. Finally, the speaker adaptation is adopted to train a set of speaker-dependent DNN voice models of target emotion with target emotional training speeches. The target emotional speech is synthesized by the speaker-dependent DNN voice models. Subjective evaluations show that comparing with the traditional hidden Markov model (HMM)-based method, the proposed method can achieve higher opinion scores. Objective tests demonstrate that the spectrum of the emotional speech synthesized by the proposed method is also closer to the original speech than that of the emotional speech synthesized by the HMM-based method. Therefore, the proposed method can improve the emotion express and naturalness of synthesized emotional speech.**

I. Introduction

There are many ways to synthesize emotional speech effectively including waveform unit selection method [1], [2], prosodic feature modification method [3] and statistical parametric speech synthesis method [4]. Each method has its advantages and disadvantages. The waveform unit selection method needs a large emotional speech database that is not easy to establish [5]–[7]. The prosodic feature modification method realizes emotional speech synthesis by modifying prosodic features that will reduce the quality of synthesized speech [8]. The HMM-based speech synthesis can be successfully applied to scalability tasks by speaker adaptation techniques and has been shown to significantly improve the perceived quality of synthesized speech [9]. Because the HMM-based statistical parametric speech synthesis can use interpolation [10], emotion vector multiple regression [11] and adaptive techniques [12] to easily transform or modify the speaker's style or emotion, this method has become the main method in emotional speech synthesis. Although the HMM-based statistical parametric speech synthesis works reasonably well in statistical speech synthesis, it has well known limitations. Firstly, decision-tree-based input-to-cluster mapping in HMM-based speech synthesis is inefficient for expressing complex context dependencies, such as the exclusive OR (XOR) problem. This may lead to overfitting to the training data because of the data partitioning issue. Secondly, the cluster-to-feature mapping using single Gaussian distributions with diagonal covariance matrices is established based on two independence assumptions: 1) conditional independence between frames given the state and 2) independence of acoustic features within a frame. This leads to reconstructed spectral envelopes being over-smoothed and the quality of synthetic speech is degraded [13].

Since 2006, deep learning has emerged as a new area of machine-learning research and has also attracted the attention of many signal processing researchers. Both unconditional deep models such as restricted Boltzmann machines (RBMs), deep belief networks (DBNs) and conditional deep models such as DNNs have been intensively studied and explored in statistical speech synthesis. These models can learn a direct, layered, non-linear model from linguistic features to acoustic parameters without decision tree clustering. Compare to the DBN models, DNN models only requires a single computing pass for feature prediction, making it more suitable for real-time synthesis. On the other hand, DNN models the conditional probability instead of the joint probability as in the DBN model, which is more intuitive for the feature mapping task. However, current studies on DNN models are mainly speaker-dependent, which required a significant amount of data from a single speaker to build a stable acoustic model. Therefore, the multi-speaker adaptation methods based on DNN is proposed.

In the research of emotional speech synthesis, Ref. [14] adopts the end to end prosodic conversion method to synthesize speech of different text prompts. Ref. [15] adopts the end to end method to realize the synthesis of emotional speech animation. Ref. [16] and Ref. [17] use DNN and RNN-LSTM to realize the emotional speech synthesis. The above deep learning-based emotional speech synthesis method can produce more natural emotional speech, but the synthesized speech is still unable to meet the emotional requirements when it is trained with multi-speaker or the smaller emotion speech corpus.

The paper proposes a DNN-based emotional speech syn-

thesis method to improve the quality of synthesized emotional speech by speaker adaptation with a multi-speaker and multi-emotion speech corpus. Firstly, a set of speaker-independent DNN average voice models were trained with the contextual labels and acoustic features. Then the speaker adaptation is adopted to train a set of speaker-dependent DNN voice models of target emotion with target emotional training speeches. Subjective evaluations show that comparing with the traditional HMM-based method, the proposed method can achieve higher opinion scores. Objective tests demonstrate that the spectrum of the emotional speech synthesized by the proposed method is also closer to the original speech than that of the emotional speech synthesized by the HMM-based method.

II. Framework of DNN-based emotional speech synthesis by speaker adaptation

The framework of the DNN-based emotional speech synthesis by speaker adaptation is shown in Fig. 1. At the stage of training speaker-independent DNN average voice models, we use a multi-speaker and multi-emotion speech corpus to train a set of speaker-independent DNN average voice models. In speaker-independent DNN average voice models, hidden layers are shared across all the speakers in training corpus. It can be considered as the global linguistic feature transformation shared by all the speakers. Conversely, each speaker has his own output layer, so-called regression layer, to modeling the specific acoustic space of himself. Due to the different from the conventional DNN, it's very important to train the network for all the speakers simultaneously, which means that each mini-batch data should be selected from all the multi-speaker and multi-emotion speech corpus during the stochastic gradient decent (SGD) procedure [18].

In the training procedure for speaker adaptation, we used target emotional training speeches to train a set of speaker-dependent DNN voice models of target emotion, which can be considered the sub-modules of the speaker-independent DNN average voice models by only taking the target speaker regression layer and shared hidden layers. Due to training data for adaptation is very limited, the shared hidden layers from speaker-independent DNN average voice models should be fixed and only the regression layer will be updated. So we use the least squares method, instead of SGD algorithm, to minimize the errors between prediction and ground-truth.

For the above two DNN-based models, linguistic context and acoustic features were treated as input and output respectively. The output of hidden unit is the nonlinear transformation of the former layer, as follow:

$$h_j^{(k)} = \tanh(b_j^{(k)} + \sum_i w_{ij}^{(k)} h_i^{(k-1)}) \quad (1)$$

Where $h_j^{(k)}$ is the jth hidden unit at the kth layer , $b_j^{(k)}$ is the bias of the jth unit at the kth layer, $w_{ij}^{(k)}$ is the weight associated with the link from $h^{(k-1)}$ to $h^{(k)}$.

$$y_j = b_j + \sum_i w_{ij} h_i^{(l)} \quad (2)$$

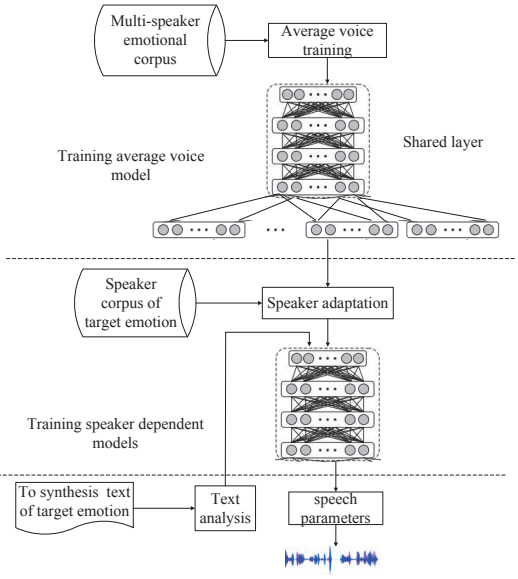

Fig. 1. Framework of DNN-based emotional speech synthesis by speaker adaptation.

Where y_j is the jth unit at the output layer, b_j is the bias of the jth unit at the output layer, w_{ij} is the weight associated with the link from the last hidden layer to the output layer. The $tanh$ function is used as an activation function of hidden layers.

In synthesis stage, the linguistic contextual features of front-end linguistic analysis on input text is fed to the speaker-dependent DNN voice models of target emotion, and then output the acoustic features normalized to zero mean and unit variance. Maximum likelihood parameter generation (MLPG) is applied to generate smooth parameter trajectories from the de-normalised neural network outputs, then spectral enhancement in the cepstral domain is applied to the MCCs to enhance naturalness. WORLD is used to extract the acoustic features and generate the speech waveform [19].

III. Experiments

A. Experimental setup

We use psychological methods to stimulate the emotional speech by an inner stimulated situation. 9 female Mandarin speakers who are not a professional actress are selected to record the emotional speech in a sound proof studio. There are 11 kinds of emotions including sadness, relax, anger, anxiety, surprise, fear, contempt, docile, joy, disgust and neutral. Each speaker records 300 Mandarin sentences for one emotion. All recordings are saved in the Microsoft Windows WAV format as sound files (mono-channel, signed 16 bits, sampled at 16 kHz). We select all 8 speaker's speeches as the emotional training corpus to train a set of speaker-independent DNN average voice models. 1 speaker target emotional speech corpus are used target emotional training corpus to train a set of speaker-dependent DNN voice models of target emotion. There are 50

utterances in target emotional training speeches, 40 of which were training sets and 10 utterances were test sets. WORLD is used to extract 60-dimensional Mel-Cepstral Coefficients (MCCs),1-dimensional band aperiodicities (BAPs), and log F0 at 5 msec step.

The input of a DNN contain 416 binary linguistic features, 9 numerical features and 1 binary feature to represent gender information. The linguistic features include initial, final, syllable, part-of-speech, prosodic word, prosodic phrase and positional information within a syllable, prosodic word, prosodic phrase, etc. The 9 numerical features involved frame position in the HMM state and phoneme (initial or final), state position in phoneme, state and phoneme duration. The output acoustic features comprise 60-dimensional MCCs,1-dimensional BAPs, and 1-dimensional log F0, their corresponding delta and delta-delta features, and a voice/unvoiced binary value. In total, the acoustic feature vector was 187 dimension. The input features were normalised to the range of [0.01, 0.99], and the output features were normalised by speaker-dependent mean and variance. The MLPG algorithm was used to generate smoothed parameter trajectories, followed by spectral enhancement post-filtering in the cepstral domain.

The speaker-independent DNN average voice models and speaker-dependent DNN voice models has 6 hidden layers, and each hidden layer has1024 units. The minibatch size is set to 256, and momentum is adopted to accelerate convergence. For the first 10 epochs, the momentum is set to 0.6, and is then increased to 0.9. A fixed learning rate of 0.0008 is used in the first 10 epochs for speaker-independent DNN average voice models. The learning rates are halved at each epoch after 10 epochs. L2 regularization is applied to the weights with a penalty factor of 0.00001. The maximum number of epochs was set to 25. In the implementation, we used the Open Source toolkit Merlin.

In order to evaluate the method proposed, three sets of comparative experiments are included DNN(ADP), DNN and HMM(ADP). In DNN(ADP), the method proposed, the speaker adaptation is trained by using multi-speaker and multi-emotion speech corpus. In DNN, we use 1 speaker target emotion speech corpus to train speaker-dependent DNN voice models. In HMM(ADP), the speaker adaptation is trained by using multi-speaker and multi-emotion speech corpus based on HMM.

TABLE I
RMSE OF F0 AND DURATION (DUR) FOR THREE EMOTIONAL SPEECH SYNTHESIS METHODS.

Emotional	F0/Hz			Dur/s		
	HMM (ADP)	DNN	DNN (ADP)	HMM (ADP)	DNN	DNN (ADP)
Sadness	58.6	65.5	52.0	0.232	0.238	0.202
Anger	66.6	62.8	60.5	0.118	0.111	0.090
Relax	25.0	31.6	23.1	0.176	0.183	0.106
Anxiety	67.5	67.6	47.5	0.124	0.136	0.087
Surprise	66.4	71.3	64.3	0.171	0.153	0.132
Fear	68.6	78.8	62.8	0.133	0.138	0.088
Contempt	51.8	51.3	49.1	0.132	0.146	0.095
Docile	33.8	42.2	32.4	0.149	0.153	0.105
Joy	72.2	75.0	61.9	0.106	0.113	0.106
Disgust	61.8	64.6	58.8	0.143	0.145	0.094
Neutral	41.4	43.6	39.8	0.143	0.139	0.102

TABLE II
MCD, BAPS AND V/UV ERROR OF THREE EMOTIONAL SPEECH SYNTHESIS METHODS.

Emotional	Method	MCD(dB)	BAPD(dB)	V/U%
Sadness	HMM(ADP)	6.21	0.14	8.66
	DNN	7.15	0.18	10.96
	DNN(ADP)	5.75	0.14	8.35
Anger	HMM(ADP)	9.01	0.39	23.85
	DNN	9.71	0.47	25.49
	DNN(ADP)	8.58	0.36	20.78
Relax	HMM(ADP)	6.94	0.23	16.75
	DNN	7.56	0.28	18.17
	DNN(ADP)	6.77	0.22	15.08
Anxiety	HMM(ADP)	7.07	0.32	8.91
	DNN	7.92	0.41	13.51
	DNN(ADP)	7.01	0.36	9.81
Surprise	HMM(ADP)	7.78	0.40	17.33
	DNN	8.20	0.42	19.14
	DNN(ADP)	6.99	0.32	13.69
Fear	HMM(ADP)	8.04	0.26	20.62
	DNN	8.64	0.29	22.32
	DNN(ADP)	7.18	0.19	10.19
Contempt	HMM(ADP)	6.70	0.26	14.76
	DNN	8.14	0.34	19.82
	DNN(ADP)	6.65	0.25	11.30
Docile	HMM(ADP)	7.10	0.24	8.02
	DNN	7.67	0.34	12.29
	DNN(ADP)	6.60	0.25	7.23
Joy	HMM(ADP)	7.69	0.32	12.70
	DNN	8.46	0.41	14.57
	DNN(ADP)	7.58	0.31	10.14
Disgust	HMM(ADP)	7.25	0.30	10.38
	DNN	8.14	0.39	15.78
	DNN(ADP)	6.60	0.28	10.72
Neutral	HMM(ADP)	7.87	0.31	10.06
	DNN	9.08	0.44	17.52
	DNN(ADP)	7.53	0.28	9.42

B. Objective evaluations

We conduct objective evaluations to analyse the performance of the emotional speech synthesized. The root mean square error (RMSE) of duration and F0 for the synthesized emotional speech as shown in Table I. The Mel-Cepstral Distortion (MCD), BAPs distortion and V/UV error for the synthesized emotional speech as shown in Table II. It is observed that DNN(ADP) and HMM(ADP) achieve better objective result than DNN.

C. Subjective evaluations

In the subjective evaluation, we conduct AB preference test, mean opinion scoring (MOS) and emotional mean opinion scoring (EMOS) to evaluate the naturalness and emotional similarity of the synthesized speech.

We synthesize 3 categories of emotional speech with same speaker. Each category has 11 emotions. In each category, 10 utterances are synthesized for each emotion. A total number of 330 utterances are synthesized. 10 native Mandarin listeners were asked to evaluate the synthesized emotional speech.

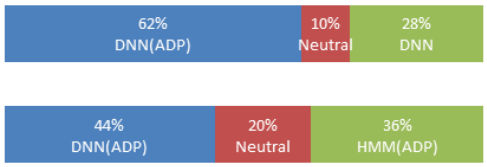

Fig. 2. Preference score of emotional speech synthesis.

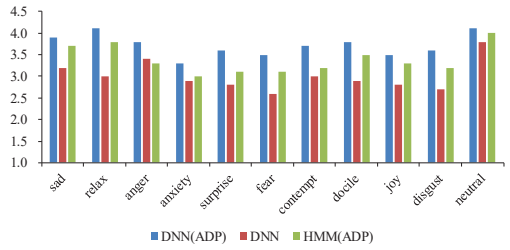

Fig. 3. MOS score of emotional speech synthesized by three methods with 95% confidence interval.

In AB preference test, 10 native Mandarin listeners are asked to give preference (the former better, the latter better and neutral) on 10 pairs of the synthesized emotional speeches from the above approaches. The preference score, shown in Fig. 2, indicates DNN (ADP) can significantly ($p<0.05$) improve the quality of synthesized speech. In MOS and EMOS tests, 10 native Mandarin listeners are also asked to respectively rate the naturalness and emotional similarity of the synthesized speech using five-point scale (i.e. 5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The results are shown in Fig. 3 and Fig. 4. The MOS and EMOS scores of DNN(ADP) are higher the other approaches.

IV. Conclusions

In this work, we present a DNN-based emotional speech synthesis to improve the quality of synthesized emotional speech by speaker adaptation with a multi-speaker and multi-emotion speech corpus. Subjective evaluations and objective test show that comparing with the conventional HMM-based method and DNN-based method, the proposed method can improve the emotion express and naturalness of synthesized emotional speech. In the next step, we plan to use different deep learning methods such as Bidirectional recurrent neural network (RNN), long and short-term memory (LSTM) based RNN to realize emotional speech synthesis, and evaluate the emotional speech synthesized by different methods and different scale corpus.

Acknowledgment

The research leading to these results was partly funded by the National Natural Science Foundation of China (Grant No. 11664036, 61263036), High School Science and Technology Innovation Team Project of Gansu (2017C-03), Natural Science Foundation of Gansu (Grant No. 1506RJYA126), and Student Innovation Project of Northwest Normal University(CX2018Y162).

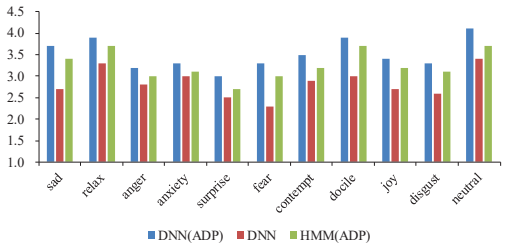

Fig. 4. EMOS score of emotional speeches synthesized by three methods with 95% confidence interval.

References

[1] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Proceedings of the 1992 IEEE International conference on Acoustics, speech and signal processing* vol. 1, pp. 145-148, 1992.

[2] J. Adell, D. Escudero, and A. Bonafonte, "Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence," *Speech Communication* vol. 54, no. 3, pp. 459-476, March 2012.

[3] W. Hamza, E. Eide, and R. Bakis, "Reconciling pronunciation differences between the front-end and the back-end in the IBM speech synthesis system," *Proceedings of the International Conference on Spoken Language Processing*, October 2004.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication* vol. 51, no. 11, pp. 1039-1064, November 2009.

[5] J. F. Pitrelli, R. Bakis, E. M. Eide, and R. Fernandez, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Transactions on Audio, Speech, and Language Processing* vol. 14, no. 4, pp. 1099-1108, July 2006.

[6] B. Murtaza, S. Shrikanth, and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," *Proceedings of the International Conference on Spoken Language Processing*, pp. 1265-1268, September 2002.

[7] E. Eide, "Preservation, Identification, and Use of Emotion in a Text-to-speech System," *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, pp. 127-130, September 2002.

[8] V. Strom , and S. King, "Investigating festival's target cost function using perceptual experiments,"*Proceedings of Conference of the International on Speech Communication Association*, pp. 1873-1876, September 2008.

[9] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems* vol. 88, no. 3, pp. 502-509, 2005.

[10] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE transactions on information and systems* vol. 88, no. 11, pp. 2484-2491, 2005.

[11] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems* vol. 90, no. 9, pp. 1406-1413, 2007.

[12] 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, no. 1, pp. 66-83, 2009.

[13] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, and L. Deng, "Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and

future trends," *IEEE Signal Processing Magazine* vol. 32, no. 3, pp. 35-52, 2015.

[14] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, and R. A. Saurous, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[15] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, and I. Matthews, "A deep learning approach for generalized speech animation," *Acm Transactions on Graphics* vol. 36, no. 4, pp. 93, 2017.

[16] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "An investigation to transplant emotional expressions in DNN-based TTS synthesis," *Proceedings of Asia-Pacific Signal and Information Processing Association Summit and Conference* pp. 1253-1258, December 2017.

[17] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," *Proceedings of Asia-Pacific Signal and Information Processing Association Summit and Conference* pp. 1613-1616, December 2017.

[18] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* pp. 4475-4479, April 2015.

[19] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," *Proceedings of ISCA Speech Synthesis Workshop* pp. 202-207, September 2016.