MULTI-SPEAKER EMOTIONAL ACOUSTIC MODELING FOR CNN-BASED SPEECH SYNTHESIS

Heejin Choi, Sangjun Park, Jinuk Park, Minsoo Hahn

School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{change, psj, windclay, mshahn2}@kaist.ac.kr

ABSTRACT

In this paper, we investigate multi-speaker emotional acoustic modeling methods for convolutional neural network (CNN) based speech synthesis system. For emotion modeling, we extend to the speech synthesis system that learns a latent embedding space of emotion, derived from a desired emotional identity, and we use emotion code and mel-frequency spectrogram as an emotion identity. In order to model speaker variation in a text-to-speech (TTS) system, we use speaker representations such as trainable speaker embedding and speaker code. We have implemented speech synthesis systems combining speaker representation and emotion representation and compared them by experiments. Experimental results have demonstrated that the multi-speaker emotional speech synthesis approach using trainable speaker embedding and emotion representation from mel spectrogram achieves higher performance when compared with other approaches in terms of naturalness, speaker similarity, and emotion similarity.

Index Terms— Text-to-speech, expressive speech synthesis, multi-speaker acoustic modeling, convolutional neural network

1. INTRODUCTION

Deep neural networks (DNNs) have been widely adopted in various speech processing tasks, including speech synthesis [1, 2]. Currently, the demand for improving flexibility and controllability in speech synthesis system is increasing greatly for human-computer interactions.

In relation of speaker variability, Fan et al. [3] proposed a DNN-based multi-task learning for multi-speaker modeling that has speaker specific outputs, and Pascual et al. [4] utilized similar approach using recurrent neural networks (RNN)-long short term memory (LSTM) architectures [5]. Hojo et al. [6] introduced the use of speaker codes, and Zhao et al. [7] compared the performance of feeding speaker identity vectors, namely, i-vectors and speaker codes, into the input layer. In addition, Li et al. [8] presented the multi-language multi-speaker text-to-speech (TTS) system. Inspired by the controlling speaker variability, the emotion control techniques have been studied. An et al. [9] suggested an approach that retraining a neutral neural network model by adding emotion codes to each layer of the model. Inoue et al. [10] investigated how to control speaker variability and emotional variability at the same time.

Recently, Skerry-Ryan et al. [11] have demonstrated prosody transfer via a learned representation of prosody directly from acoustic signals, namely mel spectrograms. They showed that conditioning Tacotron [12] on the learned embedding resulted in synthesized audio that matched the prosody of the reference signal even when the reference and synthesis speakers were different. The predicted mel-frequency spectrograms [13] were synthesized via WaveNet vocoder [14] to improve audio fidelity.

Motivated in a way similar to these previous studies, including modeling of speaker identity, we augment speech synthesis system with reference encoder [11] which extracts a fixed-length learned representation from emotion identity and reproduce the desired speaker's emotional speech audio. In this work, we conduct a comparative study on the controllability of different speaker representation and emotion representation. Besides, we synthesize speech signals from predicted mel spectrograms using WaveNet vocoder and evaluate the performance in quality, naturalness, and similarity.

2. MULTI SPEAKER EMOTIONAL SPEECH SYNTHESIS

The overall architecture of the multi-speaker emotional speech synthesis system is illustrated in Figure 1. In the convolutional neural network (CNN)-based speech synthesis system, CNN takes the linguistic features as input and acoustic features (mel spectrograms) as output and learns the mapping between the linguistic and acoustic space. To take long contextual information, we use dilated convolutional network [24] instead of recurrent neural network (RNN). Similar to [15], the convolutional block consists of a 1-D convolution, a gated linear unit as a learnable nonlinearity [25], and a residual connection to the input. The speaker representation corresponding to the true speaker of the signal is added as a bias to the convolution filter output after a softsign function. To explicitly control emotion, we add the reference encoder module to learn the latent representation of emotion. As with the speaker representation, the emotion representation is used across the convolutional layers.

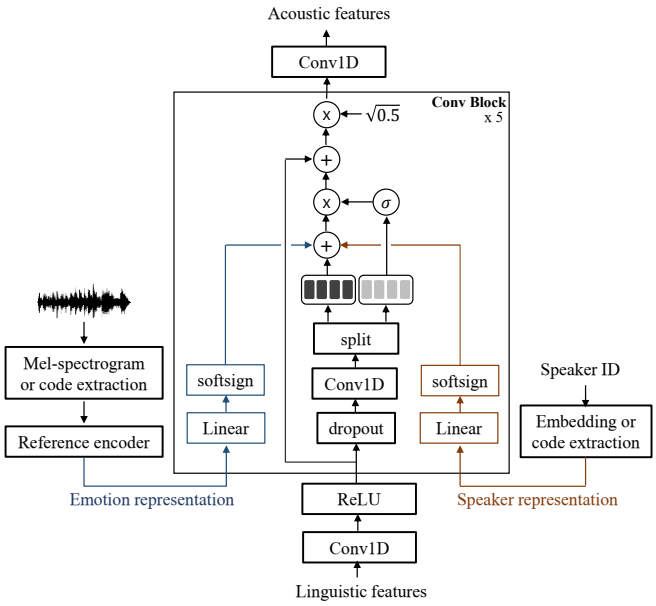

Fig. 1 Multi-speaker emotional speech synthesis system

2.1. Emotion representation modeling

We extend the CNN architecture by adding the reference encoder module [11] that takes an emotion identity as input and extracts a fixed-length embedding from it. During inference, we can use the reference encoder to encode any desired emotional mel spectrogram sequence or emotion code. For the reference encoder architecture, we use a 6-layer 2-D convolutional network that the number of filters in each layer are 32, 32, 64, 64, 128, and 128. Each layer is composed of 3 x 3 filters with 2 x 2 stride, same padding, and ReLU activation. Batch normalization [16] is applied to every layer. The output of the final convolutional layer is passed into a single Gated Recurrent Unit (GRU) [17] layer containing 128 units, followed by tanh activations.

Given an emotion identity, the reference encoder retrieves the corresponding emotion representation and this representation is used across the CNN architecture.

2.1.1. Emotion estimation from mel spectrogram

A mel-frequency spectrogram [13] includes various kinds of information not only linguistic but also non-linguistic such as prosody. As the prosody contains emotional information, mel spectrogram can be used as emotion identity. Hence, estimation of emotion representation from mel spectrogram sequences utilizing the reference encoder is investigated in this paper.

2.1.2. Emotion estimation from emotion code

Emotion code has achieved promising results in emotional speech synthesis [9, 10]. As described in [10], if there are M emotions' corpora for model training, the emotion code $E^{(j)}$ for the j-th emotion is defined as $E^{(j)} = ({e_1}^{(j)}, {e_2}^{(j)}, \ldots, {e_M}^{(j)})$, where each value ${e_m}^{(j)}$ is expressed as follows.

$$ {e_m}^{(j)} = \begin{cases} 1 \ (m = j) \\ 0 \ (m \neq j) \end{cases} \quad (1) $$

2.2. Speaker representation modeling

Learning multi-speaker models via conditioning on speaker representation is straightforward. In order to synthesize speech using multiple speakers, we describe trainable speaker embedding method similar to that presented in [15], and speaker code method [6, 7, 10].

2.2.1 Speaker embedding

We augment our acoustic model with a single low-dimensional speaker embedding vector per speaker. The weights of speaker embedding are initialized randomly from a normal distribution with mean 0 and standard variation 0.01. The speaker embedding is trained jointly with the model, and thus the hidden layers are shared among all speaker types. The i-th speaker representation $SE^{(i)}$ is the corresponding embedding vector of speaker index i.

2.2.2 Speaker code

To control speaker, speaker code method based on neural networks has been demonstrated as a valid multi-speaker modeling method. As described in Section 2.1.2, we can simply use $S^{(i)} = ({s_1}^{(i)}, {s_2}^{(i)}, \ldots, {s_N}^{(i)})$ to represent the i-th speaker code, and each ${s_n}^{(i)}$ is defined as follows.

$$ {s_n}^{(i)} = \begin{cases} 1 \ (n = i) \\ 0 \ (n \neq i) \end{cases} \quad (2) $$

3. EXPERIMENTS

3.1. Experimental setup

In the experiments, we used speech data in Korean from 10 speakers (5 female and 5 male speakers). The dataset consisted of 200 phonetically balanced sentences in 4 speaking styles, happy, anger, sad, and neutral. We randomly divided 4000 sentences into 3400 utterances for training, 400 utterances for valid, 200 utterances for test. Speech signals were sampled at 22.05 kHz, 16 bits. For acoustic features, we used 80 mel spectrograms and 1024 spectrograms, through a short-time Fourier transform (STFT) using a 1024 window length and 256 hop length. Acoustic features were scaled to log-magnitude and then normalized to the range of (0, 1]. For linguistic features, we extracted a rich set of textual features including phoneme information, prosodic boundary, state information and the corresponding position index, represented by a 342-dimensional vector with binary and/or numerical features. The state information were obtained by forced alignment using the Hidden Markov Model Toolkit (HTK) [18]. Numerical linguistic features were normalized to

have zero mean and unit variance and then rescaled to the range of (0, 1]. After the mapping from linguistic features to acoustic features was learned, the generated acoustic features were fed to WaveNet vocoder [13] to synthesize speech waveform. We separately trained speaker-dependent WaveNet vocoder on all ground-truth mel spectrograms and waveforms except the utterance in the test set. For fast waveform reconstruction, the generated mel spectrograms were synthesized by first learning a linear spectrogram prediction network [15], and then applying Griffin-Lim spectrogram inversion [19].

To evaluate the performance of the objective and subjective evaluations, we trained five systems:

• SC-EC: The model using speaker code and emotion code [10]

• SC-EEC: The model using speaker code (Section 2.2.2) and emotion representation from emotion code (Section 2.1.2)

• SC-EEM: The model using speaker code (Section 2.2.2) and emotion representation from mel spectrogram (Section 2.1.1)

• SE-EEC: The model using speaker embedding (Section 2.2.1) and emotion representation from emotion code (Section 2.1.2)

• SE-EEM: The model using speaker embedding (Section 2.2.1) and emotion representation from mel spectrogram (Section 2.1.1)

In common, all systems were composed of 7 dilated convolutional layers with 256 filters. 1-D convolutional layers consisted of 3 kernels with causal padding, but the first and last layers consisted of 1 kernel with no padding. The numbers of dilations in each convolutional layer were 1, 1, 3, 9, 27, 1, and 1. Dropout with probability 0.05 was applied except the first convolutional layer. The systems were implemented using PyTorch [20] and Adam [21] algorithm was employed as the optimizer.

3.2 Objective evaluation

To objectively evaluate the performance of the above systems, we adopted four measures, mel cepstral distortion (MCD), band aperiodicity distortion (BAPD), fundamental frequency (F0) distortion in the root mean squared error (RMSE), and voice/unvoiced (V/UV) error rate. We extracted 59-dimensional mel cepstral coefficients (MCEPs) plus log energy, 2-dimensional BAPs, logarithmic F0, and V/UV decision from the target and synthesized waveforms using WORLD vocoder [22].

3.3 Subjective evaluation

To evaluate the performances of synthetic speech, subjective evaluation was performed for naturalness, speaker similarity, and emotion similarity. A mean opinion score (MOS) test for naturalness and two types of degradation mean opinion score (DMOS) test for similarity were conducted. 14 Korean subjects listened to the synthesized speech in each of the other five systems through headphones. 160 sentences (4 sentences covering 4 emotions from 10 speakers) were synthesized using each method. A five-point scale (from 1: very unnatural to 5: very natural) was adopted for MOS. In the DMOS tests, the quality of speech synthesized by the five systems was compared to the reference speech vocoded by WaveNet in terms of speaker similarity and emotion similarity. A five-point scale was used to judge the speaker similarity and emotion similarity for DMOS (from 1: very dissimilar to 5: very similar).

Systems	MCD (dB)	BAPD (dB)	F0 RMSE (Hz)	V/UV error (%)
SC-EC	6.70	4.34	41.16	13.26
SC-EEC	6.76	4.40	43.35	13.66
SC-EEM	6.52	**4.23**	39.74	13.32
SE-EEC	6.74	4.40	41.30	**13.17**
SE-EEM	**6.51**	4.34	**39.07**	13.25

Table 1. Results of objective evaluations.

3.4 Results and discussion

The objective performance comparison result is illustrated in Table 1. When employing emotion code as input for the reference encoder, SC-EEC and SE-EEC are on par with or slightly worse than SC-EC baseline. But when mel spectrogram is attached to emotion identity, SC-EEM and SE-EEM achieve better performance than SC-EC baseline. Also, SC-EEM and SE-EEM significantly outperform SC-EEC and SE-EEC, respectively. By utilizing trainable speaker embedding, SE-EEC slightly outperforms SC-EEC, and SE-EEM also marginally outperforms SC-EEM except for BAPD. Moreover, we can see that the results of objective measures are not noticeably different across all systems because WaveNet vocoder generates buzzy voice sometimes, which can be considered as the lack of training data for WaveNet vocoder.

Fig. 2 presents the MOS results for naturalness. SC-EEC and SE-EEC are comparable with SC-EC. Fig. 3 and Fig. 4 show the DMOS results for speaker and emotion similarity, respectively. SC-EEM and SE-EEM achieve significantly better performance than the other systems. The performances of SC-EEC and SE-EEC are better than that of SC-EC. In particular, we can see that emotion similarity for sad is higher than that for other emotions on all systems. It may mean that using the same embedding for each frame has a limitation in maintaining emotion similarity, because the other emotional utterances have greater differences in prosody over time than sad utterances. We can see that the scores of sad similarity DMOS tests have overall the lowest values through all systems.

One interesting observation is that even though SC-EC achieves better performance than SC-EEC in the objective

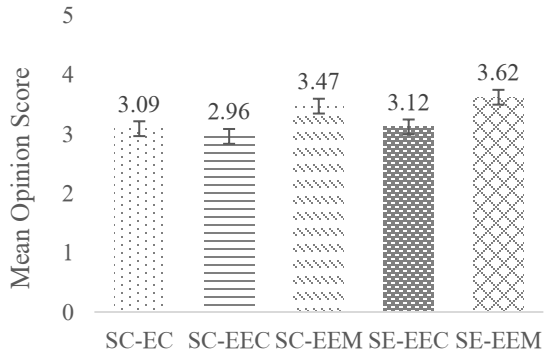

Fig. 2 Naturalness test result with their 95% confidence interval.

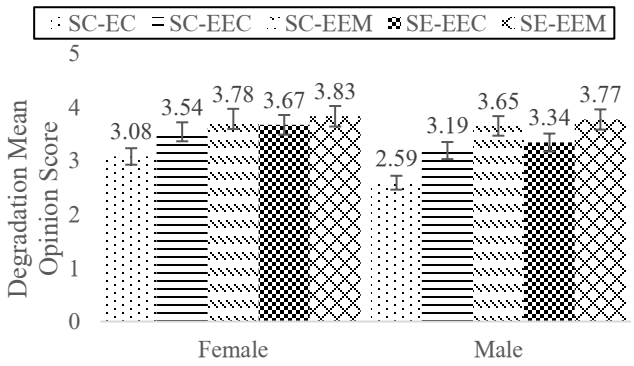

Fig. 3 Speaker similarity test result with their 95% confidence interval.

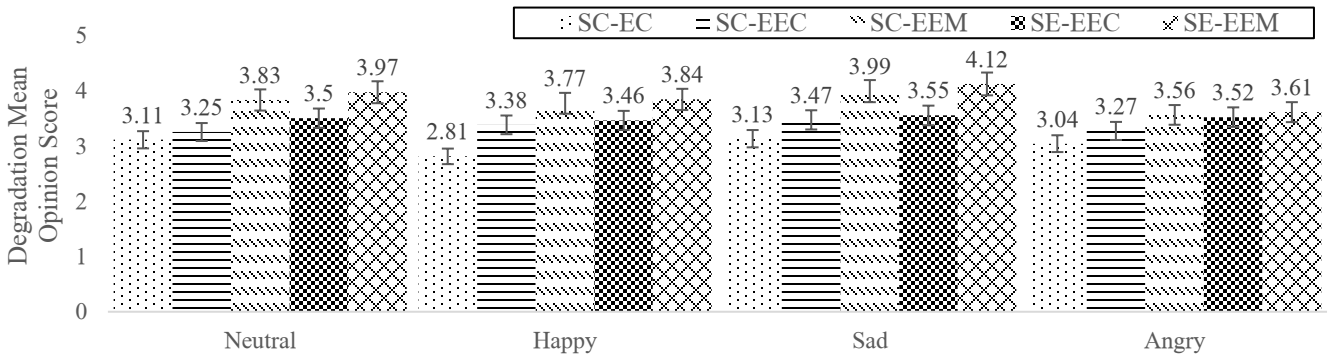

Fig. 4 Emotion similarity test result with their 95% confidence interval.

and naturalness evaluations, the subjective similarity results suggest that SC-EEC is significantly better than SC-EC. It implies that the emotion modeling with emotion code is trained to match the target speech signals, but in effect, is not trained to give the various expressional signal in comparison to the modeling with embedded emotion code.

For all subjective evaluation, the performance of SE-EEM is significantly better than that of other systems. These results suggest that the expressive speech from SE-EEM is the closest to the desired speaker and intended expression, and emotion representation from mel spectrogram can help to maintain the naturalness. Furthermore, SC-EEM achieves better performance compared with the other systems except SE-EEM, while SC-EEM shares a similar quality with SE-EEM by employing mel spectrogram to model emotion identity. We can also see that the systems using emotion representation from mel spectrogram achieve better performance than those using the emotion representation from emotion code for both objective and subjective evaluations. It implies that mel spectrogram greatly affects not only emotion modeling, but also speaker modeling.

4. CONCLUSION

In this paper, we have investigated the performance of multi-speaker emotional speech synthesis systems, according to speaker modeling method and emotion modeling method. Experimental results showed that the optimal architecture is CNN architecture using trainable speaker embedding and emotion representation from mel spectrogram which are used across the convolutional layers. By using emotion representation from mel spectrogram, we accomplish good quality for naturalness, speaker similarity, and emotion similarity.

In future works, we would like to investigate the potential and possible improvement of the exploiting embedding representation to input for different layers. Further, we will try to apply style modeling method in [23], which augment a style token layer to the reference encoder, and investigate how global style tokens correspond to emotional meaning, respectively. Also, we plan to enable WaveNet vocoder to generate better steady voice, and further investigate the speech synthesis system based on mel spectrogram.

5. ACKNOWLEDGMENT

This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (No. 10080667, Development of conversational speech synthesis technology to express emotion and personality of robots through sound source diversification).

6. REFERENCE

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

[2] H. Zen and H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” *Proc. ICASSP*, pp. 4470–4474, April. 2015.

[3] Y. Fan, Y. Qian, F. K. Soong and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," *Proc. ICASSP*, pp. 4475-4479, 2015.

[4] S. Pascual and A. Bonafonte, "Multi-output RNN-LSTM for multiple speaker speech synthesis and adaptation," *Proc. EUSIPCO*, pp. 2325-2329, 2016.

[5] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” *Neural computation* 9.8, pp. 1735-1780, 1997.

[6] N. Hojo, Y. Ijima, and H. Mizuno, “An Investigation of DNN-Based Speech Synthesis Using Speaker Codes,” *Proc. Interspeech,* pp. 2278-2282, September, 2016.

[7] Y. Zhao, D. Saito, and N. Minematsu, “Speaker Representations for Speaker Adaptation in Multiple Speakers’ BLSTM-RNN-based Speech Synthesis,” *Proc. Interspeech,* pp. 2268-2272, September, 2016.

[8] B. Li, and H. Zen, “Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis,” *Proc. Interspeech,* pp. 2468-2472, September, 2016.

[9] S. An, Z. Ling, and L. Dai, “Emotional statistical parametric speech synthesis using LSTM-RNNs,” *Proc. APSIPA ASC*, pp.1613-1616, 2017.

[10] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, “An investigation to transplant emotional expressions in DNN-based tts synthesis,” *Proc. APSIPA ASC*, pp.1253-1258, 2017.

[11] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron,” *arXiv preprint arXiv:1803.09047*, 2018.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: A fully end-to-end text-tospeech synthesis model,” *Proc. Interspeech*, pp. 4006–4010, 2017.

[13] S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[14] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent wavenet vocoder,*” Proc. Interspeech*, pp. 1118–1122, 2017.

[15] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep Voice 3: Scaling text-to-speech with convolutional sequence learning,” *Proc. ICLR*, 2018.

[16] S. Ioffe, and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *Proc. ICML*, pp. 448–456, 2015.

[17] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The htk book,” *Entropics Cambridge Research Lab*., 2002.

[19] D. Griffin, and Jae Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236-243, April 1984.

[20] A. Paszke, S. Gross, S. Chintala, and G. Chanan, “Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration,” 2017.

[21] D. P. Kingma, and J. L. Ba, “Adam: A method for stochastic optimization,” *Proc. ICLR*, 2015.

[22] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[23] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” *arXiv preprint arXiv:1803.09017*, 2018.

[24] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *Proc. ICLR*, 2016.

[25] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” *Proc. ICML*, 2017.