Emotional Speech Synthesis Based on Style Embedded Tacotron2 Framework

Ohsung Kwon[1], Inseon Jang[2], ChungHyun Ahn[2] and Hong-Goo Kang[1]
[1]*Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea*
[2]*Media Research Division, ETRI, Daejeon, Korea*
osungv@dsp.yonsei.ac.kr, {jinsn, hyun}@etri.re.kr, hgkang@yonsei.ac.kr

Abstract

In this paper, we propose a speech synthesis system that effectively generates multiple types of emotional speech using the concept of global style token (GST); where the emotion-related style information is presented by an additional style embedding vector. Although the GST is not a new idea, no one has been utilized the idea for an emotional speech synthesis task. We explicitly combine the GST idea with the Tacotron2 framework to implement an emotional text-to-speech system. The analysis results demonstrate that the proposed GST structure successfully transfers various types of emotional information to the synthesized speech. Subjective listening tests to evaluate the naturalness and emotional expression of synthesized speech are conducted to verify the superiority of the proposed algorithm.

Keywords: End-to-end text-to-speech, emotional speech synthesis, global style token, Tacotron2

1. Introduction

A text-to-speech (TTS) is a technique to convert input text into natural speech signal. This technique is essential for providing convenient user interface environment in human-computer interaction (HCI) system. For last few decades, TTS have been mainly designed with unit selection synthesis (USS) [1] and statistical parametric speech synthesis (SPSS) [2] frameworks. Due to advantages in terms of flexibility [3] and relatively small training data needed, the SPSS framework has grown in popular. The hidden Markov model [4] or deep neural network [5,6] are popular approaches in that category.

Recently, end-to-end TTS systems [7–10] that utilize the power of deep learning network itself opened a new research direction because they provide very high performance even without having strong domain expertise in speech processing and laborious design mechanisms in the SPSS framework.

With the success of the end-to-end TTS system, there is a need to develop expressive speech synthesis for various applications such as audiobook, broadcasting, stylish news narration and so on. For example, prosody [11] and style transfer [12,13] approaches were introduced on the Tacotron1 framework. Specifically, the concept of global style token (GST) is utilized to transfer diverse speaking styles of training speech corpus into synthesized speech [13]. Since the GST has not been applied to the domain of emotion expression yet, the detailed analysis on the effectiveness of the GST to emotional speech synthesis has not been verified.

In this paper, we propose an emotional speech synthesis system using the GST based Tacotron2 framework. Our contributions are as follows: 1) We combine the GST module with the Tacotron2 framework. 2) By applying the GSTs to emotional speech synthesis framework for the first time, we confirm the fact that the GSTs are helpful to synthesize emotional speech. The subjective experiment results prove the superiority of the proposed system.

2. Background

2.1. Tacotron2 framework

Tacotron1 framework, proposed as a complete toward end-to-end TTS system [7], directly generates mel-spectrogram from given input text sequence. Then, the mel-spectrogram is converted into linear-scale spectrogram, and eventually changed to the time-domain waveform using Griffin-Lim algorithm [14]. The principal merit of the Tacotron1 is its reasonable quality in terms of prosody generation.

Through these advances, Tacotron2 framework was designed for obtaining a higher synthesis quality while keeping the merit of the conventional version in the end-to-end framework [15]. It consists of spectrogram prediction network and

WaveNet vocoder. The spectrogram prediction network has a structure for an encoder-and-decoder with attention module. In the encoder network, the input text sequence is first changed to character embedding sequence, and then the embedded sequence is encoded to hidden linguistic sequence by three convolution layers followed by a bi-directional long short-term memory (LSTM) network. And then, the location sensitive attention module [16] passes the context vector to the two LSTM layers in the decoder network. In the decoder network, a pre-net and the two LSTM layers generate both low resolution mel-spectrogram and a stop token by passing hidden acoustic representation through a different linear projection network. Post-net improves the resolution of spectrogram by adding residual components. The spectrogram generation process is recursively operated at every decoding step until the stop token output is larger than the pre-defined threshold value. Completed mel-spectrogram is converted to the time-domain waveform by the WaveNet vocoder.

2.2. Global style token

GSTs are a bank of embedding vectors to capture hidden speaking style information in the reference audio signal. Style architecture including the GSTs consists of a reference encoder and a style token layer and is trained jointly with the Tacotron1 framework using the reconstruction loss [13]. The reference encoder extracts high-level prosody embedding vector from the input reference audio signal. And then, the multi-headed attention module [17] in the style token layer is trained to learn similarity between the prosody embedding and each GST. The extracted weights represent how much each GST contributes a specific speaking style of the synthesized speech. The GSTs learn hidden acoustic information related to speaking style. Finally, the linear combination of the GSTs and weights generates a style embedding vector. The style embedding is conditioned on the hidden linguistic features extracted by the encoder network in the Tacotron1 framework to reflect a specific speaking style in the synthesized speech.

3. Proposed GST-Tacotron2

As illustrated in Fig.1, the proposed system consists of a style architecture for emotion-related style embedding, and the Tacotron2 framework for a TTS task. The style architecture extracts style embedding vectors from the given reference emotional speech; whereas the Tacotron2 encoder generates hidden linguistic feature vectors from the given input text sequence. These vectors are then concatenated to compose the joint style-linguistic sequence, and it is passed through the Tacotron2 decoder to predict the mel-spectrogram. Finally, the WaveNet vocoder synthesizes the emotional speech waveform from the predicted mel-spectrogram.

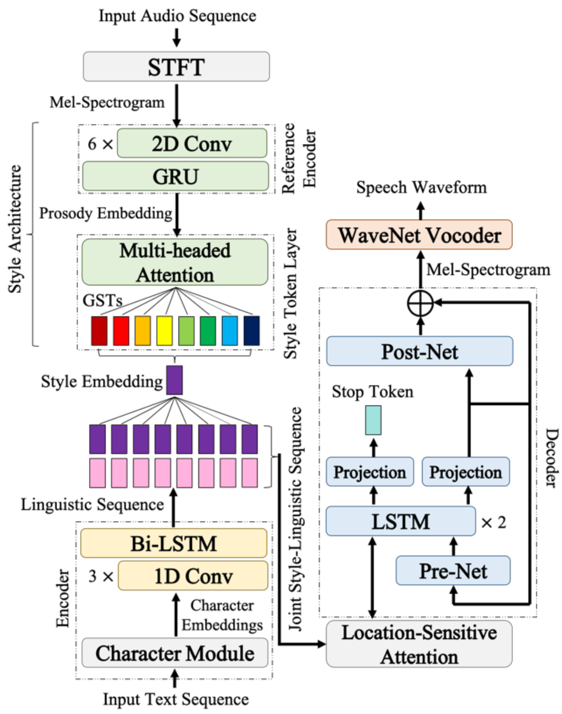

Fig.1: Block diagram of the proposed system.

4. Experiments

4.1. Database and features

A speech corpus recorded by a Korean male professional speaker was used for the experiments. The corpus consisted of four emotions: happy, angry, sad, and neutral, each with approximately 800 utterances, composing 2,956 utterances in total (3.9 hours). The speech signals were sampled at 16 kHz and each sample was quantized by 16 bits. In total, 2,668 utterances (3.5 hours) were used for training and 288 utterances (0.4 hours) for testing.

In the training phase, text and audio data were respectively input to the Tacotron2 encoder and the style architecture to generate a joint style-linguistic feature sequence. Then, the Tacotron2 decoder decoded the encoded sequence to predict the mel-spectrogram. Finally, predicted spectrogram was converted into emotional speech waveform by the WaveNet vocoder. At the inference phase, reference audio signal is used to generate emotion-related style embedding to be conditioned on the transcript embedding.

Table 1: Subjective MOS test results with a 95% confidence interval for the various emotional speech types.

System	Happy	Angry	Sad
Recorded	4.68 ± 0.22	4.69 ± 0.22	4.54 ± 0.31
Synthesized	3.62 ± 0.35	2.99 ± 0.37	2.66 ± 0.29

Table 2: Subjective emotion evaluation results with a 95% confidence interval. The following five-point responses were used (1: Completely different, 2: Slightly different, 3: Slightly similar, 4: Almost the same, and 5: Completely identical).

System	Happy	Angry	Sad
Synthesized	3.51 ± 0.28	3.26 ± 0.17	3.58 ± 0.16

4.2. Generation performance

We evaluated the perceptual quality of the proposed method by using the mean opinion score (MOS). Ten native Korean listeners were asked to make quality judgements about the synthesized speech using the following five-point MOS responses (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent). In total, 36 utterances were randomly selected from the test set and were then synthesized by using the proposed GST framework. The MOS test results summarized in Table 1 confirm that the proposed synthesis system provided reasonable quality even if the amount of training speech corpus was relatively small. These results need to be further investigated by focusing on the characteristic difference of each emotion and the quality variation depending on the size of the speech corpus.

4.3. Emotional speech synthesis

The quality of the synthesized emotional speech was also evaluated by using the MOS test. The test setups were the same as the normal MOS test as presented in the previous section, but the listeners were asked to evaluate the likeability of the emotional expression in the synthesized speech. Specifically, one- or two-point scores should be awarded if the listeners felt that the synthesized speech expressed a different type of emotion than the intended emotion. The evaluation results shown in Table 2 demonstrate that the proposed GST-based approach effectively generates various types of emotional speech.

5. Conclusion

In this paper, we introduced a GST-based emotional speech synthesis scheme implemented in the end-to-end TTS system, Tacotron2. The GSTs in the style architecture were helpful to synthesize emotional speech in the end-to-end paradigm. The subjective evaluation results demonstrated the feasibility of using the proposed approach.

Future work includes further investigating the relationship between the GST and each emotion, and the quality variation depending on the size of training corpus.

6. Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (2015-0-00860, Development of assistive broadcasting technology for invisible and deaf people's media accessibility)

References

[1] Andrew Hunt and Alan W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[2] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol.51, no. 11, pp. 1039–1064, 2009.

[3] Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP*, 2001, pp. 805–808.

[4] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[5] Heiga Ze, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[6] Yuchen Fan, Yao Qian, Feng-Long Xie and Frank K. Soong, "TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks", in *Proc. INTERSPEECH*, 2014, pp.1964–1958.

[7] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp.4006–4010.

[8] A ̈aron 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 rawaudio.," in *arXiv:1609.03499, 2016.*

[9] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joo F. Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, “Char2Wav: End-to-end speech synthesis,” in *ICLR workshop submission,* 2017.

[10] Sercan˙O. Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, XianLi, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi, “Deep voice: Real-time neural text-to-speech,” in *Proc. ICML,* 2017, pp. 195–204.

[11] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, RobClark, and Rif A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in *Proc. ICML*, 2018, pp.4693–4702.

[12] Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stan-ton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A. Saurous, “Uncovering latent style factors for expressive speech synthesis,” in *Proc. NIPS ML4Audio Workshop*, 2017.

[13] 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,” in *Proc. ICML*, 2018, pp.5180-5189.

[14] Daniel Griffin and Jae Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol.32, no. 2, pp. 236–243, 1984.

[15] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif. A Saurous, Yannis Agiomyrgiannakis, and Yonghui We, “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018, pp. 4779–4783.

[16] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, “Attention-based models for speech recognition,” in *Proc. NIPS*, 2015, pp. 577–585.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Proc NIPS*, 2017, pp. 5998–6008.