세밀한 감정 음성 합성 시스템의 속도와 합성음의 음질 개선 연구

[a)]*엄세연 *오상신 **장인선 **안충현 *강홍구

*연세대학교 전기전자공학과

**한국전자통신연구원 미디어연구본부

[a)]syum@dsp.yonsei.co.kr

A study on the improvement of generation speed and speech quality for a granularized emotional speech synthesis system

*Um, Se-Yun *Oh, Sangshin **Jang, Inseon **Ahn, Chung-hyun *Kang, Hong-Goo

* Yonsei University, Department of Electrical and Electronic Engineering, Seoul, South Korea

** Electronics and Telecommunications Research Institution, Daejeon, South Korea

요약

본 논문은 시각 장애인을 위한 감정 음성 자막 서비스를 생성하는 종단 간(end-to-end) 감정 음성 합성 시스템(emotional text-to-speech synthesis system, TTS)의 음성 합성 속도를 높이면서도 합성음의 음질을 향상시키는 방법을 제안한다. 기존에 사용했던 전역 스타일 토큰(Global Style Token, GST)을 이용한 감정 음성 합성 방법은 다양한 감정을 표현할 수 있는 장점을 갖고 있으나, 합성음을 생성하는데 필요한 시간이 길고 학습할 데이터의 동적 영역을 효과적으로 처리하지 않으면 합성음에 클리핑(clipping) 현상이 발생하는 등 음질이 저하되는 양상을 보였다. 이를 보안하기 위해 본 논문에서는 새로운 데이터 전처리 과정을 도입하였고 기존의 보코더(vocoder)인 웨이브넷(WaveNet)을 웨이브알엔엔(WaveRNN)으로 대체하여 생성 속도와 음질 측면에서 개선됨을 보였다.

1. 서론

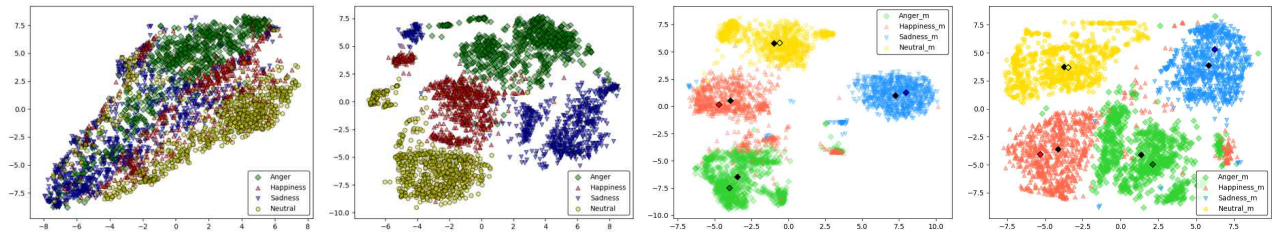
최근 급격하게 발전 중인 딥 러닝(deep-learning) 기법으로 인해 이를 활용하여 개발된 음성 합성 시스템은 입력 텍스트에 포함된 문맥의 의미를 명확하게 전달할 수 있는 음성을 합성할 수 있었다. 더 나아가 운율(prosody)과 관련된 음향적 특성(acoustic feature)인 피치(pitch), 강세(stress), 말하는 속도 등을 활용하여 사람의 음성과 유사한 생동감 있는 음성을 제공할 수 있다 (e.g. Tacotron [1]).

Skerry-Ryen et al.[2]은 GRU(Gated Recurrent Unit) cell로 구성된 참조 인코더(reference encoder)를 통해 참조 오디오의 멜 스펙트로그램(mel-spectrogram)에서 운율 임베딩(prosody embedding)을 추출하고, 이를 텍스트 인코더(text encoder)의 임베딩과 결합함으로써 감정을 표현할 수 있는 음성을 합성하였다. Wang et al. [3]는 참조 인코더를 활용하여 전역 스타일 토큰들의 가중 합으로 운율 임베딩을 측정하였고 각 스타일 토큰들의 가중치를 조절함으로써 합성 음성의 운율을 제어할 수 있었다. 이와 관련하여 Kwon et al. [4]은 동일한 감정(운율)을 나타내는 전역 스타일 토큰들의 가중치들이 서로 클러스터(cluster)된다는 사실을 기반으로 각 감정 클러스터의 평균을 계산하여 표현하고자 하는 감정을 합성된 음성에 명확하게 나타내었다. 한편, Um et al .[5] 은 세밀한 감정 표현을 위해 감정 클러스터의 특성을 고려한 SA-I2I (Spread Aware Inter-to-Intra distance ratio)를 제시하여 감정의 세기를 조절하였다. 그러나, 음성을 합성하기 위해 필요한 시간이 길고, 감정의 세기가 강해질수록 합성음의 음질이 저하되는 등의 문제가 있었다.

본 논문에서는 SA-I2I 알고리즘의 단점이었던 느린 합성 속도와 음질의 저하를 개선하기 위한 방법으로 각 감정 데이터의 특성을 고려한 데이터 전처리 과정을 추가하였고, WaveNet[6] 뉴럴 보코더 알고리즘을 WaveRNN[7] 으로 대체하였다. 효과적인 전처리 과정을 통해 합성음의 음질을 저하시키는 클리핑 현상을 제거하였고, WaveRNN을 통해 기존 알고리즘보다 매우 빠른 합성 속도를 얻었다. 두 모델의 성능 차이를 비교하기 위한 청취 평가를 통해 제안한 방법의 우수성을 입증하였다.

2. 본론

본 논문에서는 음성 합성 시스템에 전역 스타일 토큰을 추가한 GST-Tacotron을 기준 모델로 사용하였으며 보코더로 WaveRNN을 사용하였다. 타코트론 모델은 텍스트를 입력받아 인코더를 통해 텍스트 임베딩으로 변환한 뒤, 타겟(target) 오디오의 운율 정보를 갖고 있는 스타일 임베딩과 결합하여 어텐션(attention) 모듈로 입력하고 디코더(decoder)와의 얼라인먼트(alignment)를 학습하여 보코더의 입력인 멜 스펙트로그램을 생성한다. 이때 스타일 토큰 레이어(style token layer, STL)에서 레퍼런스 인코더(reference encoder)가 타겟 오디오의 멜 스펙트로그램을 입력받아 운율 정보를 모델링하고 멀티 헤드(multi-head) 어텐션을 통해 전역 스타일 토큰들과의 유사도를 측정한 후, 전역 스타일 토큰들의 가중 합으로 스타일 임베딩을 생성한다.



[그림 1] 스타일 임베딩의 가중치 벡터들에 대한 t-SNE. 왼쪽부터 순서대로: 1)데이터 전처리 적용 전, 2)데이터 전처리 적용 후, 3)SA-I2I 알고리즘으로 구한 각 감정의 대푯값(감정 클러스터 안의 짙은 색) 및 평균값(검은색), 4)기존 모델의 대푯값 및 평균값.

WaveRNN은 GRU cell과 FC(Fully Connection) layer로 구성된 자기 회귀 적 생성 모델(autoregressive generative model) 이다. 기존 모델에서 사용된 WaveNet의 여러 층의 dilated causal convolution layer를 GRU로 대체하였기 때문에 모델의 크기와 연산량이 줄어들어, 결과적으로 음성 샘플을 기존 모델보다 빠르게 합성할 수 있는 장점이 있다.

본 논문에서는 기존 모델의 문제점인 클리핑 현상을 제거하기 위한 데이터 전처리과정으로 학습에 사용할 네 가지 감정 (행복, 분노, 슬픔, 중립) 데이터 셋 (dataset)의 신호 크기 (scale)를 -3dB (decibel, 데시벨)로 조정하였다. 모든 동적 영역을 표현하기 위해 학습 데이터의 크기를 0dB로 맞춰진 경우에는, 모델에서 추정된 멜 스펙트로그램의 크기가 0dB를 초과할 수 있고, 이에 따라 클리핑 현상이 발생하기 때문이다. 이때 모든 감정들의 크기를 동일하게 감소시키면, [그림 1]의 1)에서 나타나듯이 학습된 GST-Tacotron모델이 각 감정을 뚜렷하게 구분하지 못하여 각각의 감정이 클러스터를 생성하지 못하는 현상이 발생한다. 이는 음성의 크기가 각 감정을 구별하는 하나의 특징으로 작용하기 때문이며, 이를 해결하기 위해 모든 데이터 셋의 크기를 동일한 값으로 감소시키지 않고 각 감정 음성에 따라 상대적으로 크기를 조절하였다. 그 결과, [그림 1]의 2)와 같이 학습된 모델이 네 가지 감정을 구분하여 각 감정에 대한 클러스터를 효과적으로 생성하였다.

앞서 제시한 두 모델의 성능을 평가하기 위해 두 가지 실험을 진행하였다. 먼저 합성음의 생성 속도를 비교하기 위해 각 감정 클러스터의 평균값을 스타일 임베딩으로 사용하여 동일한 문장을 합성하는데 소요되는 시간을 측정하였다. [표 1]은 네 가지 감정에 대한 기존 모델과 새로 제안한 모델의 감정 음성 합성의 속도를 나타낸다. 합성음의 길이는 약 1.2초 이며 동일한 환경에서 실험했을 시, 네 가지 감정에 대해 모두 WaveRNN이 WaveNet에 비해 평균적으로 약 8.04배 빠르다는 것을 알 수 있다. 즉, 기존 모델의 첫 번째 문제점인 생성 속도를 개선하여 모델의 성능을 향상시켰다고 볼 수 있다. 두 번째 실험으로 10명의 청취자들을 대상으로 합성음의 음질에 대한 선호도 평가를 실시하였다. 각각의 모델에서 합성한 두 개의 음성 샘플 중에 음질이 더 뛰어난 샘플을 선택하고 만약, 두 샘플의 음질에 뚜렷한 차이가 없다면 '차이 없음'을 택하도록 지시하였다. 각 감정별로 3개의 합성음을 비교하였으며 결과는 [표 2]에 나타내었다. 네 가지 감정 모두 제안된 모델의 합성음에 대한 선호도가 평균적으로 약 70%로 압도적으로 높음을 보여준다. 이는 기존 모델과 다르게 새로 제안한 데이터 전처리 과정을 통해 클리핑 현상을 제거했기 때문이다. 위의 두 가지 실험을 바탕으로 새롭게 제안된 모델이 기존의 모델과 비교했을 때, 합성음의 생성속도가 빨라지고 합성음의 음질 또한 개선됨을 확인할 수 있다.

[표 1] WaveNet과 WaveRNN의 합성음 생성 속도
(GeForce GTX 1080 Ti GPU 기준)

감정	행복	분노	슬픔	중립	평균
WaveNet	63 초	60 초	62 초	66 초	62.75 초
WaveRNN	7.8 초	7.8 초	7.8 초	7.8 초	7.8 초

[표 2] WaveNet과 WaveRNN의 합성음 음질에 대한 청취 평가 결과

감정	행복	분노	슬픔	중립	평균
기존 모델	13.3	26.7	30.0	0.0	17.5
차이 없음	13.3	16.7	16.7	10.0	14.175
제안 모델	77.3	56.7	53.3	90.0	69.325

3. 결론

본 논문에서는 GST-Tacotron과 WaveRNN을 사용하여 감정 음성을 합성하는 연구를 진행하였다. 합성음의 품질을 향상시키기 위해 데이터의 특성을 고려한 전처리 방법을 제안하였고, 음성 합성 속도를 높이기 위해 기존의 사용한 WaveNet을 상대적으로 모델의 구조가 얕고 연산량이 적은 WaveRNN으로 변경하였다. 이를 통해 클리핑 현상이 개선되고, 합성음을 생성할 때 필요한 시간이 줄어듦을 보였다.

감사의 글

본 연구 논문은 과학기술정보통신부 및 정보통신기획평가원의 출연금으로 수행하고 있는 한국전자통신연구원 시청각 장애인의 방송시청을 지원하는 감성표현 서비스 개발[2019-0-00447]의 연구결과입니다.

4. Reference

[1] 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.

[2] YuxuanWang, 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.

[3] 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.

[4] O. Kwon, I. Jang, C. H. Ahn, and H. -G. Kang, “Emotional speech synthesis based on style embedded Tacotron2 framework,” Proc. ITC-CSCC, 1-4 (2019).

[5] Um, Se-Yun, et al. "Emotional Speech Synthesis with Rich and Granularized Control." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[6] Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).

[7] Kalchbrenner, Nal, et al. "Efficient neural audio synthesis." arXiv preprint arXiv:1802.08435 (2018).