Tacotron 기반 이미지 스타일 변환을 이용한 감정 표현 음성합성

문성우, 김성현, 최용훈
광운대학교
msw3322@gmail.com, waverdeep@kw.ac.kr, yhchoi@kw.ac.kr

Tacotron-based Emotional Speech Synthesis Using Image Style Transfer

Sung-Woo Moon, Sung-Hyun Kim, Yong-Hoon Choi
KwangWoon University

요 약

본 논문에서는 Tacotron 2 기반의 이미지 스타일 변환을 이용한 특징 추출 음성합성 모델인 멜-스펙트로그램 이미지 스타일 변환 (mel-spectrogram image style transfer) Tacotron을 제안한다. 제안하는 방법은 화자에 대한 특징을 추출하는 새로운 방법으로, 오디오의 멜 스팩트로그램을 이미지화시킨 후 이미지 스타일 변환을 거쳐 특징을 추출하는 방법이다. 특징 추출에 집중하기 위해 감정 상태와 화자의 ID 값을 입력값으로 사용했다. 실험결과 gross pitch error (GPE), voicing decision error (VDE), F0 frame error (FFE)를 확인했을 때 제안하는 기법의 성능이 기존의 모델인 global style token (GST) Tacotron과 Variational Autoencoder (VAE) Tacotron보다 오차값이 낮음을 확인했다. 음원에 대한 주관적인 평가에서 제안한 모델의 mean opinion score (MOS) 값이 가장 큼을 확인했다.

Ⅰ. 서 론

음성합성 기술의 발전으로 화자의 감정과 스타일이 담긴 대화체 음성합성에 대한 연구가 활발히 이루어지고 있다 [1]. 하지만 기존의 방식은 표현할 수 있는 감정의 수가 3-4개 정도로 제한적이고, 다양한 화자의 스타일을 표현할 때는 높은음과 낮은음에 대한 표현이 부족하여 표현이 어색한 음원을 만들 때가 있다. 또한 학습 과정에서 감정, 말하는 속도, 리듬 강세와 같은 특징들이 학습되지 않을 수 있는 단점이 있다.

본 논문에서는 화자의 특징을 더욱 뚜렷하게 표현하고, 기존보다 많은 감정을 표현할 수 있는 멜-스펙트로그램 이미지 스타일 변환 Tacotron을 제안한다. 제안 기법은 이미지 스타일 변환 기법 중 스타일 변환에 대한 기법과 Tacotron이 합쳐진 end-to-end text-to-speech (TTS) 모델이다. 이미지 스타일 변환 기법은 참조 오디오의 멜-스펙트로그램을 스타일 변환 시켜 참조 오디오의 특징을 학습에 사용하는 방법이다. 참조 오디오를 멜-스펙트로그램으로 변환 후 이미지 스타일 변환을 위해 이미지 파일로 만들었다. 이미지에서의 스타일을 추출하기 위해 미리 학습된 VGG19모델을 사용하여 스타일 변환 된 이미지를 추출했고, 추출된 이미지를 통해 화자의 특징값을 추출했다.

본 논문에서 제안하는 모델의 검증을 위해 단일화자 음성합성에 대한 실험과 다 화자에 대한 음성합성 실험을 진행했다. 단일화자 실험인 경우 참조 오디오와 비슷한 스타일을 가지며 감정표현이 적절한지 확인했다. 다 화자에 대한 음성합성 실험은 참조 오디오에 대한 스타일과 감정표현이 적절히 이루어지는지 확인했고 참조 오디오의 목소리와 같은지 확인했다.

본 논문에서 제안하는 기법의 실험에 대한 검증을 위해 객관적 평가와 주관적 평가를 했다. 객관적 평가는 오디오의 F0를 이용한 오차 gross pitch error (GPE), voicing decision error (VDE), F0 frame error (FFE)를 확인하여 모델의 성능을 평가했다. 주관적 평가는 만들어진 음원이 화자의 스타일과 그에 맞는 감정표현이 적절히 이루어 졌는지 Mean opinion score (MOS) 평가 방법을 이용하여 평가를 진행했다.

Ⅱ. 본론

2.1 멜-스펙트로그램 이미지 스타일 변환

스타일 전송(style transfer)은 두 개의 이미지 (컨텐츠 이미지, 스타일 이미지)가 주어졌을 때 주된 이미지의 형태는 컨텐츠 이미지를 유지하면서 이미지의 스타일만 우리가 원하는 스타일 이미지와 유사하게 바뀌는 것을 말한다. 멜-스펙트로그램에서 스타일 이미지를 얻는 과정은 다음과 같다. 이미지에서 스타일을 추출할 수 있기 때문에 멜-스펙트로그램을 이미지로 변환시켰다. 음성 데이터에서 각 오디오 클립의 길이가 다르기 때문에, 멜-스펙트로그램 이미지의 크기도 오디오 클립 길이에 비례하여 생성했다. 생성된 멜-스펙트로그램 이미지는 사전 훈련된 CNN 구조를 가진 이미지 분류 모델의 6개 층을 통과하며, 그램 행렬은 각 층의 특징에 대해 얻어진다.

층 k에서의 전체 손실 E_k는 멜-스펙트로그램의 원 이미지 $\boldsymbol{a}$를 스타일 변환한 M_{ij}^k와 노이즈 이미지 $\boldsymbol{x}$를 업데이트시켜 생성한 이미지 G_{ij}^k를 이용해 다음과 같이 구하고

$$E_k = \sum_{ij}(G_{ij}^k - M_{ij}^k)$$

전체 스타일 손실은 k번째 층에 가중치를 줄 수 있는 W_k를 곱해 다음과 같이 구할 수 있다. 본 논문에서는 모두 1을 곱해 실험을 진행했다.

$$L_{style}(\boldsymbol{a},\boldsymbol{x}) = \sum_{k=0}^{L}(W_k E^k).$$

2.2 멜-스펙트로그램 이미지 스타일 변환

멜-스펙트로그램을 만드는 기본 틀로 sequence-to-sequence (seq2seq) 구조를 가지고 있는 Tacotron 2를 사용했다. 제안하는 모델의 전체적인

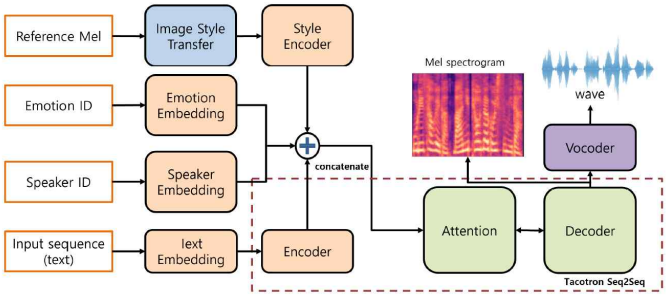

그림 1. 제안 모델의 블록 다이어그램

fig 1. Block diagram of the propoesd model

구조는 그림 1과 같다. 화자의 스타일을 추출하기 위해 멜-스펙트로그램에 이미지 스타일 변환기법을 사용해 스타일이 담긴 스타일 이미지를 생성했다. 생성된 스타일 이미지를 Style Encoder의 입력으로 사용해 스타일 이미지에서 주요한 특징값들을 추출했다. Style Encoder의 구조는 6개의 합성곱 층으로 구성되어 있고 층마다 배치 정규화, 활성화 함수로 ReLU를 사용하였다. 합성곱 층을 통과한 특징값은 gated recurrent unit (GRU)를 통해 시간적 순서를 고려한 특징값을 뽑아낸다. 화자의 스타일 특징을 학습하는 것이 목표이기 때문에 감정에 대한 학습은 감정이 레이블링 된 값을 임베딩하여 사용했다. 한국어의 경우 음소 단위 입력이 각 소리의 문자 단위 입력보다 좀 더 명료하고 화자의 음색이 더 살아나는 경향이 있어 입력 텍스트를 모두 음소 단위로 바꿔 입력으로 사용했다. 레이블링 된 화자의 ID값을 임베딩 시켜 임베딩된 감정 ID 값과 인코더를 통과한 텍스트, Style Encoder를 거쳐 만들어진 특징값을 모두 연결해 Attention 모듈의 입력으로 사용했다. 이 방법을 통해 화자 목소리와 감정은 참조 음원으로부터 독립되고 참조 음원으로부터 화자의 말하는 속도, 리듬 강세와 같은 특징만을 뽑아올 수 있게 된다.

Ⅲ. 실험

3.1 실험환경

30대 전문 여성 성우 1인이 7가지 감정(중립, 행복, 슬픔, 화남, 역겨움, 놀람, 공포)에 대해서 각각 3,000개의 발화로 구성되어 있고, 녹음 분량은 약 30시간이다. 다 화자 데이터는 Pitchtron논문 [2]에서 제공하는 데이터셋을 사용했다. 일반 낭독체와 대화체가 섞여 있는 데이터로 감정(일반, 기쁨, 화남, 슬픔) 대본을 사용한 녹음 데이터도 포함되어있다. 약 24,000개의 발화로 구성되어 있고 녹음 분량은 약 34시간이다. 두 개의 데이터 모두 주파수는 22,050Hz를 가지고 있지만 Tacotron 2에서 16,000Hz로 실험을 진행했기 때문에 기존의 주파수를 16,000Hz로 바꿔 실험을 진행했다. 데이터별로 검증용 데이터로 100개, 테스트를 위한 데이터로 500개를 사용했다. 테스트 데이터에는 여러 감정과 여러 명의 화자가 섞여 있게 구성했다. 600개를 제외한 모든 데이터는 학습 데이터로 사용했다.

3.2 실험결과

표 1은 기존의 방법과 제안한 방법의 오차율을 나타낸 객관적 평가 결과표이다. 제안 모델을 GST Tacotron, VAE Tacotron과 비교했다. 결과를 보면 제안한 모델이 GPE, VDE, FFE 모두 기존의 모델보다 좋은 모습을 확인할 수 있다. 제안하는 모델이 참조 음원으로부터 스타일을 잘 추출해 가장 작은 오차 값을 가지는 것이라 판단된다. 표 2는 스타일과 감정에 대한 MOS이다. 표 2를 보면 VAE의 MOS값이 GST의 MOS 값보다 더 높은 점수를 받은 것을 확인할 수 있다. GST가 다화자 한국어 데이터 셋에 대한 학습시킬 때 손실 값이

표 1. 단일화자와 다화자에 대한 객관적 평가

Table 1. Objective evaluations for single speaker and multi-speaker

	Single speaker			Multi-speaker		
	GPE(%)	VDE(%)	FFE(%)	GPE(%)	VDE(%)	FFE(%)
GST	16.13	34.37	50.50	19.72	32.92	52.64
VAE	16.51	35.00	51.51	24.46	36.67	61.14
Ours	**12.87**	**29.48**	**42.35**	**18.67**	**32.30**	**50.97**

표 2. 스타일과 감정에 대한 단일화자 다화자의 MOS 평가

Table 2. MOS evaluations for single speaker and multi speaker about style and emotion

	Single speaker		Multi-speaker	
	Style MOS	Emotion MOS	Style MOS	Emotion MOS
GT	4.37(±0.18)	4.02(±0.24)	4.51(±0.18)	4.15(±0.20)
GST	3.54(±0.25)	2.43(±0.20)	2.93(±0.15)	2.86(±0.21)
VAE	3.66(±0.23)	3.33(±0.16)	3.38(±0.19)	3.20(±0.19)
Ours	**3.88(±0.23)**	**3.93(±0.19)**	**3.93(±0.20)**	**3.86(±0.20)**

NaN으로 표출되어 학습이 적절히 이루어지지 못했고, 참조 음원과 다른 목소리로 나와 점수가 낮게 나온 것이라 판단된다. 주관적인 평가는 객관적 평가에서 사용되었던 VDE에 대한 것을 판단하기 어렵다. 그렇기 때문에 주관적 평가에서 높은 오차 점수를 받은 VAE 모델이 더 높은 MOS 값을 얻은 것이라 판단된다. 제안하는 모델의 경우 화자의 감정과 화자에 대한 ID 값을 입력으로 사용해 화자에 따른 감정을 정확하게 표현할 수 있음을 확인했다. 또한 GST와 VAE에서는 높은 음역대를 표현하지 못하지만 제안하는 모델에서는 높은음까지 나타내며 자연스러운 발화가 가능하다. 화자의 감정과 화자에 대한 ID 값을 입력으로 사용해 화자의 특징이 적절히 전달된 것이라 판단된다.

Ⅳ. 결론

본 논문에서는 Tacotron 2기반 멜-스펙트로그램 이미지 스타일 변환을 이용한 seq2seq 모델을 제안했다. 실험 결과 F0를 이용한 오차를 확인했을 때 제안한 모델의 성능이 기존의 모델보다 오차율이 낮음을 확인했다. 또한 만들어진 음성을 들어봤을 때 기존의 모델보다 화자의 특징을 더욱 뚜렷하게 표현하는 것을 확인할 수 있었다. 화자의 감정 표현 측면에서도 기존의 방법들에 비해 감정표현을 더 확실히 하는 것을 확인했다.

추후 연구로는 최신의 이미지 스타일 변경 방법을 사용해 화자의 스타일을 더욱 뚜렷하게 잘 전달하는 방법에 대한 연구를 진행할 것이다.

ACKNOWLEDGMENT

이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1064080).

본 연구(혹은 프로젝트)는 산업통상자원부의 산업기술혁신사업으로부터 지원을 받아 수행된 연구(No. 10080667, 음원 다양화를 통하여 로봇의 감정 및 개성을 표현할 수 있는 대화음성합성 원천기술 개발)의 결과물인 공개 음성 데이터베이스를 사용하였음.

참 고 문 헌

[1] Wang, Yuxuan, et al. "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis." International Conference on Machine Learning. PMLR, 2018.

[2] Jung, Sunghee, and Hoirin Kim. "Pitchtron: Towards audiobook generation from ordinary people's voices." arXiv e-prints (2020): arXiv-2005.