멜-셉스트럴 왜곡(MCD)를 활용한 딥러닝 기반 목소리 합성 기술의 성능 평가 연구

한재상[1], 이하연[2], 강윤서[3], 나상우[4,]
[1234]숭실대학교 소프트웨어학부
hhncn4471@soongsil.ac.kr, papepopepe@gmail.com, robbyra@gmail.com,
ysys1130@gmail.com

A performance evaluation study of a deep learning-based voice synthesis technique using Mel-Conceptual Distortion (MCD).

Jaesang Han[1], Yunseo Kang[2], Sangwoo Na[3] Hayeon Lee[4]
[1234]School of Software, Soongsil University

요 약

노래 음성 변환(Singing Voice Conversion, SVC)은 오디오 처리 분야에서 최근 활발히 연구되는 분야 중 하나로, 원래의 멜로디와 가사를 유지하면서 소스 가수의 노래 음성을 대상 가수의 음성으로 변환하는 것을 목표로 한다. 본 논문에서는 딥러닝 기반 SVC 모델을 중심으로 멜 셉스트럴 왜곡 지표를 활용해 모델 간 성능 평가를 진행한다. 이를 통해 엔터테인먼트, 교육 등 분야에서 효율적인 SVC 모델을 찾아 활용할 수 있을 것이다.

1. 서론

최근 인공지능 기술의 발전으로 딥러닝 기반 목소리 합성 기술(Singing Voice Conversion, 이하 SVC)에 관한 연구가 활발히 진행되고 있다. [1] SVC는 멜로디나 가사를 유지한 채로 대상 가수의 목소리를 원본 가수의 목소리로 변환하는 기술이다. 다양한 분야에서 SVC 기술이 활용되고 있지만, SVC 모델의 합성 결과에 대한 평가는 여전히 정성적 측면에 의존하는 경향이 있다. 본 연구에서는 멜 셉스트럴 왜곡(Mel Cepstral Distortion, 이하 MCD)을 이용하여 딥러닝 기반 SVC 모델 간의 정량적 성능 평가를 수행하고, 가장 효과적인 모델을 결정하고자 한다.

2. 연구 방법

2.1 모델 선정

본 논문에서는 비교를 위해, 최근 연구(2021년~)에서 공개된 네 가지의 SVC 모델을 선정하였다.

Model	핵심 키워드
FastSVC	Feature-wise Linear Modulation
DiffSVC	Diffusion probabilistic models
Assem-VC	Any-to-many non-parallel VC system
StarGANv2-VC	Generative Adversarial network, GAN

각 모델은 공통적으로 딥러닝 모델을 활용해 음성을 합성하지만, 변환 과정에서의 기술 구조와 작동 방식은 상이하다. 음성 합성 품질 간의 차이를 비교하기 위해 서로 다른 기술에 기반한 SVC 모델을 선택하였다.

2.2 연구 설계

데이터셋으로 공개 음성 데이터셋을 사용하였으며, 원본 가수의 음성인 source, 대상 가수의 음성인 target 파일로 구분하여 실험을 진행하였다. 또한 모든 음성 파일은 비손실 압축 형식을 따르는 .wav 파일을 사용하였다. MCD 성능 측정은 Ubuntu 20.04 가상 환경에서 python3를 활용해 진행하였다.

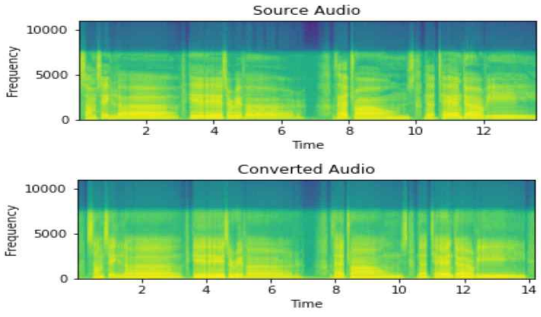

그림 1 Spectograms - Source/Converted

2.3 성능 비교 방법

먼저 음성 데이터를 특징 벡터화하기 위해 각 프레임에서 Mel-cepstral-coefficients(MCEPs)를 추출하였다. 이후 성능 비교를 위해 멜 - 셉스트럴 왜곡(Mel Cepstral Distortion, 이하 MCD)을 계산하였다. MCD는 두 음성 신호 사이의 왜곡을 측정하기 위한 지표로서, 원본 음성과 변환된 음성의 차이를 객관적인 수치로 나타낼 수 있다. [2] MCD 값은 다음과 같은 식을 통해 얻을 수 있다.

$$MCD(v^{targ}v^{ref}) = \frac{\alpha}{T}, \sum_{t=0}^{T-1}\sqrt{\sum_{d=s}^{D}(v_d^{targ}(t) - v_d^{ref}(t))^2}$$

$$\alpha = \frac{10\sqrt{2}}{\ln 10} = 6.14185$$

합성된 두 오디오를 비교하는 과정에서는 Dynamic Time Warping(DTW)을 사용하였다. [3] DTW는 두 개의 시퀀스 간 최적 정렬을 찾는 데에 이용되는 기법이다. 만약 두 음성 신호 간 유사성을 유클리드 거리로 측정하는 경우, 같은 시간 선상에 대해서만 거리를 계산하여 MCD 점수에 불이익이 생길 수 있다. 따라서 정확한 MCD 점수를 얻기 위해 각 음성 신호의 DTW를 계산하였다.

3. 연구 결과

Method	MCD Score (dB)
FastSVC	11.16084
DiffSVC	11.14365
Assem-VC	9.76408
StarGANv2-VC	10.97964

표 3 SVC 모델의 MCD 점수

분석 결과 모델 Assem-VC의 MCD Score가 가장 낮았다. MCD의 값이 낮을수록 왜곡이 적은 것을 의미하므로, 더 높은 품질의 합성 결과로 판단할 수 있다. 이는 Assem-VC의 음성 합성 품질이 다른 모델보다 높음을 의미한다. [3] Assem-VC의 경우 PPG-VC, Cotatron-VC, Mellotron-VC 등 기존 VC의 장점을 접목해 고안된 모델이란 점에서 가장 뛰어난 성능을 보인 것으로 추측할 수 있다. 한편, StarGANv2-VC는 두 번째로 낮은 MCD 값을 가졌으며, 서로 근소한 차이를 보인 DiffSVC와 FastSVC가 뒤를 이었다. FastSVC의 경우 실시간성에 중점을 둔 SVC 모델이기에 음성 합성 품질이 가장 낮았던 것으로 해석된다.

4. 결론

본 논문에서는 딥러닝 기반 SVC 모델의 성능을 평가하기 위해 FastSVC, DiffSVC, Assem-VC, StarGANv2-VC 의 네 가지 모델을 선정하였다. 이후 각 모델 간 성능 평가를 위해 MCD Score를 계산하여 정량적인 수치로 나타내었다. 그 결과, 여러 VC 기술의 장점을 접목한 Assem-VC의 성능이 가장 높게 나타났다. 그러나 연구에 사용된 데이터셋의 크기와 종류가 제한적이기에, 일반화된 성능 평가를 위해서는 다양한 데이터셋을 확보해야 한다는 한계점이 존재한다. 또한 본 연구는 합성된 음성의 품질에만 초점을 맞추었으나, SVC 모델의 변환 속도와 효율성 등 요인은 고려하지 않았다. 따라서 후속 연구를 통해 더 다양한 SVC 모델과 데이터셋에 대한 평가가 이루어질 필요가 있다. SVC 기술은 서로 다른 VC 기술과 융합하며 성능 향상을 이뤄내고 있다. 따라서 이러한 성능 평가는 더 높은 품질의 SVC 모델을 개발하고, 다양한 분야에서의 SVC 활용 가능성을 높이는 데에 도움이 될 것이다.

ACKNOWLEDGMENT

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음” (2018-0-00209)

참고문헌

[1] Chen, X., Chu, W., Guo, J., & Xu, N. (2019, March). Singing voice conversion with non-parallel data. In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 292-296). IEEE.

[2] Kominek, J., Schultz, T., & Black, A. W. (2008, May). Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In SLTU (pp. 63-68).

[3] Müller, M. (2007). Dynamic time warping. Information retrieval for music and motion, 69-84.

[4] Kim, K. W., & Lee, J. (2021). Controllable and Interpretable Singing Voice Decomposition via Assem-VC. arXiv preprint arXiv:2110.12676.