2025년도 한국통신학회 추계종합학술발표회

cVAE 와 Flow Matching 기법을 활용한 가창 음성 합성 기술 개발 연구

윤민혁, 최용훈*
광운대학교, *광운대학교

gurals9368@gmail.com, * yhchoi@kw.ac.kr

Development of Singing Voice Synthesis Technology Using CVAE and Flow Matching

Minhyeok Yun, Yong-Hoon Choi*
Kwangwoon Univ., * Kwangwoon Univ.

요 약

본 논문은 기존의 본 논문은 가창 음성의 미세하고 다이나믹한 특징을 효과적으로 반영하여 고품질의 출력물을 생성할 수 있는 가창 음성 합성 구조를 제안한다. 기존 cVAE(Conditional Variational AutoEncoder) 기반 가창 음성 합성 모델들은 악보 정보(스코어)를 입력 조건으로 사용하여 실제 가창 음성으로부터 추출된 Posterior 잠재 변수와 유사한 Prior 잠재 변수를 생성하고자 하였으나, 가창 음성의 세밀한 표현을 담아내는 데는 여전히 한계가 있다. 이에 본 연구에서는 cVAE 를 통해 생성된 Prior 분포로부터 샘플링 된 잠재 변수를 Posterior 분포로부터 샘플링 한 잠재 변수에 더욱 가깝게 변환시키는 경로를 학습하는 CFM(Conditional Flow Matching) 기법을 추가적으로 도입해 미세 특징 표현력 및 출력 품질 향상이 가능한 새로운 가창 음성 합성 구조를 제안한다.

Ⅰ. 서 론

기존 가창 음성 합성 연구는 주로 악보와 음성 특징 간의 매핑을 학습하는 데 중점을 둔다. 특히, VISinger[1]와 그 후속 모델인 VISinger2[2]는 음성 합성 분야에서 자주 사용되는 cVAE 기반의 구조를 사용한 End to End 모델 구조를 제안하며 주목받았다. 이 모델들은 악보 정보를 조건으로 사용하여 cVAE 의 잠재 공간을 통해 가창 음성의 특징을 모델링하고자 한다. 그러나 이러한 cVAE 기반 접근 방식에는 근본적인 한계가 존재한다. cVAE 구조의 악보 정보에만 의존하여 생성된 Prior 잠재 변수가 Mel Spectrogram 으로부터 추출된 Posterior 잠재 변수의 세밀한 특징을 정확하게 포착하기 어렵다는 점이다. 이 Prior 와 Posterior 분포로부터 샘플링 된 잠재 변수 간의 매핑 불일치로 인해 합성 결과물에 세밀한 특징이 충분히 반영되지 못한다. 따라서, 잠재 변수의 표현력을 극대화하여 고품질의 가창 특징을 반영하는 새로운 방법론의 도입이 필요하다.

Ⅱ. 본론

본논문에서는 기존 cVAE 기반 가창 음성 합성 모델의 잠재 공간 매핑 불일치 문제를 해결하기 위해, Conditional Flow Matching (CFM)[3] 기법을 도입한 새로운 가창 음성 합성 모델 구조를 그림 1 에 제안한다. 기존 연구인 VISinger2 의 cVAE 구조를 통해 생성된 Prior 잠재 변수를 실제 가창 음성의 미세 특징을 담고 있는 Posterior 잠재 변수로 이동시키는 최적의 경로를 학습한 CFM(Conditional Flow Matching) 을 통해 출력 결과물에 가창 음성의 세밀한 특징이 반영될 수 있도록 설계했으며, 자세하게는 depth-wise separable convolutional (DDSConv)가 임의의 시점에서의 Prior 잠재 변수가 Posterior 잠재 변수로 이동하기 위해서 가져야 하는 변화량을 예측해 정답 변화량과의 회귀 학습 방식을 사용했다. 이러한 접근으로 추론 시 사용될 Prior 잠재 변수의 표현력을 획기적으로 향상시켰으며, Mel-Spectrogram 의 시각적 비교로 제안 구조의 최종 출력물에 정답 가창 음성의 바이브레이션이 잘 반영된 것을 그림 2 에서 확인 가능하다.

모델 성능은 MCD(유사성), MOS(음질 평가)로 정성적,

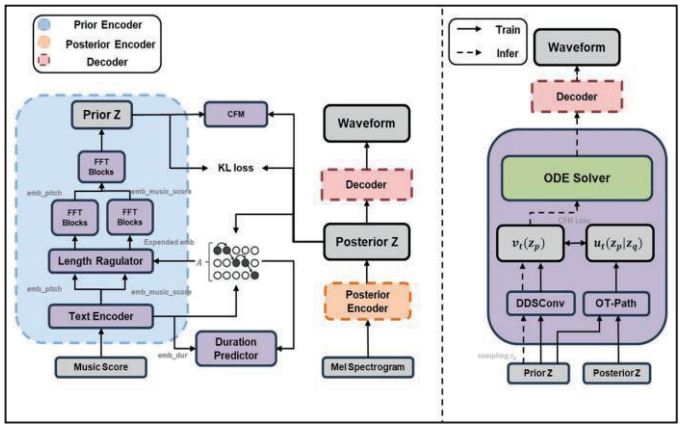

그림 1. cVAE 구조에 CFM 이 적용된 제안 구조 개요

정량적 평가가 모두 진행되었으며, 표 1 에 평가 결과를 나타내었다. 결과적으로 기존 모델에 비해 진행된 모든 평가에서 성능의 향상을 확인 가능했으며 특히 MOS 평가지표에서는 4.039 점을 달성하며 제안 구조가 기존 구조에 비해고품질의 가창 음성을 생성하는 것을 확인할 수 있다.

Model	MCD ↓	MOS ↑
Ground Truth	-	4.592 (±0.05)
VISinger2	6.328 (±0.04)	3.347 (±0.07)
VISinger2 NF	5.784 (±0.04)	3.569 (±0.07)
VISinger2 CFM (ours)	**4.815 (±0.03)**	**4.039 (±0.06)**

표 1. 각 구조의 MCD, MOS 평가 결과 표

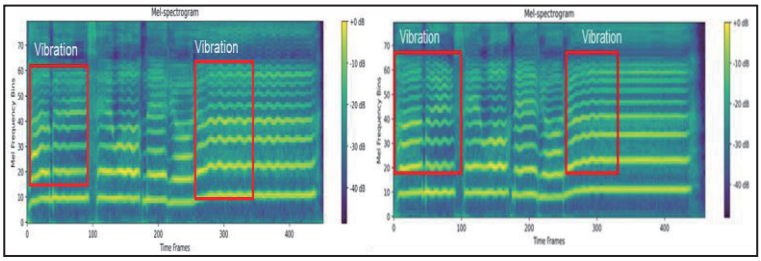

그림 2. 정답(좌측)과 제안구조(우측)의 Mel Spectrogram 비교

Ⅲ. 결론

본논문에서는 cVAE 구조 기반의 가창 음성 합성 모델에 한계를 극복하기 위해 잠재 변수간의 Conditional Flow Matching 방식을 적용 했다. 결과적으로 모든 수행된 모든 정성적, 정량적 평가에서 기존의 성능을 능가하는 것을 확인 가능했다. 추후 다국어 데이터에 대해서도 적용 가능한 모델로 발전시켜 나갈 계획이 있다.

ACKNOWLEDGMENT

본 연구는 국토교통부/국토교통과학기술진흥원의 2025 년도 지원으로 수행되었음(과제번호 :RS-2025-02532980).

참 고 문 헌

[1] ZHANG, Yongmao, et al. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022. p. 7237-7241.

[2] ZHANG, Yongmao, et al. Visinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer. *arXiv preprint arXiv:2211.02903*, 2022.

[3] LIPMAN, Yaron, et al. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.