벡터 양자화 기반 엔드-투-엔드 원샷 음성 변환

심영준, 윤진성, 서영주

포항공과대학교

youngjunsim@postech.ac.kr, truestar2001@postech.ac.kr, yjsuh@postech.ac.kr

End-to-End One-Shot Voice Conversion Based on Vector Quantization

Youngjun Sim, Jinsung Yoon and Young-Joo Suh

Pohang University of Science and Technology (POSTECH)

요 약

화자의 음성 특징을 모방하기 위한 음성 변환 연구는 최근 상당한 발전을 보이고 있다. 원샷 음성 변환(One-shot voice conversion)은 훈련 데이터셋에 포함되지 않은 화자의 음성을 모방하기 위한 연구로, 대상 음성의 한 가지 예만 참조로 사용하여 화자 특징에 충실한 고품질의 변환된 음성을 생성하는 것을 목표로 한다. 대부분의 음성 변환 모델은 대상 화자의 목소리로 변환된 멜-스펙트로그램(Mel-spectrogram)을 생성하고, 별도로 학습된 보코더(Vocoder)의 입력으로 사용하여 음성을 생성하는 2 단계 파이프라인 모델이다. 하지만 이러한 2 단계 파이프라인 음성 변환 모델은 보코더의 학습 데이터셋과의 분포 불일치에 의한 성능 저하를 야기한다. 본 논문에서는 음성의 언어적 특징과 화자별 특징을 분리하기 위한 벡터 양자화(Vector quantization) 기법과 보코더 구조를 결합한 모델을 학습하여 엔드-투-엔드(End-to-end) 원샷 음성 변환을 수행한다. 실험의 MOS (Mean Opinion Score), SMOS (Similarity Mean Opinion Score) 및 R-score를 활용한 음성 변환 성능 평가에서 가장 높은 점수를 달성하였으며, 높은 대상 화자 유사도를 가지는 고품질의 음성을 생성함을 보였다.

Ⅰ. 서 론

음성 변환이란 대상 화자의 음성 특징을 모방하기 위한 연구로, 소스 음성의 언어적 의미는 유지한 채, 대상 음성의 화자 목소리로 변환하는 기술이다. 이는 소스 음성에서 추출한 콘텐츠 정보(Content information)와, 대상 음성에서 추출한 화자 정보(Speaker information)를 결합하여 수행할 수 있다. 최근 관심이 많아지고 있는 원샷 음성 변환 분야는 학습 데이터에 포함되어 있지 않은 화자의 음성을 대상 음성으로 입력하였을 때에도 좋은 성능을 유지하는 것을 목표로 한다.

이를 위해서는 음성 데이터에서 언어적 특징 성분인 콘텐츠 정보와, 화자 특징 성분인 화자 정보를 성공적으로 분리해야 한다. 대표적인 방법으로는 정보 병목(Information bottleneck) 기법, 벡터 양자화 기법, 정규화(Normalization) 기법 등이 있다. 그 중 벡터 양자화 기법은 VQ-VAE [1] 구조를 사용한 기법으로 콘텐츠 인코더의 출력을 양자화 하여 화자의 특징 정보를 효과적으로 제거한다.

대부분의 음성 변환 모델은 2 단계의 파이프라인으로 이루어진다. 첫 번째 단계는 소스 음성의 콘텐츠 정보와 대상 음성의 화자 정보를 추출한 후 변환된 멜-스펙트로그램으로 합성하는 단계이다. 두번째 단계는 생성된 주파수 영역의 멜-스펙트로그램을 사전학습(Pretrained) 된 보코더의 입력으로 전달하여 시간 영역의 음성 파형을 생성하는 단계이다. 그러나 2 단계 파이프라인 음성 변환의 경우 두 모델이 각각 따로 학습되고, 음성 변환 모델에 의해 생성된 멜-스펙트로그램의 분포와 보코더가 학습한 데이터의 분포가 일치하지 않아 성능 저하를 야기할 수 있다.

따라서 본 논문에서는 벡터 양자화 기법을 통해 콘텐츠 정보와 화자 정보를 효과적으로 분리하고 보코더를 결합, 동시에 학습하여 높은 대상 화자 유사도를 가지는 고품질의 발화 음성을 생성할 수 있는 엔드-투-엔드 원샷 음성 변환 모델을 제안한다.

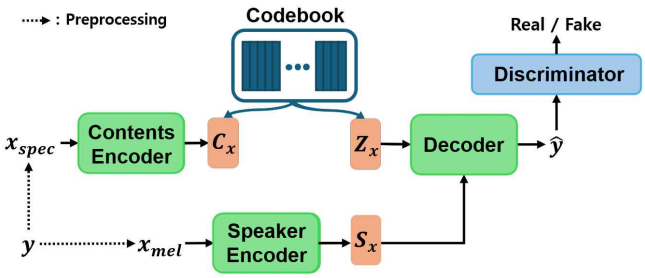

그림 1. 음성 변환 모델의 학습 절차 모식도

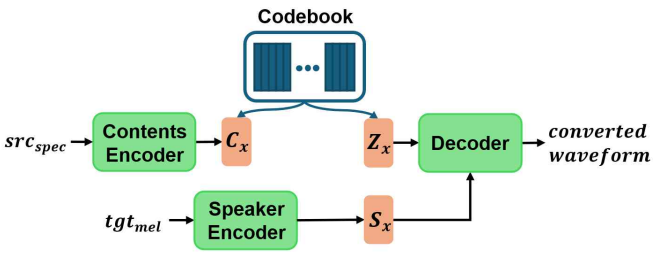

그림 2. 음성 변환 모델의 추론 절차 모식도

Ⅱ. 모델 설계

1. 모델 구조

그림 1은 모델 학습 시의 모식도다. 입력 음성 파형 y 는 전처리 과정을 거쳐 스펙트로그램 x_{spec} 와 멜-스펙트로그램 x_{mel} 로 변환하여 콘텐츠 인코더, 스피커 인코더의 입력으로 전달된다. 특징 차원 D 를 가지는 콘텐츠 임베딩 $C_x = \{c_1, c_2, \ldots, c_T\}$, $c_t \in R^{D\times 1}$ 는 $Z_x = \{z_1, z_2, \ldots, z_T\}$,

$z_t \in R^{D\times 1}$로 양자화 되고, Z_x는 스피커 임베딩 $S_x \in R^{D\times 1}$ 과 디코더의 입력으로 전달된다. 디코더의 출력 $\hat{y}$는 생성된 원본 음성 파형이다. 그림 2는 모델 추론(Inference) 시의 모식도를 나타낸 것으로, 소스 음성 src_{spec} 과 대상 음성 tgt_{mel} 이 입력으로 전달되면 소스 음성의 언어적 정보는 유지하며 대상 음성의 화자 정보로 변환한 음성을 생성한다.

2. 벡터 양자화

벡터 양자화 기법은 연속된 콘텐츠 임베딩을 개별적인 코드북 벡터로 대체함으로써 중요한 콘텐츠 특성을 유지하며 데이터를 효과적으로 압축한다. 이 과정을 통해 화자 정보를 효과적으로 제거할 수 있다. 콘텐츠 임베딩 $C_x \in R^{D\times T}$는 식 (1)을 통해 다음 $VQ(C_x) = \{z_0, z_1, ..., z_T\}$와 같이 양자화 된다. D 는 256이다.

$$VQ(c_t) = z_t,\ where\ z_t = \underset{q\in codebook}{\arg\min}(\parallel c_t - Z_x \parallel_2^2) \quad (1)$$

3. 인코더 / 디코더

콘텐츠 인코더와 스피커 인코더는 각각 콘텐츠 정보와 화자 정보를 추출한다. 콘텐츠 인코더는 WaveGlow [2]의 non-causal WaveNet residual block를 사용하였으며, 스피커 인코더는 가지는 3개의 LSTM 레이어로 구성하였다. 디코더는 HiFi-GAN [3] 보코더 구조를 사용한다.

4. 손실함수

총 손실함수(Loss function)은 식 (2)와 같다. L_{recon}은 재구성(Reconstruction) 손실함수로, $\hat{y}$을 멜-스펙트로그램으로 변환하여 x_{mel} 과의 L_1 거리를 계산한다. L_G, L_D, L_{fm} 은 각각 생성자(Generator), 판별자(Discriminator), 특징 매칭(Feature matching) 손실함수로 HiFi-GAN [3]의 손실함수이다. L_{latent}, $L_{codebook}$ 는 벡터 양자화 [1] 손실함수로 콘텐츠 임베딩과 코드북 사이의 거리를 계산한다. 본 논문의 실험에서 손실함수의 가중치 $\alpha, \beta, \gamma, \delta$로 각 45, 2, 0.5, 2를 사용했다.

$$L_{total} = \alpha L_{recon} + L_G + L_D + \beta L_{fm} + \gamma L_{latent} + \delta L_{codebook} \quad (2)$$

Ⅲ. 실험

1. 실험 설계

표 1 은 AutoVC [4], VQMIVC [5] 모델과 비교하여 성능을 평가한 것이다. 정성 평가로 15 명의 참가자를 통해 음성의 자연스러움과 유사도를 1 - 5 의 점수의 MOS와 SMOS로 측정하였다. VCTK와 LibriTTS 데이터셋에서 각각 남, 여 6명의 화자를 무작위로 선택하여 총 24명의 화자로 진행하였다. VCTK 데이터셋의 화자는 학습 데이터에 포함된 화자의 목소리로 음성을 변환하는 seen-to-seen, LibriTTS 데이터셋의 화자는 원샷 음성 변환 unseen-to-unseen 의 평가에 사용하였다.

정량 평가는 가짜 음성 탐지(Fake speech detection)을 위해 학습된 오픈소스 툴킷(Open source toolkit) (1)을 사용해 진행하였다. 변환된 음성에 대해 0 - 1의 점수 R-score를 부여한다. 점수가 높을수록 더 좋은 품질의, 대상 화자와 유사한 목소리를 생성한다는 의미이다.

2. 실험 결과

제안한 모델은 seen-to-seen 과 unseen-to-unseen 상황에서 MOS, SMOS, R-score 점수 모두 기존 모델 대비 가장 높은 점수를 달성하였다. MOS 점수의 경우 seen-to-seen 3.91, unseen-to-unseen 3.66 으로 기존의 2단계 파이프라인 모델들과 달리, 엔드-투-엔드 구조를 사용하여 원샷 음성 변환 상황에서도 품질이 좋은 음성을 생성하였다. 특히, 4.44, 4.13 의 높은 SMOS 점수는 콘텐츠 인코더와 벡터 양자화 기법이 콘텐츠 임베딩에서 화자 정보를 효과적으로 제거하여 간단한 스피커 인코더 구조에서도 충분히 화자 정보를 추출할 수 있음을 보였다.

Methods	Seen-to-Seen VC			Unseen-to-Unseen VC		
	MOS	SMOS	R-score	MOS	SMOS	R-score
AutoVC[4]	2.68	2.18	0.52	2.18	2.08	0.48
VQMIVC[5]	3.51	2.84	0.57	2.97	2.55	0.56
Ours	**3.91**	**4.44**	**0.77**	**3.66**	**4.13**	**0.67**

표 1. 모델 별 성능 비교

Ⅳ. 결론

본 논문에서는 훈련 데이터에 포함되지 않은 화자의 단일 샘플을 사용하여 해당 화자의 목소리를 모방할 수 있는 원샷 음성 변환 모델을 제안하고 검증한다. 벡터 양자화 기법과 내장(Built-in) 보코더를 결합하여 엔드-투-엔드 모델을 구축하였고, 간단한 스피커 인코더 구조를 사용하여 변환된 음성 파형을 생성했다. 실험의 seen-to-seen, unseen-to-unseen 상황 및 정성 평가, 정량 평가 모두 높은 점수를 달성하며 좋은 품질의 음성을 생성해냈으며, 특히 높은 유사도 점수를 통해 콘텐츠 임베딩에 포함된 화자 정보 제거에 탁월한 성능을 보였다.

ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2019-0-01906,인공지능대학원지원(포항공과대학교))과 2024년도 정부(교육부)의 재원으로 한국연구재단의 지원(No.2022R1A6A1A03052954, 기초연구사업)을 받아 수행된 과제입니다.

참 고 문 헌

[1] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. Advances in neural information processing systems, 30, 2017.

[2] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3617–3621. IEEE, 2019.

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in neural information processing systems, 33:17022–17033, 2020.

[4] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In International Conference on Machine Learning, pages 5210–5219. PMLR, 2019.

[5] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. arXiv preprint arXiv:2106.10132, 2021.

(1) https://github.com/resemble-ai/Resemblyzer