Diffusion 기법을 통한 Zero Shot 기반 Voice Cloning Speaker Embedding 잡음 제거 연구

정준섭, 양수빈, 김동환, 이성주*, 김학재
*상명대학교 소프트웨어학과, ㈜플레토로보틱스

202021034, 2024D4005, 2024D1009@sangmyung.kr, *peacfeel@smu.ac.kr, kjh@follettorobotics.com

Diffusion-Based Denoising of Speaker Embeddings in Zero-Shot Voice Cloning

Jeong Jun Seop, Yang Su Bhin, Kim Dong Hwan, Lee Sung Ju*, Kim Hak Jae
*SangMyung Univ., Folletto Robotics Incorporated.

요 약

최근 AI 기술의 발전과 음성 인식 기술, 음성 합성 기술에 대한 연구가 활발히 이루어지고 있으며, 그 중 별도의 화자 적응 과정 없이 소량의 음성만으로도 새로운 화자 목소리를 합성할 수 있는 Zero-Shot 기반의 Voice Cloning 기법이 주목받고 있다. 그러나, Zero-Shot 방식은 직접 Speaker Embedding 을 추출하기 때문에 배경 잡음이나 비의도적 특성이 Speaker Embedding 에 반영될 수 있어 합성 음질의 품질 저하를 유발할 수 있다. 따라서, 본 논문에서는 Zero-Shot 기반의 Voice Cloning 환경에서 Diffusion 을 이용하여 Speaker Embedding 내 잡음을 제거하는 방법을 제안한다. 제안방법과 잡음이 포함된 Speaker Embedding 을 Voice Cloning 하여 원본 음성과 코사인 유사도를 비교한 결과, 남성과 여성은 각각 평균 3%, 4% 개선되었다. 따라서, Diffusion 기법을 사용하여 Speaker Embedding 내 잡음을 제거하였을 때, 합성 음성 생성 과정에서 화자의 음질과 화자 표현의 정확성을 개선하여 합성 음성의 품질을 향상시킬 수 있음을 확인한다.

Ⅰ. 서 론

최근 AI 기술이 발전하여 ASR(Automatic-Speech Recognition)과 같은 음성 인식 기술과 Voice Cloning 과 같은 음성 합성 기술이 발전하고 있다[1]. Voice Cloning 기법 중, 별도의 학습 과정 없이 화자의 적은 음성만으로 화자의 음성을 합성할 수 있는 Zero-Shot 기반의 Voice Cloning 연구가 활발히 이루어지고 있다[2]. 그러나 Zero-Shot 방식은 화자 적응을 생략하여 배경 잡음이나 비의도적 특성이 Speaker Embedding 에 반영될 수 있으며, 이러한 특성은 음질 저하로 이어진다.

본 논문은 Diffusion 을 이용하여 Zero-Shot 기반 Voice Cloning 의 Speaker Embedding 에서 잡음을 제거하고, 성능을 코사인 유사도(Cosine Similarity)로 평가하는 방법을 제안한다.

Ⅱ. 제안방법

본 논문에서는 Speaker Embedding 내 비의도적 잡음을 제거하기 위하여 Diffusion 기반의 잡음 제거 기법을 사용한다. Diffusion 기법은 잡음을 점진적으로 제거하는 확률적 복원 과정을 통하여 특정 잡음 유형에 종속되지 않고 다양한 잡음에 대응할 수 있다. 따라서, 다양한 잡음에 대응 가능한 Diffusion 기반 잡음 제거 모델을 통하여 Speaker Embedding 을 정제하고, Voice Cloning 환경에서 합성 음성의 음질과 화자 특성을 향상시킨다.

그림 1 은 Voice Cloning 에 적용된 제안방법의 시스템 구조를 나타낸다. Speaker Encoder 를 통하여 Speaker Embedding 을 추출한다. 이후, Diffusion 기반의 Speaker Embedding Denoiser 를 이용하여 Speaker Embedding 에 포함된 잡음을 제거한다. 최종적으로, 잡음이 제거된 Denoised Speaker Embedding 을 TTS 내 Encoder 와 결합하여 화자 특성이 적용된 합성 음성인 Synthesized Speech 를 생성한다.

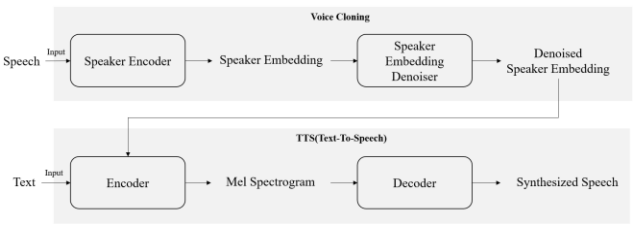

그림 1. 시스템 구조

Ⅲ. 실험 환경

본 논문에서는 Zero-Shot 기반의 Voice Cloning 시스템에서 화자의 발화 특성으로 사용되는 Speaker Embedding 의 잡음을 제거하기 위하여 MLP(Multi-

Layer Perceptron)을 이용하였다. MLP 기반 Diffusion 은 3 개의 은닉층으로 구성되며, 각 은닉층은 512 개의 노드로 구성된다. 각 은닉층은 시간 정보와 Speaker Embedding 을 결합하고, Diffusion 기법으로 Speaker Embedding 내 잡음을 점진적으로 제거한다. 본 논문에서는 Zero-Shot 기반의 Voice Cloning 모델로 MyShell 에서 제공하는 OpenVoice 를 활용하였다.

본 논문에서는 AI Hub 에서 제공하는 '자유대화 음성(일반남여)'를 활용하였으며, 100 명(남성 50 명, 여성 50 명)으로부터 수집된 평균 1 분 분량의 단일 화자 음성을 사용하였다. 또한, 실제 녹음 환경을 고려하기 위하여, 잡음이 포함되지 않은 원본 음성에 인위적으로 잡음을 첨가하였다. 잡음은 AI Hub 에서 제공하는 '도시 소리' 데이터와 '생활환경소음' 데이터 중 일부를 활용하였다. 표 1 은 사용된 잡음 유형 및 구성을 나타낸다.

표 1. 잡음 유형 및 구성

잡음 유형	잡음 종류
대중교통 소음	기차, 지하철, 항공기
생활 소음	샤워소리, 화장실소리, 문여닫는소리
가전 소음	식기세척기, 청소기, 에어컨 실외기
차량 소음	이륜차, 자동차, 덤프차, 불도저

본 논문에서는 제안방법의 잡음 제거 성능을 평가하기 위하여 코사인 유사도를 평가 지표로 사용한다. 코사인 유사도는 동일 화자 판별 시, 임계값은 0.6 에서 0.7 정도이다[3, 4]. 따라서, 본 논문에서는 동일 화자 판별 임계값을 0.6 으로 설정한다. 제안방법으로 잡음이 제거된 Speaker Embedding 과 잡음이 없는 원본 Speaker Embedding 간의 유사도, 잡음을 제거하지 않은 Speaker Embedding 과 잡음이 없는 원본 Speaker Embedding 을 비교하여, 제안방법의 잡음 제거의 성능을 확인한다.

IV. 실험 결과

본 논문에서는 제안방법으로 생성된 합성 음성을 원본 음성과 비교하였으며, 표 2 는 코사인 유사도를 통하여 잡음 유형별 Voice Cloning 성능을 비교한 결과를 나타낸다. 남성과 여성 화자의 경우, 잡음이 포함된 음성으로 생성된 합성 음성의 코사인 유사도는 각각 평균 66%, 68%이며, Diffusion 기법을 적용한 경우 각각 평균 69%, 72%로 유사도가 남성은 3%, 여성은 4% 개선되었다. 따라서, Zero-Shot 환경에서 Diffusion 을 이용하여 Speaker Embedding 내 잡음을 제거하였을 때, 원본 음성과의 유사도를 높일 수 있음을 확인한다.

표 2. 코사인 유사도를 통한 잡음 유형별 Voice Cloning 성능 비교(%)

잡음 유형	Voice Cloning		Diffusion Voice Cloning	
	남성	여성	남성	여성
대중교통 소음	66	68	69	73
생활 소음	65	66	68	70
가전 소음	67	68	69	72
차량 소음	68	69	70	71
평균	**66**	**68**	**69**	**72**

V. 결론 및 향후연구

본 논문에서는 Zero-Shot 기반 Voice Cloning 에서 Diffusion Model 을 이용하여 Speaker Embedding 내 잡음을 제거하는 방법을 제안하였다. 제안방법을 통하여 잡음 환경에서 코사인 유사도가 남성, 여성 평균 3%, 개선되었으며, 제안방법을 사용할 때, 합성 음성과 원본 음성과의 유사도를 높일 수 있음을 확인하였다. 향후 연구에서는 다중 화자 식별, 다중 화자 상황에서의 잡음 제거 등을 중심으로 연구를 확장할 예정이다.

ACKNOWLEDGMENT

이 연구는 2023 년도 산업통상자원부 및 산업기술기획평가원(KETI) 연구비 지원을 받아 수행된 연구임 (No. 20026232).

참 고 문 헌

[1] 이민영, 지현빈, 박은일, "음성 인식 기술 및 음성 합성 기술의 상용화," 정보과학회지, vol. 42, no. 9, pp. 14-20, Sep. 2024.

[2] K. Azizah, "Zero-Shot Voice Cloning Text-to-Speech for Dysphonia Disorder Speakers," IEEE Access, May. 2024.

[3] H. Zhang, Z. Cai, X. Qin, and M. Li, "Sig-vc: A speaker information guided zero-shot voice conversion system for both human beings and machines," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6567-6571, May. 2022.

[4] I. Christop, and M. Kubis, "ClonEval: An Open Voice Cloning Benchmark," 2025, arXiv preprint arXiv:2504.20581.