실환경 음성 데이터를 통한 디퓨전 음성 합성 모델 파인튜닝에 관한 연구

김세민, 김선욱, 강주연, 김남수

서울대학교 전기정보 공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{smkim21, sukim, jykang}@hi.snu.ac.kr, nkim@snu.ac.kr

A Study on fine-tuning diffusion Diffusion-based Text-to-Speech model utilizing in-the-wild speech data

Semin Kim, Seonuk Kim, Ju Yeon Kang, and Nam Soo Kim

Human Interface Laboratory,

Department of Electrical and Computer Engineering and INMC,

Seoul National University

요 약

본 논문은 사전 학습된 디퓨전 기반 음성 합성 모델을 새로이 수집한 실환경 음성 데이터로 파인튜닝하였을 때의 성능 변화를 분석한다. 본 연구에서는 LibriTTS 데이터셋으로 학습된 F5-TTS 모델을 초기 체크포인트로 사용하고, 음성 향상, 화자 분리, 음성 인식 등의 단계를 거쳐 정제한 유튜브 기반 음성 데이터를 활용하여 추가 학습을 수행하였다. 실험 결과, 실환경 음성 데이터로 파인튜닝한 모델이 발음 안정성과 강건성 측면에서 성능 향상을 보임을 확인하였다.

Ⅰ. 서 론

최근 디퓨전 모델을 기반으로 한 음성 합성 기술은 자연스러운 음질과 안정적인 발화 품질을 바탕으로 빠르게 발전하고 있다 [1-3]. 특히 Ditto-tts[2], F5-TTS[3]와 같은 zero-shot 디퓨전 음성 합성 모델의 경우 입력된 음성 프롬프트와 같은 화자의 음성을 원하는 텍스트에 대해 생성할 수 있는 능력을 높은 수준으로 끌어올렸다. 또한, 모델이 발전함에 따라 스튜디오에서 녹음된 “쉬운” 프롬프트에 대해서 충분한 성능을 보유한 것이 이미 검증되었다. 이에 따라 최근의 zero-shot 음성 합성의 경우 여러 배경 잡음, 화자 특성을 가진 다양한 음성 프롬프트에 대해 강건한 성능성 보이는 것이 주요한 지향점 중 하나가 되었다.

이에 따라 본 논문에서는 사전학습된 디퓨전 기반 음성 합성 모델을 실환경 음성 데이터로 파인튜닝하였을 때의 효과를 분석한다. 이를 위해 LibriTTS[4]로 학습된 F5-TTS 모델을 초기 모델로 사용하고, 정제한 음성 데이터를 활용하여 추가 학습을 수행하였다. 실험을 통해 실환경 데이터 기반 파인튜닝이 음성 합성 품질에 미치는 영향을 정량적으로 분석하였다.

Ⅱ. 본 론

본 연구에서는 실환경 음성 데이터로부터 음성 합성 모델 학습에 적합한 고품질 음성-텍스트 페어를 구축하기 위해 다단계 데이터 정제 파이프라인을 설계하였다. 전체 파이프라인은 음성 정규화, 음성 향상, 화자 분리, 음성 구간 탐지, 음성 인식의 순서로 구성된다.

먼저, 수집된 원시 음성 데이터에 대해 샘플레이트 변환 및 amplitude 스케일 정규화를 수행하여 입력 음성의 통계적 분포를 통일하였다. 이후 음성 향상 단계에서는 배경 잡음 및 환경 잡음을 제거하기 위해 사전 학습된 딥러닝 기반 음성 향상 모델을 사용하였다. 다음으로 화자 분리 단계에서는 하나의 음성 파일에 여러 화자가 포함된 경우를 처리하기 위해 화자 분리 모델을 활용하여 화자별 음성 구간을 분리하였다. 이를 통해 단일 화자 음성 단위로 데이터를 재구성하였다. 이후 음성 탐지 단계에서는 음성이 존재하지 않는 무음 구간을 제거하기 위해 음성 활동 탐지 모델을 적용하여 발화 구간만을 추출하였다. 마지막으로 자동 음성 인식 모델을 사용하여 각 음성 구간에 대응되는 텍스트를 생성하였다. 이 과정을 통해 최종적으로 30 초 이하 길이의 음성-텍스트 페어를 구성하였으며, 음성 합성 학습에 적합하지 않은 품질의 샘플은 추가적으로 필터링하였다.

본 연구에서는 파이프라인을 통해 유튜브 수집 데이터인 YODAS[5]의 일부를 정제하여 약 3000 시간 음성 데이터를 확보하였으며 이를 파인튜닝에 활용하였다.

본 연구에서 사용한 베이스라인 모델은 F5-TTS 로, 디퓨전 과정을 기반으로 멜 스펙트로그램을 생성하는 음성 합성 모델이다. F5-TTS 는 입력 텍스트와 화자 음성 프롬프트를 조건으로 사용하여, 목표 화자의 음색과 발화 특성을 유지하면서 새로운 텍스트에 대한 음성을 생성할 수 있는 zero-shot 음성 합성 모델이다.

F5-TTS 는 확률적 디퓨전 모델의 변형으로, 직접적으로 노이즈를 예측하는 대신 velocity 를 예측하는 방식을 사용한다. 학습 과정에서는 정답 멜 스펙트로그램에 단계적으로 노이즈를 추가한 후, 모델이 각 단계에서의 velocity 를 예측하도록 학습된다. 추론 시에는 초기 노이즈 상태에서 시작하여 반복적인 denoising 과정을 통해 최종 멜 스펙트로그램을 생성한다.

사전 학습된 F5-TTS 모델은 약 500 시간 규모의 다화자 음성 합성 데이터셋인 LibriTTS 를 사용하여 학습되었다. 학습은 약 40 만 스텝 동안 진행되었으며, 이 체크포인트를 파인튜닝의 초기 모델로 사용하였다.

파인튜닝 단계에서는 새롭게 정제한 유튜브 기반 실환경 음성 데이터를 사용하였으며, 사전 학습과 동일한 학습 설정을 유지한 채 추가로 10 만 스텝 학습을 수행하였다. 이를 통해 모델이 기존의 정제된 스튜디오 음성뿐 아니라, 실제 환경에서 발생하는 다양한 발화 특성과 잡음 조건에 적응하도록 유도하였다.

모델 성능 평가는 LibriTTS 테스트 데이터 중 300 문장을 음성 프롬프트로 사용하고, 서로 다른 300 문장을 생성하는 방식으로 진행하였다. 생성된 음성에 대해 발음 안정성과 화자 유사도를 평가하였다.

발음 안정성은 Whisper 기반 음성 인식 모델을 사용하여 WER(Word Error Rate)을 측정하였으며, 화자 유사도는 WavLM-large[7] 모델을 이용하여 cosine similarity 기반의 화자 임베딩 유사도(SIM)를 계산하였다.

Model	WER	SIM
Baseline	2.96	0.617
Proposed	2.84	0.638

표 1. 베이스 라인 모델과 파인튜닝한 모델의 WER, SIM

사전 학습 모델과 파인튜닝된 모델을 비교한 결과, 표 1 에서 볼 수 있듯이 WER 과 SIM 지표 모두에서 유의미한 성능 향상이 확인되었다. 특히 사전 학습 모델이 평가에 사용된 데이터셋과 동일한 도메인의 데이터로 학습되었음에도 불구하고, 실환경 음성 데이터를 추가로 학습하여 파인튜닝하는 것만으로 zero-shot 음성 합성 성능에 뚜렷한 개선 효과가 나타남을 확인하였다.

Ⅲ. 결론

본 논문에서는 사전 학습된 디퓨전 기반 음성 합성 모델을 실환경 음성 데이터로 파인튜닝하고 그 효과를 분석하였다. 실험 결과, 유튜브 기반 실환경 음성 데이터를 활용한 파인튜닝이 발음 안정성과 강건성 측면에서 유의미한 성능 향상을 가져옴을 확인하였다. 이는 실제 서비스 환경에서 음성 합성 모델의 성능을 개선하기 위해 실환경 데이터 기반 학습이 중요함을 시사한다.

ACKNOWLEDGMENT

이 논문은 2026 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Jeong, Myeonghun, et al. "Diff-tts: A denoising diffusion model for text-to-speech." arXiv preprint arXiv:2104.01409 (2021). [2] Borsos, Zalán, et al. "Soundstorm: Efficient parallel audio generation." arXiv preprint arXiv:2305.09636 (2023).

[2] Lee, Keon, et al. "DiTTo-TTS: Diffusion transformers for scalable text-to-speech without domain-specific factors." arXiv preprint arXiv:2406.11427 (2024).

[3] Chen, Yushen, et al. "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching." Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025.

[4] Zen, Heiga, et al. "Libritts: A corpus derived from librispeech for text-to-speech." arXiv preprint arXiv:1904.02882 (2019).

[5] Li, Xinjian, et al. "Yodas: Youtube-oriented dataset for audio and speech." 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023.

[7] Chen, Sanyuan, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1505-1518.