2024 년 봄 POSTECH 컴퓨터공학과 과제연구 연구제안서

[One-shot 음성 변환 기술 연구]

학　　번: 20200437
이　　름: 김채현
연구 지도교수: 서영주

연구 목적 (Problem statement)

본 연구는 음성 변환(Voice Conversion, VC) 연구의 일환으로, 원본 화자 음성의 언어 정보는 변경하지 않으면서 목소리, 톤, 강세, 발음만을 수정하여 목표 화자의 음성으로 변환하는 것을 목적으로 합니다. 그 중 본 연구에서 사용하는 One-shot Voice Conversion 은 음성 처리 분야 내에서 최근 주목받는 영역으로 어떤 화자의 음성을 다른 화자의 음성으로 변환하는 작업을 단 한 번의 예제만을 이용하여 수행하는 기술입니다. One-shot VC 은 다음과 같은 이유로 중요성과 필요성을 가집니다.

먼저, One-shot VC 모델은 제한된 학습 데이터를 가지고도 모델 생성이 가능합니다. 일반적으로 GMM 기반 혹은 regression 기반 모델과 같이 일대일 혹은 일대다 VC 에 사용되는 모델들은 지도 학습 모델로서, 대부분 대량의 병렬 데이터를 필요로 합니다. 하지만 음성 데이터의 경우, 특정 화자에 대한 데이터 수집이 어려운 경우가 있습니다. One-shot VC 는 위의 모델들과 달리 최소한의 학습 데이터를 활용하여 현실적이고 확장 가능한 모델을 제공합니다. 그렇기에 데이터 수집이 제한되거나 실용적이지 않은 시나리오에서도 적용할 수 있는 유연성과 적응성을 가집니다.

두 번째로, One-shot VC 모델은 content 와 speaker information 을 효과적으로 분리함으로써 음성 처리 기술의 발전을 보여줍니다. 이러한 분리는 VC 프로세스를 보다 정확하게 제어하고, 더욱 정교하게 음성 합성 시스템이 개발될 수 있도록 돕습니다.

마지막으로, One-shot VC 모델은 위와 같은 이유들로 새로운 화자에 대한 일반화가 가능합니다. 본래 One-shot VC 모델은 단 하나의 화자 데이터 예제를 통해 content 와 speaker information 을 효과적으로 분리할 수 있기 때문에 일반화가 가능하며, 덕분에 다양성이 높은 도메인에서 실용적인 적용이 가능하다는 장점을 가집니다.

연구 배경 (Motivation and background)

음성 변환 기술에서 Vector Quantization (VQ) method 를 적용하여 One-shot VC task 를 수행한 VQVC 모델은 콘텐츠와 화자 정보를 분리하는 데 성공하게 만들었습니다. 그러나 해당 모델을 통한 음성 변환에서는 생성된 음성의 naturalness 와 interpretability 측면에서 아쉬움이 있습니다. 특히, VQ 기반 방법을 통한 음성 변환 모델들은 음성의 톤과 발음이 자연스럽게 보이지 않는 문제점이 있습니다. 이러한 한계로 인해 VQ 방법을 통한 음성 변환 기술은 아직도 더 나은 음성 품질을 위한 개선이 필요한 상태입니다.

더불어, 기존의 지도 학습 방식에서 발생하는 문제 중 하나는 대량의 병렬 데이터를 수집하는 어려움입니다. 그러나 One-shot VC 모델은 제한된 학습 데이터에서 content 와 speaker information 을 분리하여 학습함으로써 이러한 한계를 극복할 수 있는 가능성을 제시하고 있습니다. 이는 음성 변환 기술에서 새로운 연구 방향을 제시하며, 데이터 수집의 어려움을 극복하여 보다 일반화되고 현실적인 모델을 개발하는 방향으로의 발전을 도모합니다.

연구 방법 (Research proposal)

1. WavLM 을 활용한 Self-Supervised Learning(SSL) Feature 생성 및 VQ 기법을 사용한 speaker identity 제거

   자기 지도학습으로 사전 훈련된 WavLM(Waveform Language Model)을 활용하여 음성 데이터의 임베딩 표현을 더욱 풍부하게 추출할 수 있습니다. 이후 WavLM 에서 얻은 SSL feature 에서 VQVC 모델의 VQ 기법을 활용하여 speaker identity 에 대한 정보를 제거함으로써 content information 만 남길 수 있습니다.

2. Pretrained model 을 활용한 contents information 제거
   Speaker verification task 에서 사용되는 pretrained model 을 활용하여 speaker embedding 얻어냅니다. 이 때 대상 화자의 contents information 을 보다 효과적으로 제거 가능하기 때문에 화자의 speaker information 을 보다 객관적으로 추출할 수 있습니다.

3. 합성 음성 생성
   위의 두 가지 방법에서 추출된 content information 과 speaker information 을 합성합니다. 이러한 방식으로 합성된 음성은 더 높은 음질과 유사성을 가지며, 보다 자연스러운 음성 합성을 가능하게 합니다.

기대 효과 (Expected output)

연구 결과로 기대되는 효과는 여러 산업 분야에서 혁신적이고 긍정적인 변화를 가져올 것으로 예상됩니다.

첫째로, 음성 변환 기술의 발전은 엔터테인먼트 분야에서 새로운 차원의 경험을 제공할 것입니다. 캐릭터의 목소리를 보다 자연스럽게 조작할 수 있는 능력은 애니메이션, 게임, 영화 등에서 사용자의 감동을 더욱 깊게 만들어 줄 것으로 기대됩니다. 사용자는 다양한 스피커로의 음성 변환을 통해 자신만의 맞춤형 콘텐츠를 창조할 수 있게 되어, 창의적인 산업 분야에서 새로운 아이디어와 경험을 활용할 수 있을 것입니다.

둘째로, 창의적 산업에서는 음성 변환 기술을 통해 다양하고 혁신적인 콘텐츠를 생산할 수 있습니다. 사용자가 자유롭게 스피커를 선택하여 음성을 변환할 수 있다면, 브랜딩, 광고, 예술 등에서 새로운 비즈니스 모델과 마케팅 전략을 구축하는 데 도움이 될 것입니다. 이러한 창의적 활용은 시장에서의 경쟁 우위를 가져올 수 있을 것입니다.

마지막으로, 음성 변환 기술의 진보는 의사 소통 시스템 분야에서도 혁신적인 변화를 이끌 것으로 기대됩니다. 실시간 통역 및 음성 인터페이스의 향상은 의료, 교육, 업무 환경 등에서의 의사 소통을 보다 효과적으로 도울 것입니다. 더불어 음성 변환 기술은 다국어 및 다문화 환경에서의 의사 소통을 원활하게 하며, 글로벌 시장에서의 응용 가능성으로 확장시킬 수 있습니다.

이러한 종합적인 효과들은 음성 변환 기술의 발전이 산업계와 학계에 긍정적인 파급 효과를 가져올 것이며, 기존의 음성 합성 및 의사 소통 기술에서의 한계를 극복함으로써 다양한 분야에서 혁신적인 변화를 이끌어낼 것으로 기대됩니다.

참고 문헌

Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. Advances in neural information processing systems, 30.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505-1518.

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018, April). Generalized end-to-end loss for speaker verification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4879-4883). IEEE.

연구 추진 일정

3/19 ~ 4/2 : 문헌 조사

4/2 ~ 4/16 : Baseline model 구현

4/16 ~ 5/7 : 새로운 기법 적용

5/7 ~ 5/21 : 성능 비교