2024 년 봄 POSTECH 컴퓨터공학과 과제연구 진행보고서

[One-shot Voice Conversion 연구]

학 번: 20200437
이 름: 김채현
연구 지도교수: 서영주
학 과: 컴퓨터공학과

---

* 아래 사항들에 대해 4-5 페이지 내외 (최대 5 페이지)로 명료하게 서술한 뒤 **pdf 로 저장**하여 제출합니다. pdf 가 아닌 형태로 제출하면 감점입니다.

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 (Design and methodology)

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.

연구 진행 상황 (Progress report)

본 연구는 4 월까지 One-shot Voice Conversion 기술 분석에 주력하였습니다. 연구 시작단계로서 문헌 조사를 통해 기존의 음성 변환 모델들을 분석하고, 데이터 셋을 정하였으며 이를 토대로 초기 Baseline 모델인 VQVC 를 개발하였습니다. 주요 연구 활동은 다음과 같습니다.

1. 데이터셋 선정 및 준비

   본 연구에서 사용한 데이터셋은 CSTR VCTK Corpus 로, 110 명의 영어 화자가 말하는 다양한 방언을 포함한 오디오 데이터로 구성되어 있습니다. 이러한 풍부한 데이터셋은 음성 변환 모델의 학습에 이상적인 환경을 제공하기에 모델이 실제 화자의 음성과 유사한 방식으로 음성을 합성하게 하는데 큰 도움이 될 것으로 예상됩니다.

2. WavLM 을 활용한 SSL Feature 생성

   본 연구에서 사용된 WavLM 모델은 원시 오디오 데이터(Waveform)를 입력으로 활용하여 self-supervised pre-training 을 수행하는 모델입니다. 입력 데이터는 PCM (pulse-code modulation) 형태의 시계열 데이터로, WavLM 은 이 데이터를 직접 처리하여 복잡한 Convolutional Layer 과 Transformer Encoder 를 통해 1024 차원의 feature vector 를

추출합니다. 이렇게 WavLM 을 이용한 SSL Feature 생성 방법은 Waveform 에서 직접 특성을 추출함으로써 다양한 특성을 포함한 Embedding 을 추출할 수 있습니다.

3. VQVC 모델 적용

VQVC 모델은 WavLM 모델로부터 추출된 특성을 기반으로 하여 음성의 content information 과 speaker identity 를 효과적으로 분리합니다. 모델은 크게 세 부분으로 구성되어 있습니다. Encoder, Vector Quantization (VQ), 그리고 Decoder 입니다.

◆ Encoder

Encoder 는 WavLM 으로부터 추출된 1024 차원의 Mel spectrogram 을 저차원의 latent vector 로 변환하는 역할을 합니다. 이는 1D Convolutional Layer 들과 Residual Blocks 을 통해 처리되고, 각 Layer 는 오디오의 시간적 및 공간적 특성을 추출하고 강조하여 최종적으로 32 차원의 feature vector 를 출력합니다.

◆ Vector Quantization (VQ)

Encoder 를 통해 얻은 연속적인 32 차원의 feature vector 들을 이산적인 Codebook vector 로 매핑합니다. 이 과정은 VQVC 모델의 Codebook 을 활용하여 가장 가까운 벡터를 찾고, 해당 벡터로 특성을 양자화하는 방식으로 수행됩니다. Codebook 의 크기는 256 으로 설정하였습니다. 양자화된 벡터는 speaker identity 로부터 독립적인 content 정보를 순수하게 표현하며, 이는 다양한 화자의 음성에 대해 일관된 음성 변환을 가능하게 합니다. 추가적으로, content 정보를 분리한 후, 남은 residual (원시 벡터에서 양자화된 벡터를 뺀 결과)은 speaker identity 정보를 포함하게 됩니다. 이 residual 정보는 speaker embedding 으로 활용되어, Decoder 에서 Mel Spectrogram 을 재합성할 때, 원하는 화자의 음성 특성을 입히는 데 사용됩니다.

◆ Decoder

양자화된 content 벡터와 별도로 추출된 speaker identity 벡터를 조합하여, 원래의 Mel Spectrogram 과 유사한 음성 신호를 재합성합니다. Decoder 는 이 두 벡터를 입력으로 받아, 복잡한 Convolutional 및 Residual Layers 를 통해 최종적인 음성 신호를 생성합니다. 이 과정은 변환된 음성이 원본 화자의 내용 특성을 유지하면서도 다른 화자의 음성 특성을 띄게 하는 것을 목표로 합니다.

연구 추진 일정 (Future plan)

4. Vocoder 를 이용한 Waveform 변환

현재 연구는 추출된 content 정보와 speaker 정보를 결합하여 Decoder 를 통해 melspectrogram 을 생성하는 것까지 진행되었습니다. 이후, 생성된 melspectrogram 은 VocGAN 이라는 고성능 vocoder 를 사용하여 최종적인 오디오 waveform 으로 변환할 것입니다. VocGAN 은 특히 자연스러움을 잘 보존할 수 있도록 설계되었으며, 따라서 변환된 waveform 은 원본의 음질과 매우 유사한 결과를 나타낼 것으로 예상됩니다.

5. Loss 함수 적용과 모델 평가

모델의 학습 과정에서는 두 가지 유형의 loss 함수를 적용할 예정입니다.

- Reconstruction Loss

  이는 입력 melspectrogram 과 출력 melspectrogram 간의 Mean Squared Error (MSE)를 계산하여, 모델이 원본 음성을 얼마나 잘 재현해내는지를 평가할 예정입니다. 낮은 MSE 값은 높은 재구성 정확도를 의미합니다.

- Latent Loss

  Encoder 의 출력과 Codebook 벡터 간의 거리를 최소화하는 loss 로, 이는 VQ-VAE 모델에서 중요한 양자화 효율과 모델의 안정성을 보장합니다. 이 loss 는 특히 Encoder 출력을 Codebook 벡터에 매핑하는 과정에서 발생하는 정보의 손실을 최소화하기 위해 사용됩니다.

현재까지의 연구를 통해 구현한 VQVC 모델은 음성의 content information 과 speaker identity 를 효과적으로 분리하고 재조합하여 하고 있습니다. 앞으로의 일정동안 VocGAN 을 사용하여 음성 변환의 자연스러움을 개선하고 모델의 정확성을 더욱 높일 예정입니다.