과제 수행 학생: 강민구
지도 교수: 윤성로 담당조교: 김희승

Voice Conversion

SpeechSplit 모델을 이용한 Voice Conversion

서울대학교 전기정보공학부 2017-19509 강민구

프로젝트 소개

프로젝트 요약

Bottleneck 이 있는 **Autoencoder** 기반
+
발화 음성을 **Rhythm**, **Pitch**, **Timbre**, **Content** 로 분리
+
VCTK + Zeroth-Korean 음성 데이터셋 활용

개발 배경 및 목적

자신의 목소리를 음성 변조하여 다양한 특색있는 목소리로 변형하고 싶었습니다.

⇒ 다양한 음성에 대한 Voice Conversion의 한계
⇒ 자신의 목소리에 대한 Voice Conversion의 한계

➤ 음성 데이터셋과 자신의 목소리 녹음본을 가지고 Voice Conversion 모델을 학습하고 Voice Conversion 진행

프로젝트 내용

Related Work

▫ WaveNet Model

- 오디오 데이터를 Dilated causal convolution을 이용하여 새로운 파형으로 생성하는 모델
- Melspectrogram을 wav 파일로 변환하는 Vocoder로 활용

▫ AutoVC Model

- Bottleneck 개념이 도입된 Autoencoder 기반의 Voice Conversion 모델
- 최초의 Zero-shot Voice Conversion 모델
- Self-Reconstruction loss 의 최소화

▫ SpeechSplit Model

- Bottleneck 개념이 도입된 Autoencoder 기반의 'AutoVC' 후속 모델
- Speech Information: Content, Timbre, Pitch, Rhythm

Background Information in Speech

Speech Information:

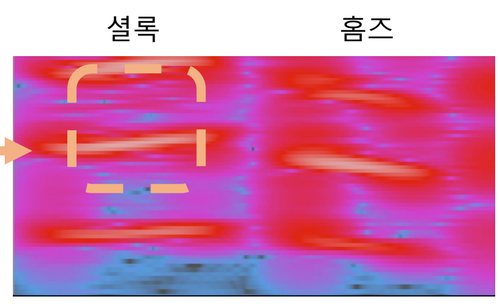
Content: 대부분의 언어에서 Content의 기본 단위를 'Phone' 이라고 하는데 각 'Phone' 이 특정한 포먼트 패턴을 형성한다.

Timbre: 스피커의 음성 특성으로 'Formant Frequency' (포먼트 주파수)에 의해 결정된다.

Pitch: 억양을 결정짓는 중요한 성분으로 발화 음성의 높낮이를 나타낸다.

Rhythm: 발화 음성의 빠르기를 나타낸다.

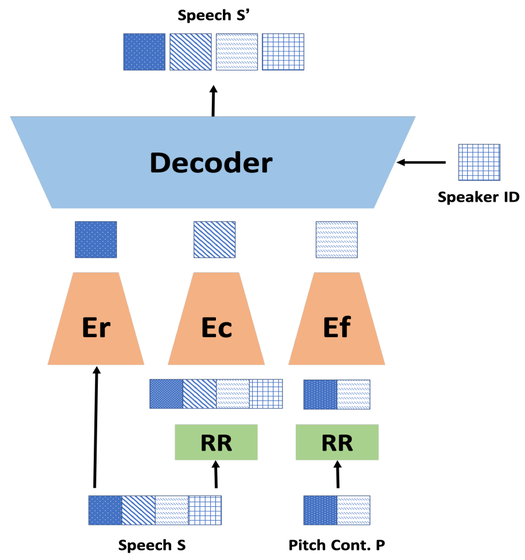
'Formant Frequency'



Framework

Framework of SpeechSplit:

Er: rhythm encoder
Ec: content encoder
Ef: pitch encoder
RR: random resampling

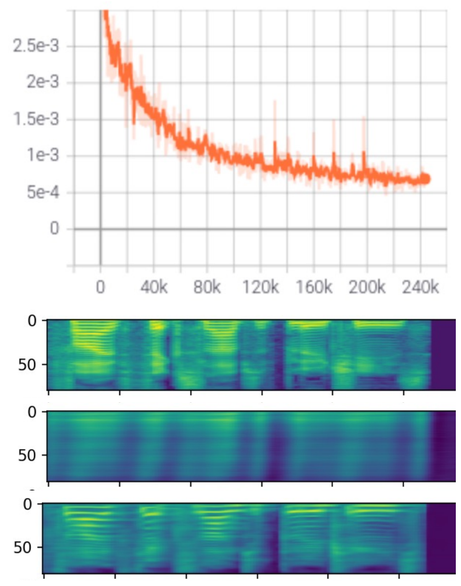


결과 및 계획

구현 결과

VCTK DataSet 의 남성 10명과 여성 10명의 음성 발화 파일, 총 20명의 스피커에 대한 음성 파일을 가지고 모델 학습 진행

➤ 남성 (10명) + 여성 (10명); 각 스피커당 약 10분 음성 파일



⇒ 각 Iteration에 대한 Self-Reconstruction Loss 변화 그래프

=> Speech **S**

=> Speech **S'** (Self Reconstruction at 1)

=> Speech **S'** (Self Reconstruction at 24800)

➤ Self-Reconstruction Loss를 이용한 모델 학습 진행

향후 계획

자신의 목소리에 대한 Voice Conversion의 한계
=> Speech-Split 모델 개선

WaveNet 모델을 활용한 Vocoder (spectrogram to wav file) 를 사용하여 Conversion하는데 약 10분 정도의 시간이 소요
=> hifi-gan 모델을 통한 빠른 Vocoder 활용

자신의 목소리를 녹음하여 학습한 경우에 대해서 Voice Conversion이 가능
=> 새로운 스피커에 대한 Voice Conversion 을 위한 Fine-Tune

기대효과

⇒ 언어 통역 분야의 Voice Conversion
⇒ 언어 교육 분야의 Voice Conversion
⇒ 오디오북 서비스의 Voice Conversion
⇒ 외국어 더빙에 대한 Voice Conversion

Reference

1. Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Mark Hasegawa-Johnson, AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. arXiv:1905.05879v2

2. Kaizhi Qian, Yang Zhang, Shiyu Chang, David Cox, Mark Hasegawa-Johnson, Unsupervised Speech Decomposition via Triple Information Bottleneck. arXiv:2004.11284v6