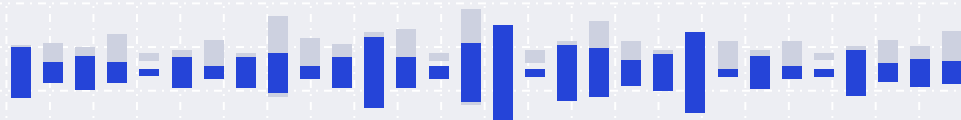




N3-VC : you can speak like a native

11기 양주원, 김여원

12기 김경환, 전종욱





Contents

01. Introduction

- why Speech-To-Speech?
- Our Goal

02. Background

- Speech-to-Speech

03. Dataset

- VTCK Corpus

04. Modeling

- Autovc
- SpeechSplit

05. Results

- 우리의 멋진 결과
- zzI 존 결과

06. Conclusion

- Limitation
- Future works



01. Introduction

Why

Speech-to-Speech?

- 음성 인식 기술에서 non-native 화자의 음성을 정확히 잡아낼 수 있을까?
- STT 기술은 많이 상용화되어있으나, 우리는 그에 앞서 발음을 교정해주는 STS가 필요하다고 판단함.

VIVA 브릿지경제 · 3주 전

신한은행, 음성인식 기술 기반 '모두를 위한 은행' 서비스 시행
신한은행, 음성인식 기술 기반 '모두를 위한 은행' 서비스 시행 신한은행이
시중은행 최초로 AI 기술기반 음성인식 기술(Speech To Text,STT)을 활용
한 상담 시각화 서비스 '모두를 위한 은행(Banking for Everyone)'을 시행...



기록관리 AI '다글로', 압도적 STT 성능으로 '위스퍼(Whisper)'와 격차
벌렸다!

상품소개

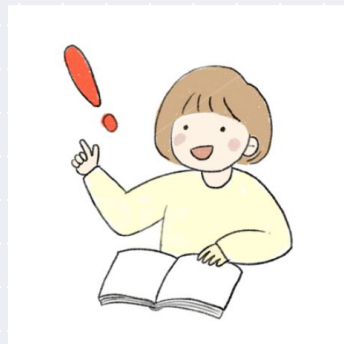
실시간 자막 구현까지!
강력해진 CLOVA Speech

"라이브 스트리밍 STT 기능 신규 추가"

01. Introduction

Our Goal :

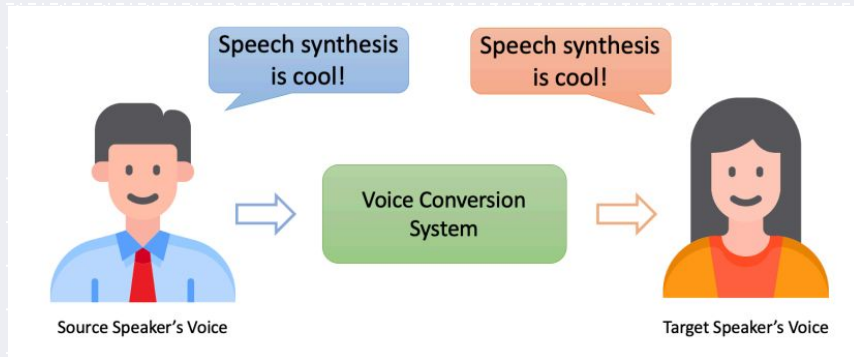
non-native 화자의 목소리는 유지하면서 발음만 native로 바꾸어주는 Speech-to-Speech 모델



02. Background

Speech-to-Speech

- **Voice Conversion** : source speaker의 음성 데이터를 target speaker의 목소리로 변환하여 출력
- 음성의 큰 두 가지 요소
 - Style (speaker의 voice identity)
 - Content (speaker의 voice identity를 제외한 나머지)
 - Content는 보존하고 style만 변환



02. Background

Speech-to-Speech

- Voice Conversion은 voice identity(style)에 집중하는 방식
- 대부분의 모델은 style 전체를 변환
- 우리의 목적인 발음, 억양 등 특정 style feature만 변환하고 나머지를 보존하는 speech-to-speech 모델은 없다 !

AutoVC

StarGAN-VC

SpeechSplit

02. Background

Speech-to-Speech

- Voice Conversion은 voice identity(style)에 집중하는 방식
- 대부분의 모델은 style 전체를 변환
- 우리의 목적인 발음, 억양 등 특정 style feature만 변환하고 나머지를 보존하는 speech-to-speech 모델은 없다 !

AutoVC

StarGAN-VC

SpeechSplit

03. Dataset

VTCK Corpus

- 110명의 다양한 악센트를 가진 사람이 각 400개의 문장을 읽은 44시간 분량의 음성 데이터셋
- 44,200개의 audio clips
- 48,000Hz sampling rate

| ID | AGE | GENDER | ACCENTS | REGION |
|-----|-----|--------|---------------|------------------|
| 225 | 23 | F | English | Southern England |
| 226 | 22 | M | English | Surrey |
| 227 | 38 | M | English | Cumbria |
| 228 | 22 | F | English | Southern England |
| 229 | 23 | F | English | Southern England |
| 230 | 22 | F | English | Stockton-on-tees |
| 231 | 23 | F | English | Southern England |
| 232 | 23 | M | English | Southern England |
| 233 | 23 | F | English | Staffordshire |
| 234 | 22 | F | Scottish | West Dumfries |
| 236 | 23 | F | English | Manchester |
| 237 | 22 | M | Scottish | Fife |
| 238 | 22 | F | NorthernIrish | Belfast |
| 239 | 22 | F | English | SW England |
| 240 | 21 | F | English | Southern England |
| 241 | 21 | M | Scottish | Perth |
| 243 | 22 | M | English | London |
| 244 | 22 | F | English | Manchester |
| 245 | 25 | M | Irish | Dublin |
| 246 | 22 | M | Scottish | Selkirk |
| 247 | 22 | M | Scottish | Argyll |
| 248 | 23 | F | Indian | |

03. Dataset

VTCK Corpus

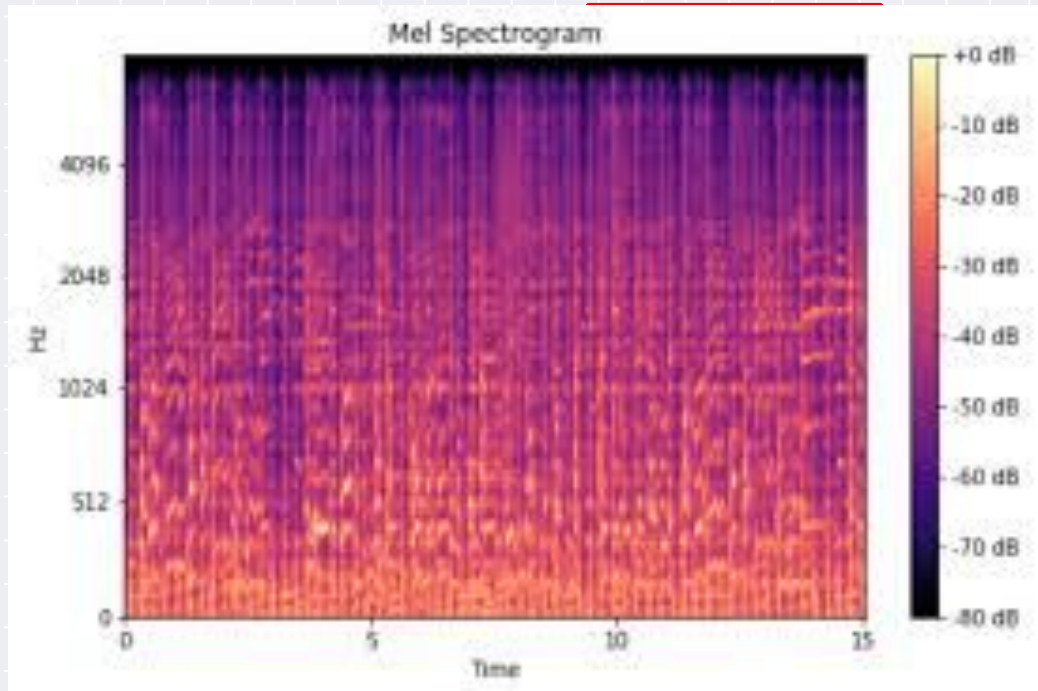
- Standard of native we set: England
- source speech : non-native regions
- target speech : native regions

| ID | AGE | GENDER | ACCENTS | REGION |
|-----|-----|--------|---------------|------------------|
| 225 | 23 | F | English | Southern England |
| 226 | 22 | M | English | Surrey |
| 227 | 38 | M | English | Cumbria |
| 228 | 22 | F | English | Southern England |
| 229 | 23 | F | English | Southern England |
| 230 | 22 | F | English | Stockton-on-tees |
| 231 | 23 | F | English | Southern England |
| 232 | 23 | M | English | Southern England |
| 233 | 23 | F | English | Staffordshire |
| 234 | 22 | F | Scottish | West Dumfries |
| 236 | 23 | F | English | Manchester |
| 237 | 22 | M | Scottish | Fife |
| 238 | 22 | F | NorthernIrish | Belfast |
| 239 | 22 | F | English | SW England |
| 240 | 21 | F | English | Southern England |
| 241 | 21 | M | Scottish | Perth |
| 243 | 22 | M | English | London |
| 244 | 22 | F | English | Manchester |
| 245 | 25 | M | Irish | Dublin |
| 246 | 22 | M | Scottish | Selkirk |
| 247 | 22 | M | Scottish | Argyll |
| 248 | 23 | F | Indian | |

04. Modeling

Speech data

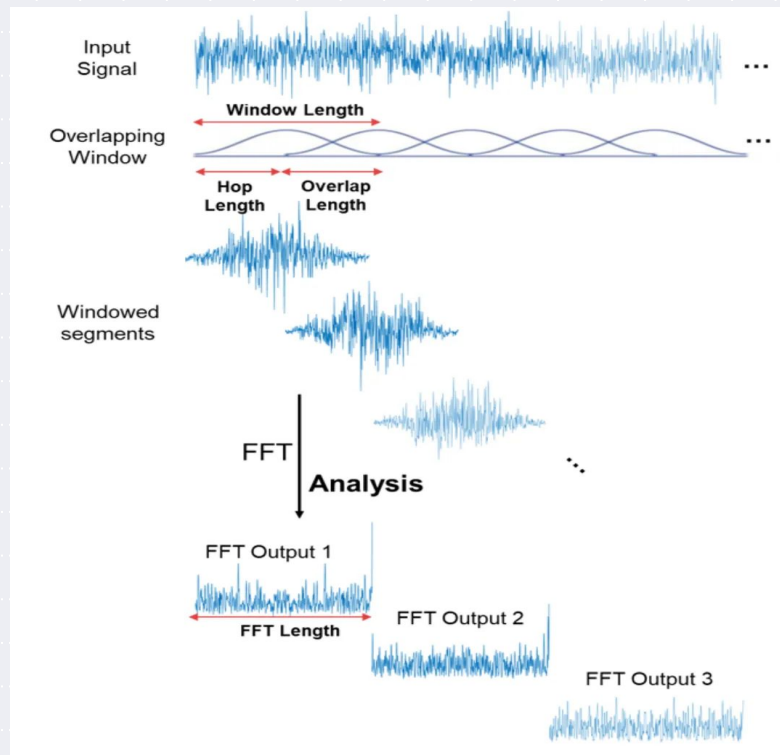
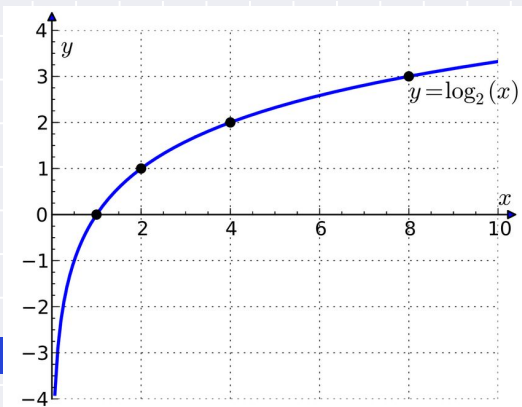
- 거의 이미지 처리와 동일
 - audio(wav) → mel spectrogram → CNN
 - 시계열 데이터이기 때문에 RNN도 사용.
- 하지만 거의 이미지 처리와 동일



04. Modeling

Speech data

- mel spectrogram의 x축은 시간. y축은 주파수
- 푸리에 변환 후 주파수에 log scaling을 하여 고주파수대역은 차이가 덜하게. 저주파수는 상대적으로 크게 만듦
- 이유는 고주파수는 사람들이 잘 구분 x



04. Modeling

Key point is disentanglement

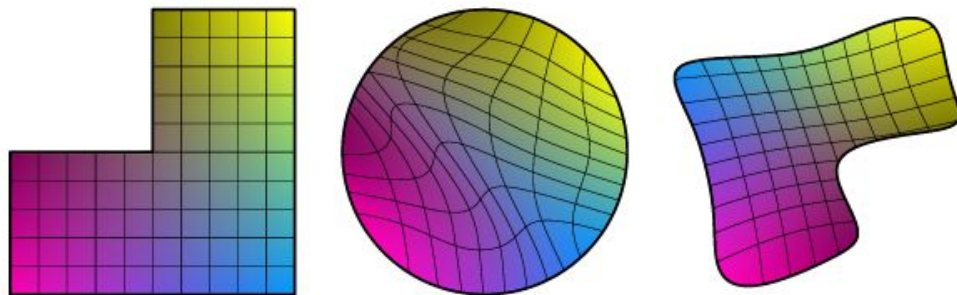
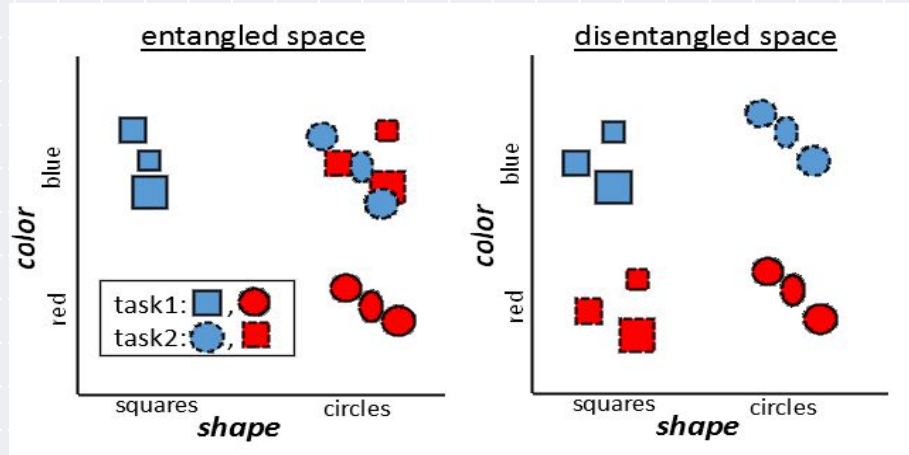
- Why disentanglement?

-원하는 건 X를 변화시켜도 Y에는 영향이
없도록

-disentanglement를

independent(orthogonal)와 비슷하다고
보면 된다

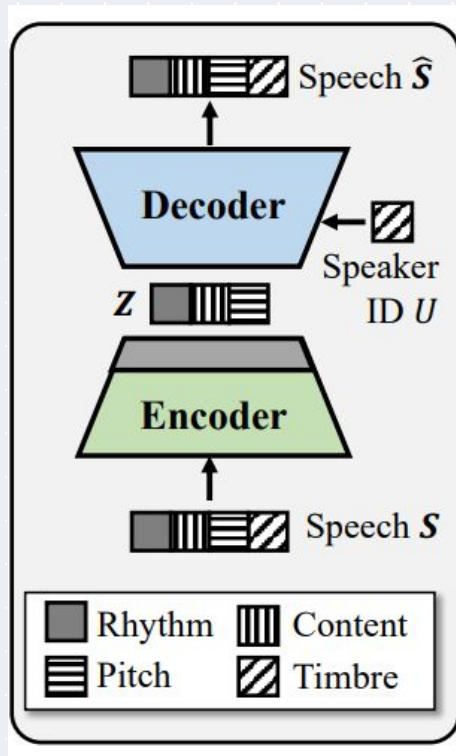
e.g. shape를 변화시켜도 color는 변하지
않음



04. Modeling

Disentanglement

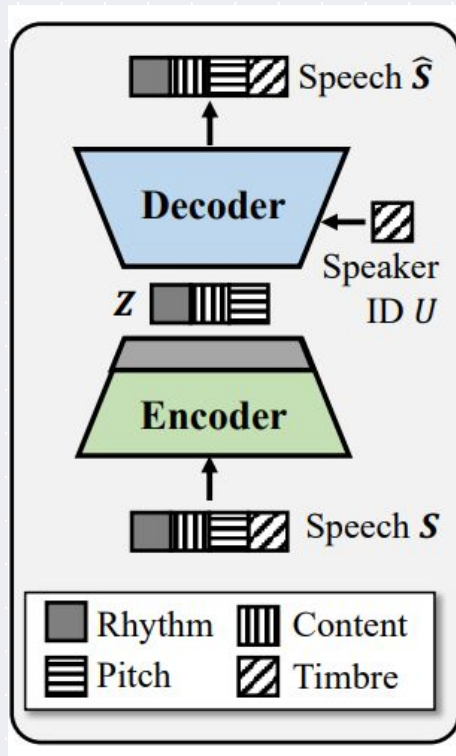
- 화자의 style은 여러 latent 변수들로 구성되어 있을 것
- 만약 accent에 해당하는 변수를 disentangle하게 추출할 수 있으면?
- 다른 정보는 보존하면서 accent만 바꿀 수 있지 않을까?
- 문제는 accent라는 것이 너무 추상적
- 선행 연구는?



04. Modeling

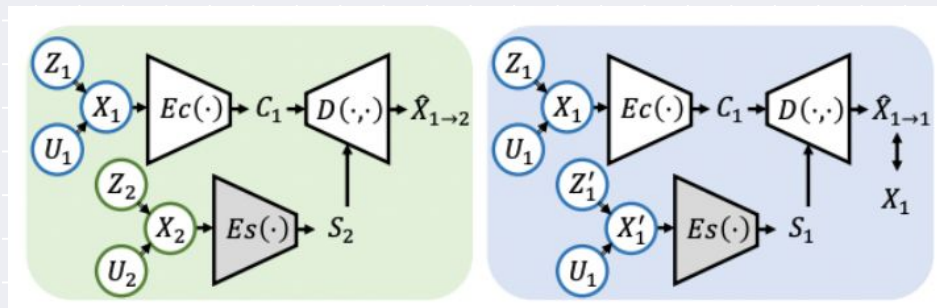
AutoVC

- autoencoder의 bottleneck 구조
기반으로 Speaker ID를 분리하여 voice를
disentangle하게 학습
content와 voice(style)를 분리함으로써 voice
conversion 진행.
- 최초로 Zero-shot Voice Conversion
수행
- Non-parallel (서로 다른 content로도
학습 가능)



04. Modeling

AutoVC

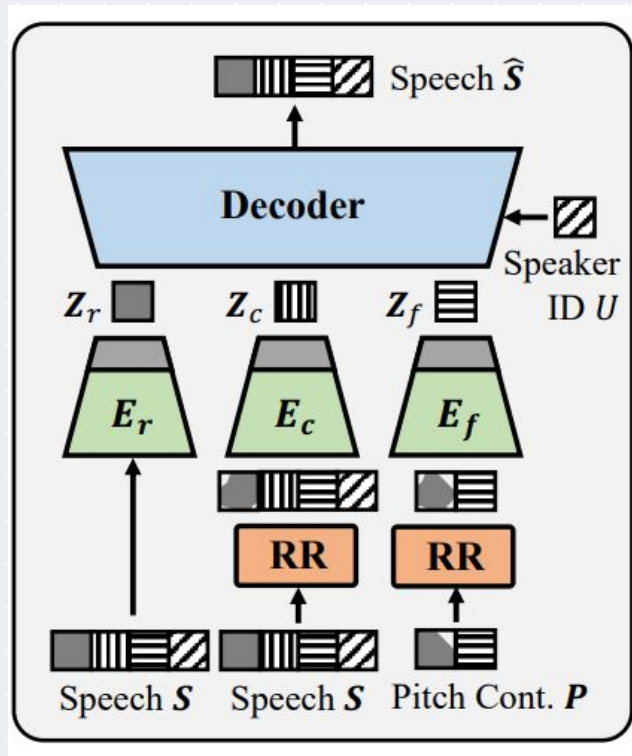


- source speaker의 content를 인코딩하는 Encoder, Ec
- target speaker의 identity를 인코딩하는 Encoder, Es (pre-trained)
- 각 content와 identity로 새로운 음성을 생성하는 Decoder

04. Modeling

SpeechSplit

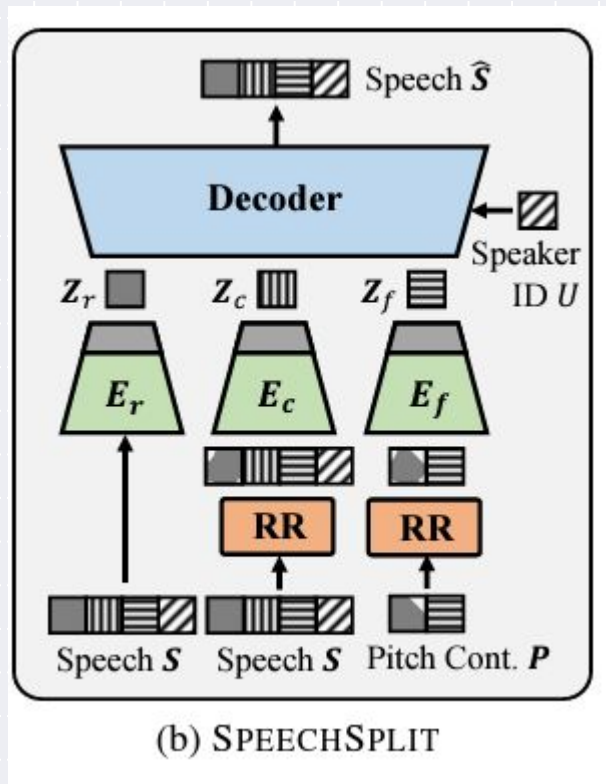
- Content, timbre, pitch, rhythm으로 분리 가능한 생성 모델
- AutoVC보다 더 좋은 성능을 보인다고 알려짐
- Content Encoder, Timbre Encoder, 그리고 style Encoder로 구성



04. Modeling

SpeechSplit

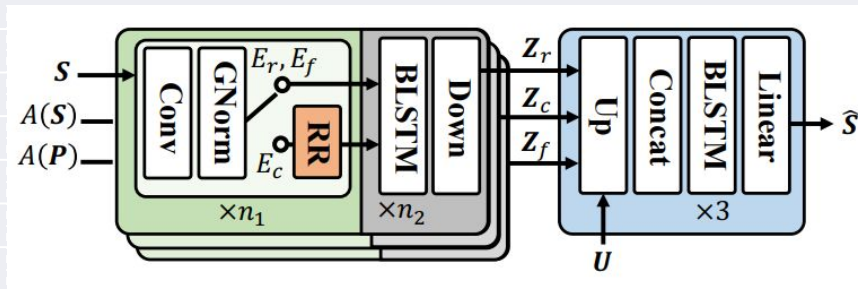
- 오디오 구성요소를 크게 네가지로 봄
- timbre(speaker 고유특성), pitch, rhythm, content
- 이 네가지 요소가 서로 disentangle한다고 가정
- 각 요소를 추출하는 encoder를 학습하여 따로 추출



04. Modeling

Why SpeechSplit?

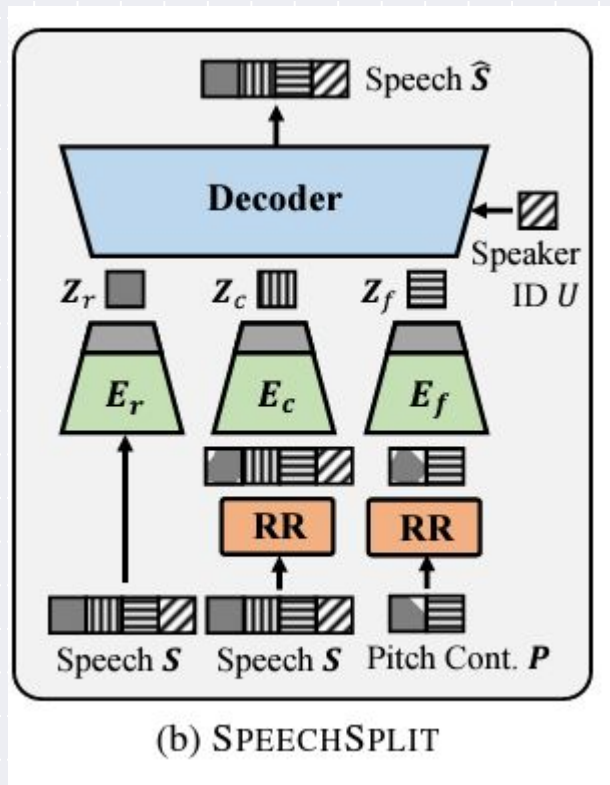
- style Encoder로 발음과 관련된 요소인 pitch와 rhythm의 특징을 뽑아낼 수 있을 것이라 판단 !



04. Modeling

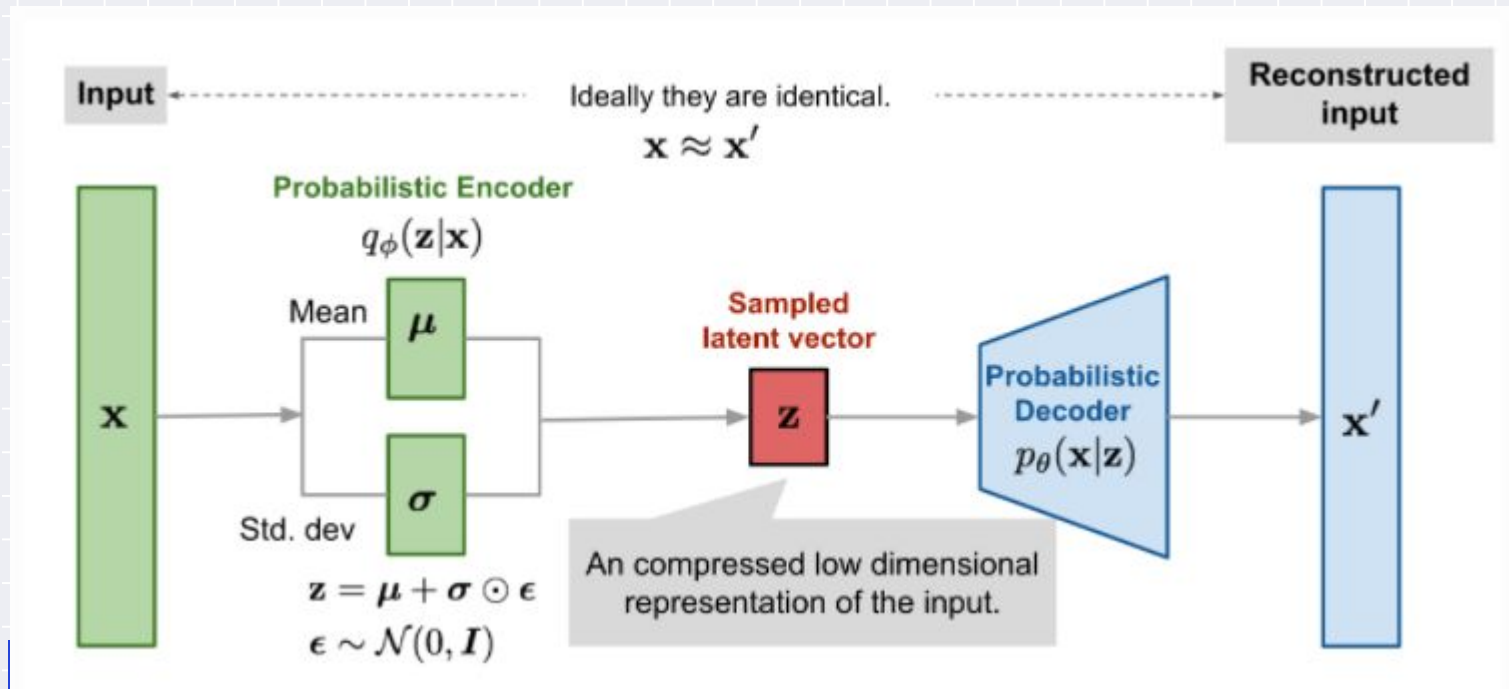
SpeechSplit

- timbre가 speaker identity 정보를 가지므로. 이 변수는 그대로 두고 나머지를 수정하면 어떨까?
- accent는 개개인의 특성보다 그룹의 특성(e.g. 한국인, 인도인, 일본인)이라고 보고 pitch.rhythm이 accent 정보를 가지고 있다고 가정



04. Modeling

VAE



04. Modeling

SpeechSplit

-Wavefile(mel -> pickle file로 변환

- wave->f0, spect vector로 분리
- vector들을 pickle파일로 저장
- wave->f0, spect vector로 분리

```
# compute spectrogram
D = pySTFT(wav).T
D_mel = np.dot(D, mel_basis)
D_db = 20 * np.log10(np.maximum(min_level, D_mel)) - 16
S = (D_db + 100) / 100

# extract f0
f0_rapt = sptk.rapt(wav.astype(np.float32)*32768, fs, 256, min=lo, max=hi, otype=2)
index_nonzero = (f0_rapt != -1e10)
mean_f0, std_f0 = np.mean(f0_rapt[index_nonzero]), np.std(f0_rapt[index_nonzero])
f0_norm = speaker_normalization(f0_rapt, index_nonzero, mean_f0, std_f0)

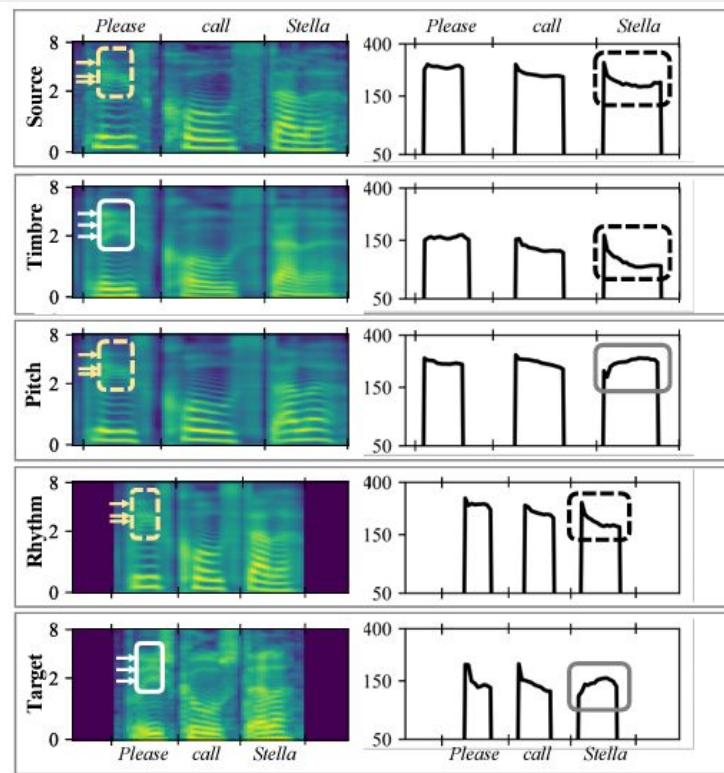
assert len(S) == len(f0_rapt)

np.save(os.path.join(targetDir, subdir, fileName[:-4]),
        S.astype(np.float32), allow_pickle=False)
np.save(os.path.join(targetDir_f0, subdir, fileName[:-4]),
        f0_norm.astype(np.float32), allow_pickle=False)
```

04. Modeling

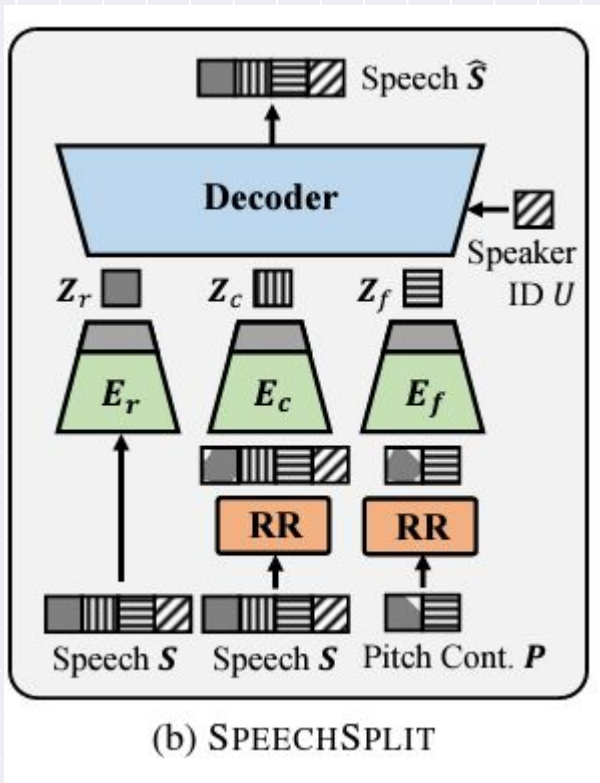
SpeechSplit

- we can get embedding vectors and 4 different vectors from melspectrogram and f0(pitch) Tensors
- changed the model.py file to only return rhythm and pitch tensors
- combine all the tensors



04. Modeling

SpeechSplit



05. Results

before



After



06. Conclusion

Limitation


- 10초짜리 음성이 출력되는데 7분
가량이 소요(SpeechSplit 자체가 무거운
모델..)
- VTCK 데이터셋 **speaker**의 **region**이
다소 제한적임(**native** 수도 많지 않음)
- 생성모델 시 **conditional**을 추가적으로
줘서 좀 더 **realistic**한 **feature**를
학습하다록 유도해야 된다고 생각
- 각 **feature** 간의 독립성을
가정하였으나, 독립적이지 않았던
것으로 보임

Future works

- 구조를 더 효율적으로 develop
- 더 다양한 **region**에 속한 **speaker**의
발화 데이터 확보



References

- Qian, K., Zhang, Y., Chang, S., Yang, D., & Hasegawa-Johnson, M. (2020). "Unsupervised Speech Decomposition via Triple Information Bottleneck." <https://arxiv.org/abs/2004.11284>
 - Chen, M. (n.d.). "Voice Conversion Guest Lecture." Retrieved from https://hajim.rochester.edu/ece/sites/zduan/teaching/ece477/lectures/GuestLecture_VoiceConversion_MelissaChen.pdf
 - Qian, K., Zhang, Y., Chang, S., Yang, D., & Hasegawa-Johnson, M. (2020). "SpeechSplit: Learning Speaker-independent Speech Representations via Disentangled Speech and Style Factors." Retrieved from <https://auspicious3000.github.io/SpeechSplit-Demo/>
 - Qian, K., Zhang, Y., Chang, S., Yang, D., & Hasegawa-Johnson, M. (n.d.). "SpeechSplit GitHub repository." Retrieved from <https://github.com/auspicious3000/SpeechSplit/tree/master>
 -
 -
- 



Thank You !

2024-2 Yonsei Data Science Lab Modeling Project

