Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation

Fadi Biadsy, Ron J. Weiss, Pedro J. Moreno, Dimitri Kanevsky, Ye Jia

Google

Interspeech 2019

Abstract

- Discrete한 Representation을 사용하지않는 Spectrogram -> Spectrogram Speech Conversion 모델. (End-to-End)
- Encoder, Spectrogram/ASR Decoder, Vocoder로 구성됨
- 모든 음성에 대해 Accent, Prosody, Noise에 관계 없이 일관된 표준 음성으로 정규화 시키는 것이 목적
- 청각장애인 등 비정상적인 음성을 정규화 된 음성으로 변환시킴으로써 음성인식이 가능하도록 함
- 이 구조는 Source Separation에서도 효과를 보임

Introduction

- Encoder-Decoder 모델은 NMT, ASR, TTS 등 Seq2Seq Task에서 좋은 성능을 보임
- 이 논문에서는 Attention 기반 Seq2Seq의 ASR과 Voice Conversion 모델을 결합해서 Speech-to-Speech 모델을 구축함, 이 모델은 Discrete한 표현이 필요하지 않음(End-to-End)
- 제안하는 모델이 여러 Noise, Accent, Prosody, 결함 등을 포함하더라도 음성을 정규화 하여 항상 동일한 음성을 생성할 수 있도록 함
- 텍스트 독립적인 Many-to-One Task에 해당됨

Introduction

- 논문은 Voice Conversion 논문이지만, 실제로 Conversion 하는 이유는 ASR을 위해서임
- 청각 장애인의 음성은 매우 비 정형적이므로 일반적인 음성인식 모델로는 인식이 불가능 함
- 비 정형적인 음성을 정규화 시켜서 음성인식이 가능하도록 함
- 또한 이 구조가 Source Separation에서도 효과가 있는지 확인할 것
- Phoneme Classification 을 통해 MTL학습을 진행함

Model Architecture

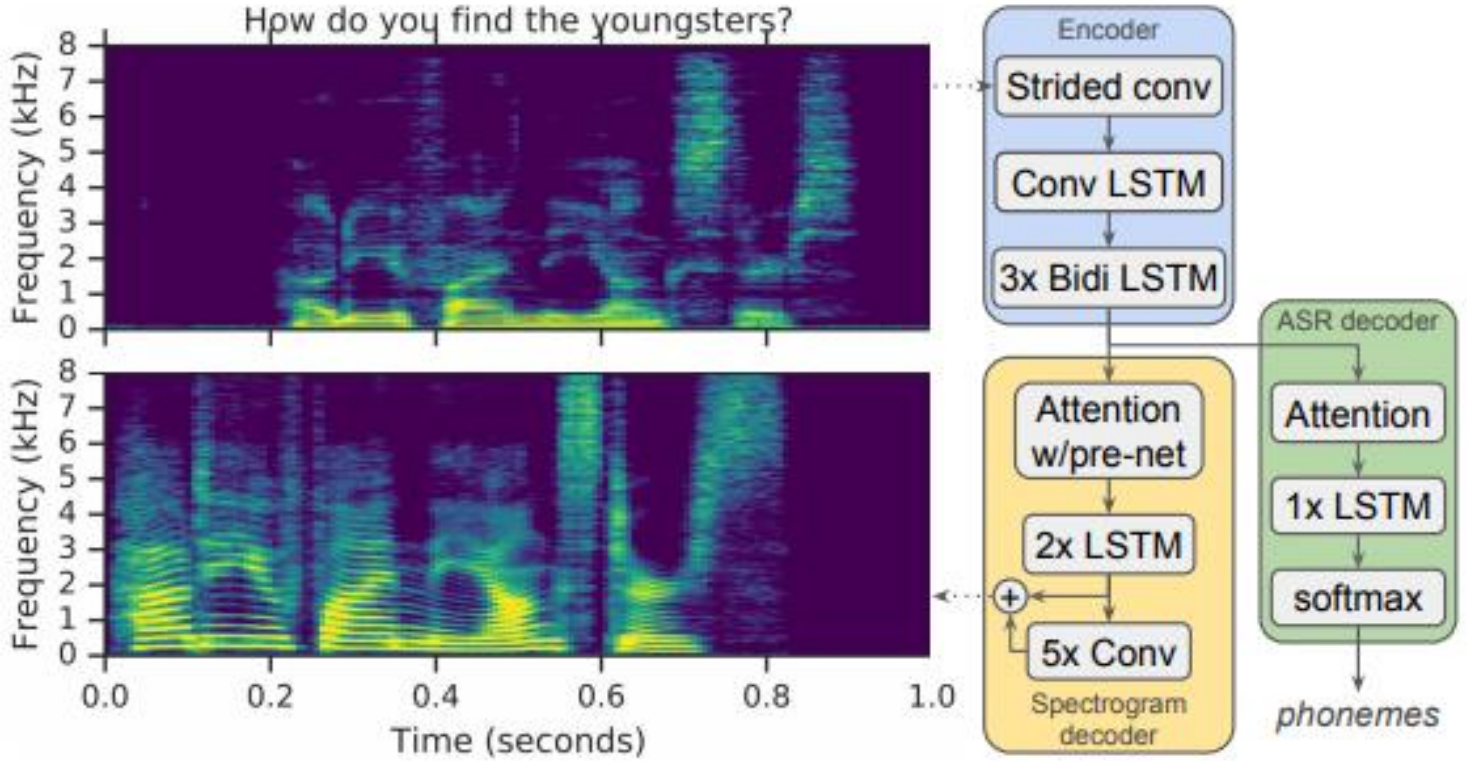
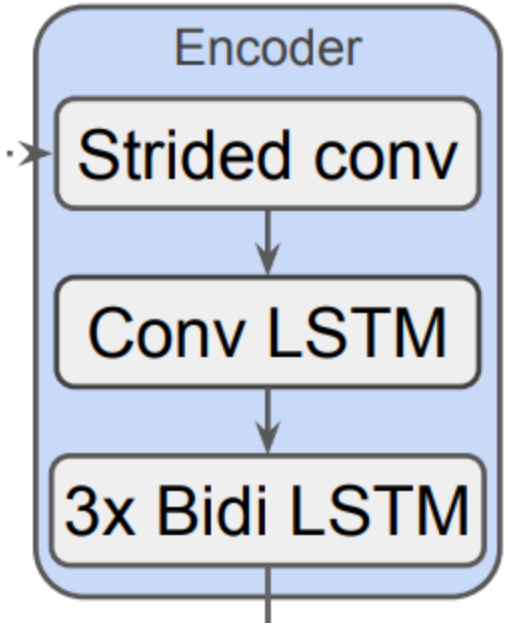


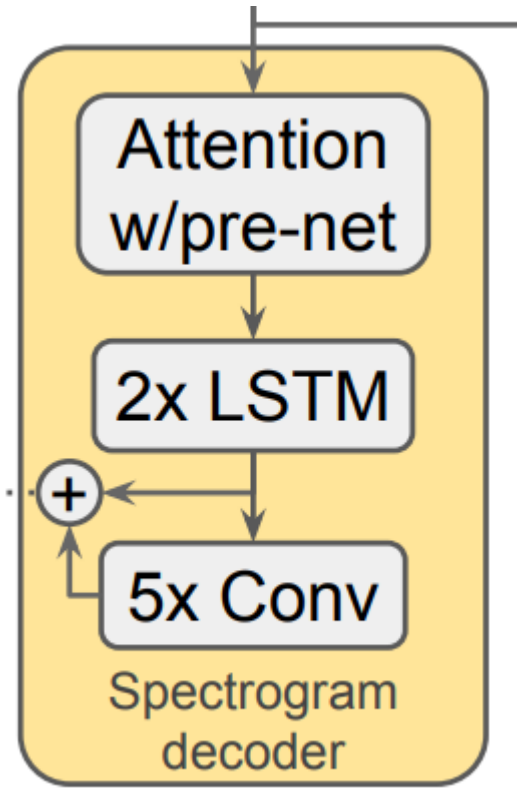
Figure 1: *Overview of the Parrotron network architecture. The output speech is from a different gender (having higher pitch and formants), and has a slightly slower speaking rate.*

Encoder



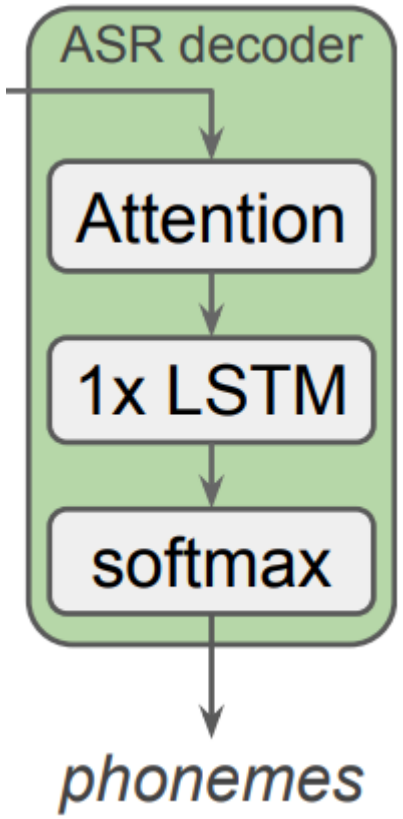
- 입력은 16kHz 음성
- 80 log-mel spectrogram
- Hann Window, 50ms, 12.5ms shift, 1024 FFT
- 32 Kernal(3x3), stride(2x2)
- Bi-ConvLSTM (1X3)
- 3 Bi-LSTM
- Encoder Hidden = 512 (256+256 concat)

Spectrogram Decoder



- 출력은 1025 Spectrogram (2048 FFT)
- Decoder는 일반적인 auto-regressive한 LSTM Decoder
- 2개의 FC, ReLU로 구성된 pre-net
- 2개의 Uni-LSTM (hidden 1024)
- LSTM out과 attention context vector를 concat 후 projection 하여 target spectrogram 추출
- Post-net으로는 5개 Conv (5x1) 사용 (residual)

ASR Decoder



- Spectrogram을 만들어 내는 것이 일반적인 소리를 만드는 것이 아니라 음성을 만드는 것이므로 ASR MTL시 도움이 됨(언어 정보에 Bias되도록)
- Phoneme/grapheme 단위로 ASR
- Location Sensitive Attention 사용

$$s_{t,i} = w^T \tanh(W d_{t-1} + V h_i + U f_{t,i} + b)$$

$$f_i = F * \alpha_{i-1}$$

Voice Normalization

- 임의의 음성을 표준 화자의 음성으로 정규화 하는 작업
- 학습을 위해 화자, 환경 등 다양한 음성이 필요하고 각각 음성은 표준 화자의 음성과 쌍으로 구성되어야 함
- 한 명의 화자가 Clean 환경에서 필요한 데이터만큼 녹음하는 것은 비용이 매우 많이 들고 비현실적임
- Google의 TTS를 사용하여 Target음성을 생성
- 일관된 억양, 무소음, 필요한 만큼 데이터를 만들 수 있는 장점 존재

Voice Normalization

- 실험에는 2,400만개의 발화로 구성된 30,000시간의 Google voice search traffic 데이터 사용
- 평가는 SOTA ASR 모델 사용하여 WER 측정

Table 1: *WER comparison of different architecture variations combined with different auxiliary ASR losses.*

ASR decoder target	#CLSTM	#LSTM	Attention	WER
None	1	3	Additive	27.1
Grapheme	1	3	Additive	19.9
Grapheme	1	3	Location	19.2
Phoneme	1	3	Location	18.5
Phoneme	0	3	Location	20.9
Phoneme	0	5	Location	18.3
Phoneme w/slow decay	0	5	Location	17.6

Voice Normalization

- 잡음이 심한 테스트셋에 대해서 실험
- WER은 실제 음성과 비슷한 결과를 보임
- MOS는 조금 떨어지나 Task의 목적은 자연스러운 발화가 아닌 Normalization

Table 2: *Performance of Parrotron models on real speech.*

Model	MOS	WER
Real speech	4.04 ± 0.19	34.2
Parrotron (female)	3.81 ± 0.16	39.8
Parrotron (male)	3.77 ± 0.16	37.5

Voice Normalization

- 랜덤하게 20개를 추출하여 8명의 영어 원어민에게 질문 후 평가
- Parrotron은 음성을 일관되게 Normalization할 수 있음

Table 3: *Subjective evaluation of Parrotron output quality.*

Survey question	Avg. score / agreement
How similar is the Parrotron voice to the TTS voice on the 5 point Likert scale?	4.6
Does the output speech	
use a standard American English accent?	94.4%
contain *any* background noise?	0.0%
contain *any* disfluencies?	0.0%
use consistent articulation, standard intonation and prosody?	83.3%

Normalization of Hearing-Impaired Speech

- 청각장애인 및 기타 언어장애인을 위한 음성 인식에 활용
- 비정형 음성을 유창한 음성으로 변환하는 작업
- 실험 대상은 러시아 출생의 10대의 청각장애를 지니고 영어를 배운 사람
- 영화 인용구 읽기에 해당하는 15.4시간의 데이터셋을 사용
- 90%(KADPT)를 Adaptation(Fine-tuning)에, 5%를 dev, 5%(KTEST)를 test에 사용
- 스크립트를 TTS로 합성하여 google ASR한 결과는 14.8% WER

Normalization of Hearing-Impaired Speech

- KTEST 에 대해 Google ASR로 평가
  - 실제 비 정형 음성 그대로 ASR
  - Parrotron으로 Normalize 후 ASR
  - Fine-tuned Parrotron으로 Normalize 후 ASR

Table 4: *Performance on speech from a deaf speaker.*

Model	MOS	WER
Real speech	2.08 ± 0.22	89.2
Parrotron (male)	2.58 ± 0.20	109.3
Parrotron (male) finetuned	$\mathbf{3.52 \pm 0.14}$	**32.7**

Speech Separation

- Parrotron이 Speech Separation에서도 효과가 있는지 실험
- 한 발화에 대해 1~7개의 발화를 무작위로 선택하여 배경 잡음으로 혼합
- SNR 12.15dB 정도
- Clean Speech보다는 높은 WER이지만 Denoising(Separation) 효과를 볼 수 있었음

Table 5: *Parrotron speech separation performance.*

Data	WER	del	ins	sub
Original (Clean)	8.8	1.6	1.5	5.8
Noisy	33.2	3.6	19.1	10.5
Denoised using Parrotron	**17.3**	6.7	2.2	8.4

Conclusion

- Spectrogram을 다른 Spectrogram으로 변환하는 End-to-End Voice Conversion 모델인 Parrotron
- 임의의 화자의 음성을 단일 화자 음성으로 Normalize하도록 학습
- 청각장애인의 비 정형 발화에 대해 Normalize하여 WER과 자연스러움을 개선할 수 있었음
- 또한, 여러 음성이 혼합된 상황에서 가장 큰 화자의 음성을 분리하여 ASR성능을 높이는데도 활용이 가능함