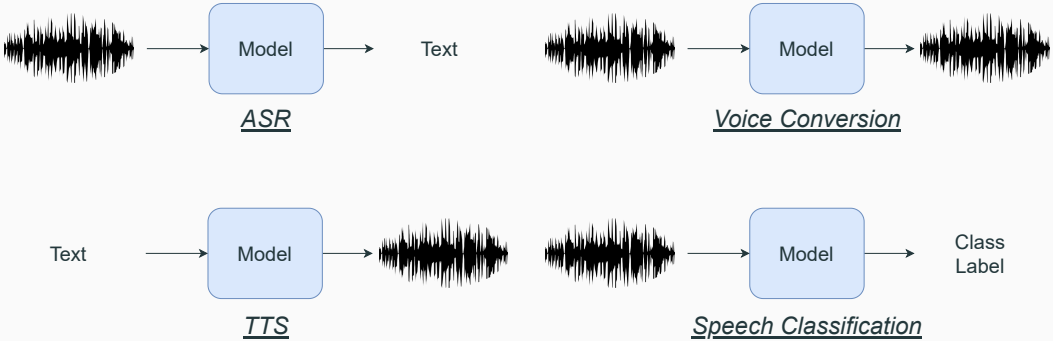
Recitation 9: Speech

ASR and Beyond

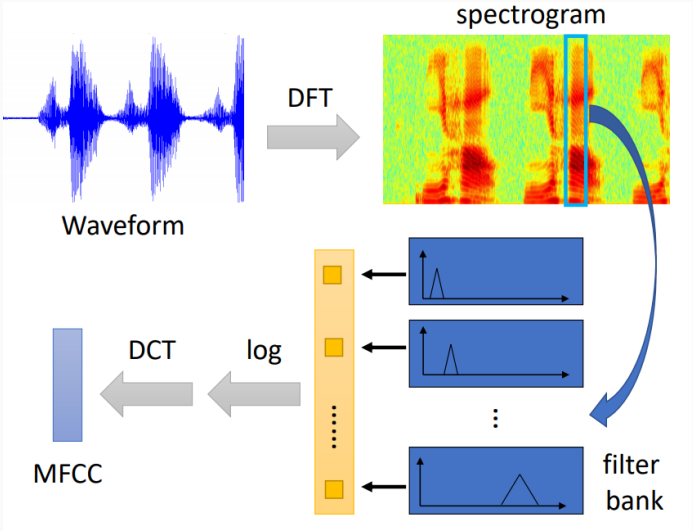
Abby Bertics & Wei Fang

MIT 6.806-6.864 Spring 2021

Overview


Model
Text
ASR
Model
Voice Conversion
Text
Model
TTS
Model
Class Label
Speech Classification

Acoustic Features

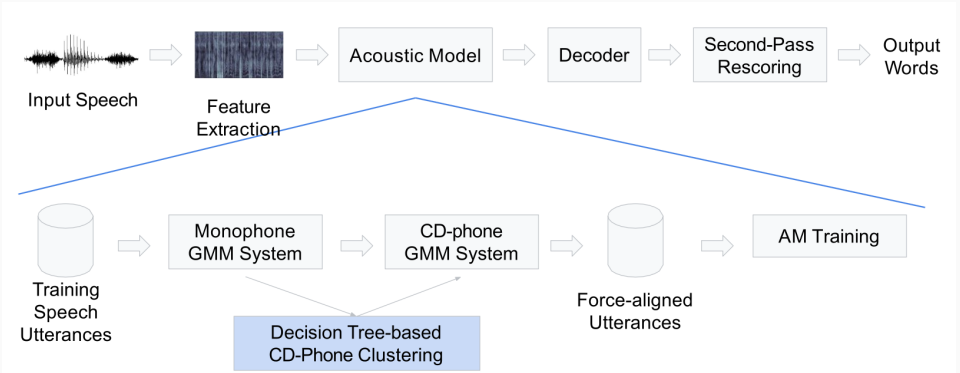


ASR

Some ASR Applications (think: voice assistants)

- Speech Recognition
- +Translation
- +Intent classification (eg. buy ticket)
- +Slot filling (eg. fill info for buying ticket)

Conventional ASR Pipeline



- Weighted FST for composing the components
- DNN Acoustic models: replace GMM (hybrid) or as inputs (Tandem)

ASR (End-to-end) - Listen, Attend and Spell (LAS)

Summary: typical seq2seq with attention

Seq2seq + Attention

- Encoder
  - Here pyramid BLSTM (can also use RNN/CNN/self-attention combinations)
  - downsampling: pyramid, pooling over time, time-delay DNN, dilated CNN, truncated self-attention
- Attention: many variations
- Decoder:
  - Training: teacher forcing, scheduled sampling
  - Decoding: beam search, LM rescoring

ASR (End-to-end) - Connectionist Temporal Classification (CTC)

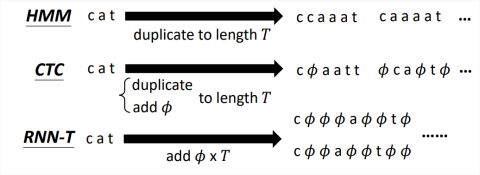
Summary: can be viewed as seq2seq with *linear* decoder

CTC

- Encoder: uni-directional RNN
- Classifier: linear, at each time step
- Training loss: not at each time step, but sum over all alignments (DP)
- Decoding: beam search

RNN-T

- Adds output dependencies with extra RNNLM



Comparisons

	LAS	CTC	RNN-T
Decoder	dependent	independent	dependent
Alignment	not explicit (soft alignment)	Yes	Yes
Training	just train it	sum over alignment	sum over alignment
On-line	No	Yes	Yes

(some) Recent Advances

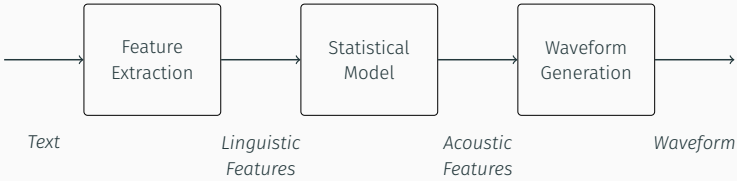
- Low-resource settings
  - unsupervised/semi-supervised/transfer learning
- Self-supervised learning / pre-training
  - speech representation learning (eg. CPC[1], APC[2], wave2vec[3], ...)
- SOTA: Huge data/models[4][5]

[1] https://arxiv.org/pdf/1807.03748.pdf
[2] https://arxiv.org/pdf/1904.03240.pdf
[3] https://arxiv.org/pdf/2006.11477.pdf
[4] https://arxiv.org/pdf/2104.03416.pdf
[5] https://arxiv.org/pdf/2104.02133.pdf

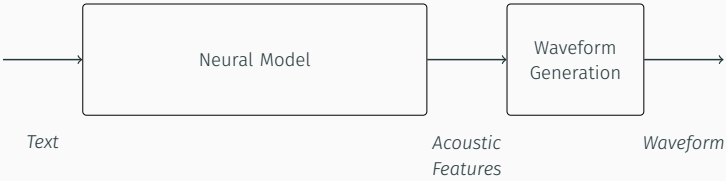
Text-to-Speech (TTS) Synthesis

Concatenative
From database of recordings

Parametric



End-to-end Neural



E2E Neural Speech Synthesis

- NN model (Enc-Dec): seq2seq + attn (Tacotron2[6])
  - Attention: modeling duration
- Vocoder (Waveform generation)
  - Rule-based: Griffin-Lim
  - Neural Network: WaveNet[7]
- Evaluation:
  - Metrics that correlate with voice quality and prosody. eg. mean cepstral distortion (MCD)
  - (much more important) Human evaluation: Mean opinion scores (MOS)

[6] https://arxiv.org/pdf/1712.05884.pdf
[7] https://arxiv.org/pdf/1609.03499.pdf

Beyond Tacotron

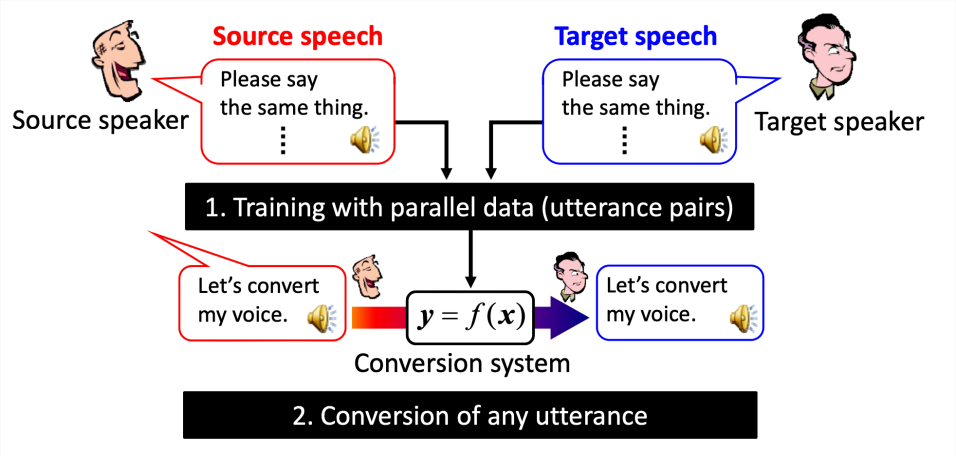
- Improvements on attention
- Richer information for encoder[8]
- Mispronunciation[9]
- Controllable generation[10]
- ...

[8] https://www.isca-speech.org/archive/Interspeech_2019/pdfs/3177.pdf
[9] https://www.isca-speech.org/archive/Interspeech_2019/pdfs/2830.pdf
[10] https://ai.googleblog.com/2018/03/expressive-speech-synthesis-with.html

Voice Conversion

Problem: transform the para-/non-linguistic characteristics included in a source speech waveform into a different one while preserving linguistic information

Voice Conversion

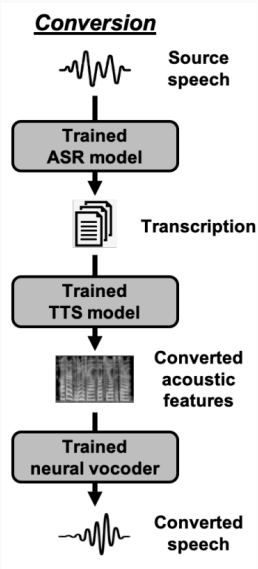


Voice Conversion

Applications:
(pls type in chat)

Voice Conversion

”Naive” approach:
Cascading ASR + TTS



Voice Conversion

Different possible models:

- feature conversion

- waveform generation

Voice Conversion

Evaluation:

Listening tests: naturalness and similarity

Speech Classification

- Language/Dialect Identification
- Speaker Verification / ID / Diarization
- Emotion Classification

Speech Classification – Language ID

i.e.: for use in call centers, or for Siri

models:

- HMMs (using i-vectors)
- Convolutional RNNs (CNN to extract spatial features, RNN/LSTM handle time and predict language)
- + attention

Speech Classification – Psychiatric Disorders

speech production is ***complex*** – slight physiological and cognitive changes potentially can produce noticeable acoustic changes

acoustic features linked to symptoms

- major depressive disorder: decrease in f0 and f0 range; also jitter, shimmer, f0 variablilty
- ptsd: slower, flatter; also reduced tonality in vowel space and f0 variability
- schizophrenia: total time talking, speech rate, mean pause duration, flat affect
- and more (see references)

Speech Classification – Psychiatric Disorder Detection

- SVMs or Gaussian Mixture models
- CNN on spectrograms

The End

Questions?

References

- Deep Learning for Human Language Processing course by Lee (2020), Lec. 2-7
- Expressive Speech Synthesis with Tacotron by Google AI (2018)
- Voice Conversion Challenge 2020
- Spoken Language Identification using ConvNets by Sarthak et al. (2019)
- Automated assessment of psychiatric disorders using speech: A systematic review by Low et al. (2020)