A Gentle Introduction to

Neural Text-to-Speech

Chae Young Lee
Oct 26, 2021

Table of Contents

- Introduction
  - ASR vs. TTS
  - Applications of TTS
  - Background on Speech Processing
  - TTS Evaluation, Datasets
  - Early Work
  - Modern TTS Pipeline
- Neural Frontend
  - Tacotron 1 & 2 (Paper 1)
  - FastSpeech 1 & 2 (Paper 2)
  - WaveGrad 2
  - Gan-TTS
  - Flowtron
- Prosody Modeling
  - What is prosody
  - Language Modeling
  - Style Modeling
- Adaptive TTS
  - Motivation
  - Multi-Speaker Vocoder
  - Adaptation to New Voice
  - Semi-Supervised Learning
- Conclusion
  - Conclusion
  - Future Work
  - Demo & Tools
  - Discussion

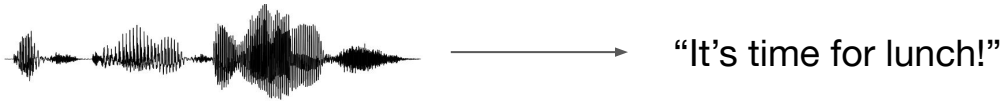
Introduction

Why speech?

- Natural interface of communications
- More spoken languages than written languages
- One of the earliest goals of computer language processing
  - Speech Recognition
  - Speech Synthesis
  - Speech Dialog System

ASR vs. TTS

Automatic Speech Recognition (ASR)

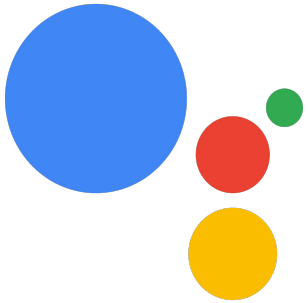


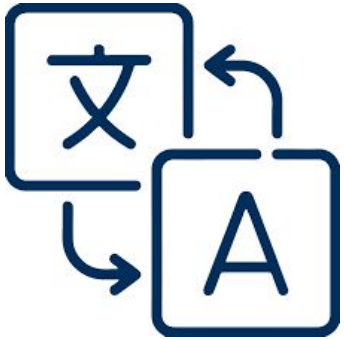
“It’s time for lunch!”

Speech Synthesis, Text-to-Speech (TTS)

“It’s time for lunch!”

Applications of TTS



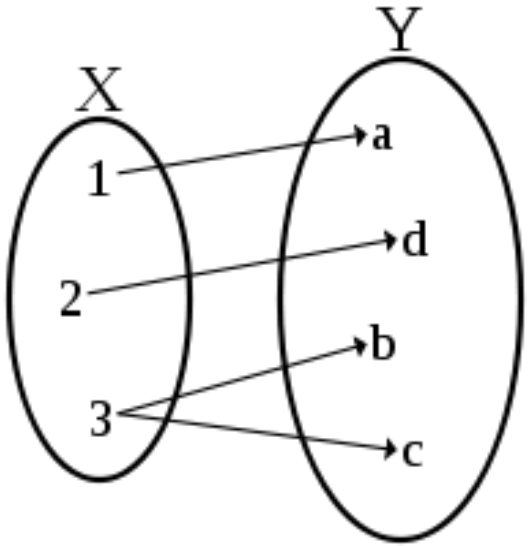


Why is TTS difficult?

- Non-standard words: “That’s 1314 gallons of milk to be delivered to 1314 St Andrew’s Dr. on 5/11/20”
- Tone / intonation: “Thanks for the help,” “Sure”
- Code switching: “Gracias for the lovely gift.”
- **Need both acoustic, language information**

One-to-Many Mapping

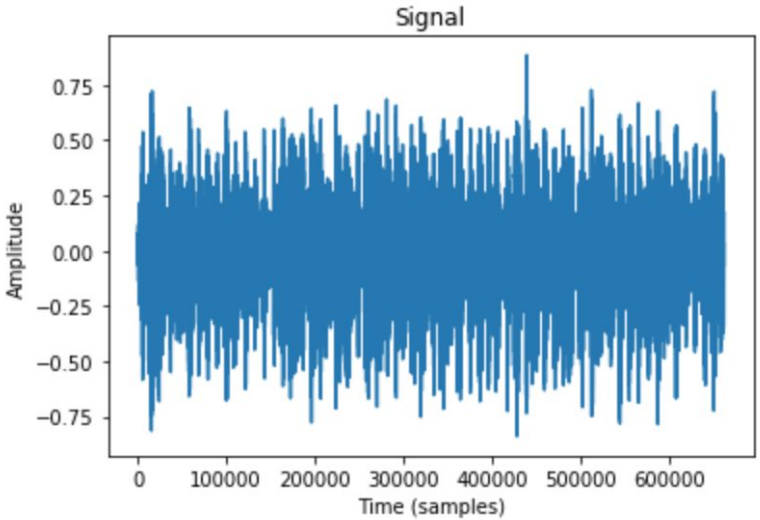
- There can be multiple forms of audio corresponding to a given text
- Variance in speaker (voice), pitch, energy, tone, accent, etc.
- Average of dataset vs. conditional modeling



Source: Wikitionary

What are speech signals?

- Variation of air pressure over time
- Analog-to-digital: most human speech below 10kHz (20kHz), telephone transmission below 4kHz (8kHz)



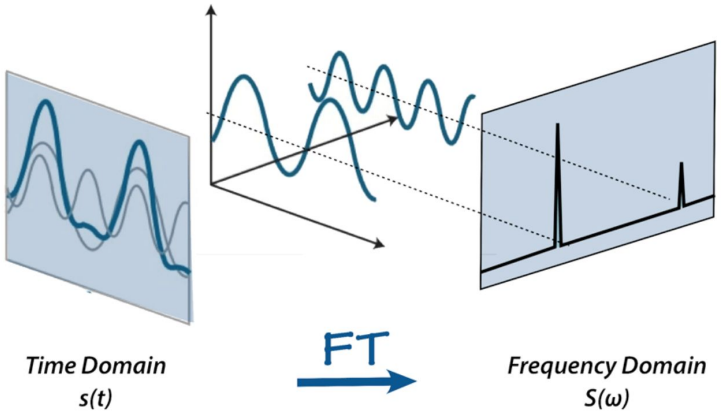
Source: Understanding the Mel Spectrogram (Medium)

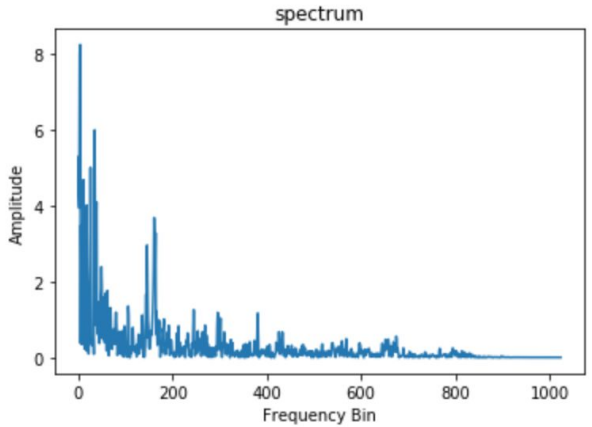
Quantization of speech signals

- Store amplitude measurement in integer
- 8 bit (values from -128 to 127) or 16 bit (values from -32768 to 32767)

The Fourier Transform

- Audio signal is composed of multiple single-frequency sound waves
- Fourier Transform decomposes signals into individual frequencies & amplitudes
- Fast Fourier Transform (FFT) used for efficient computation

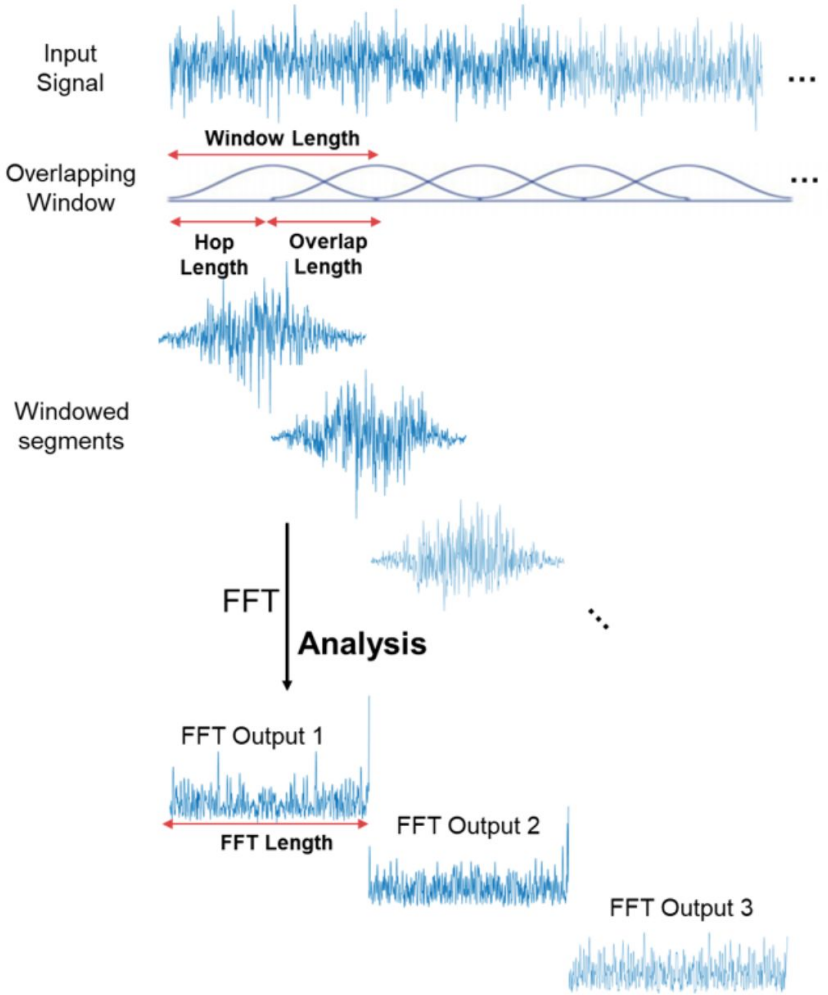




Source: https://aavos.eu/glossary/fourier-transform/c

Signals to Spectrogram

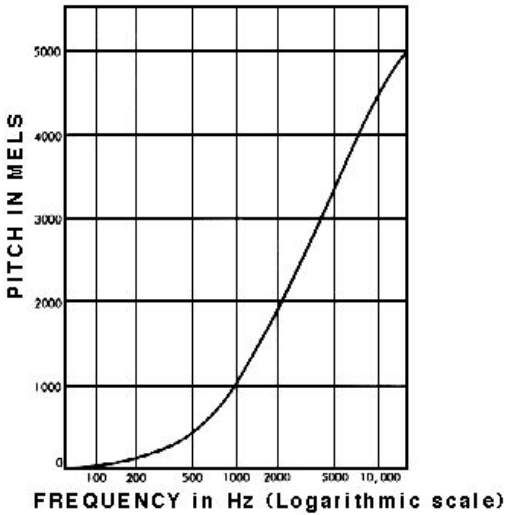
- Short-Time Fourier Transform (STFT) to represent non-periodic signals that vary over time

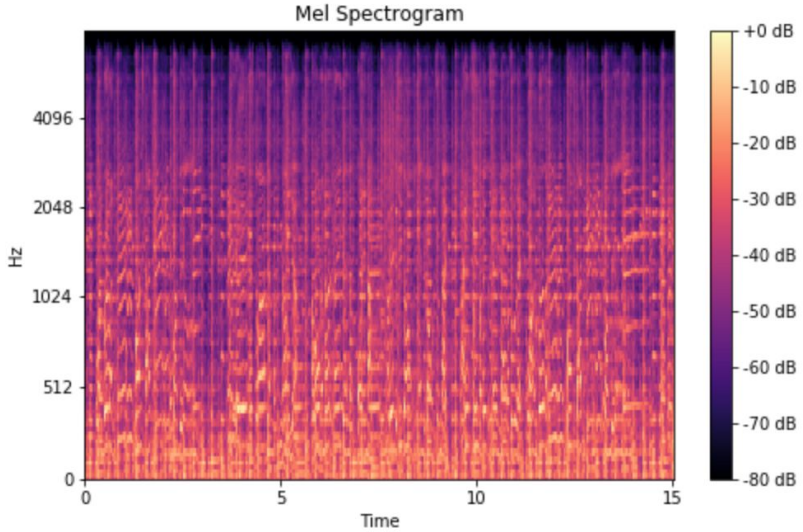


Source: MathWorks

Mel-Spectrogram

- Mel-scale is a unit of pitch such that the difference in pitch sounds equally distant to the listener





Source: Understanding the Mel Spectrogram (Medium), Simon Fraser University

Text Pre-Processing

- Text Normalization / Expansion
- Grapheme-to-Phoneme Conversion
  - Phoneme: the smallest unit of sound in a word (e.g. /k/)
  - Grapheme: the way we write a phoneme (e.g. /k/ → c, k, ck, qu, ch)

semiotic class	examples	verbalization
abbreviations	**gov't**, *N.Y., mph*	government
acronyms read as letters	**GPU**, *D.C., PC, UN, IBM*	G P U
cardinal numbers	**12**, *45, 1/2, 0.6*	twelve
ordinal numbers	*May* **7**, *3rd, Bill Gates III*	seventh
numbers read as digits	*Room* **101**	one oh one
times	*3.20,* **11:45**	eleven forty five
dates	**28/02** *(or in US, 2/28)*	February twenty eighth
years	**1999**, *80s, 1900s, 2045*	nineteen ninety nine
money	**$3.45**, *€250, $200K*	three dollars forty five
money in tr/m/billions	**$3.45 billion**	three point four five billion dollars
percentage	**75%** *3.4%*	seventy five percent

Figure 26.13 Some types of non-standard words in text normalization; see Sproat et al. (2001) and (van Esch and Sproat, 2018) for many more.

Source: Jurafsky and Martin, Speech and Language Processing

Prosody

- Study of the **intonational** and **rhythmic** aspects of language
- Use of F0, energy, and duration to convey pragmatic, affective, or conversation-interactional meanings
- Why important?
  - Mark discourse structure
  - Mark saliency of word / phrase
  - Convey emotions

or a final drop in F0 (called a **final fall**) to indicate a declarative intonation:

Languages make wide use of tune to express meaning (Xu, 2005). In English,

Source: Jurafsky and Martin, Speech and Language Processing

Evaluation

- Mean Opinion Score (MOS)
  - Human rating from 1 to 5 in evaluating the general speech quality
- AB Testing
  - Directly compare two systems and choose the preferred one
- Comparison MOS (CMOS)
  - Rating from 0 to 3 in evaluating the magnitude of the difference (the higher the bigger difference)

Rating	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

MOS Score Chart

Source: CROWDMOS

Evaluation

Mean Opinion Score (MOS)

Method	MOS
GT	4.30 ± 0.07
GT (Mel + PWG)	3.92 ± 0.08
Tacotron 2 (Shen et al., 2018) (Mel + PWG)	3.70 ± 0.08
Transformer TTS (Li et al., 2019) (Mel + PWG)	3.72 ± 0.07
FastSpeech (Ren et al., 2019) (Mel + PWG)	3.68 ± 0.09
FastSpeech 2 (Mel + PWG)	3.83 ± 0.08
FastSpeech 2s	3.71 ± 0.09

(a) The MOS with 95% confidence intervals.

Method	CMOS
FastSpeech 2	0.000
FastSpeech	-0.885
Transformer TTS	-0.235

(b) CMOS comparison.

Source: FastSpeech2

TTS Datasets

- LJSpeech (English, 1 female, 24 hours, Ito & Johnson)
- LibriTTS (English, 2456 speakers, 585 hours, Google Research)
- VCTK (English, 109 speakers, 44 hours, University of Edinburgh)
- AISHELL-3 (Mandarin, 218 speakers, 85 hours, Beijing Shell Shell Tech)
- KSS Dataset (Korean, 1 female, 12 hours, Park)
- JSUT (Japanese, 1 female, 10 hours, Takamichi)
- Pavoque (German, 1 male, 12 hours, DFKI GmbH)
- Tundra (14 European), M-AILABS (9 European), CSS10 (10 languages)
- LibriVox (crowdsourced)

Speech Conferences

- ICASSP (The International Conference on Acoustics, Speech, and Signal Processing): IEEE; every June; topics include acoustics, speech, and signal processing.
- Interspeech: every late August/early September; all areas related to the science and technology of speech communication.
- ICLR (International Conference on Learning Representations): every May; deals with machine learning in general.
- ICML (International Conference on Machine Learning): every July.
- NeurIPS (Neural Information Processing Systems): every December; all aspects of the use of machine learning networks and artificial intelligence.

Early Work

- Von Kempelen’s synthesizer in 1791
  - Blows air through physical tubes

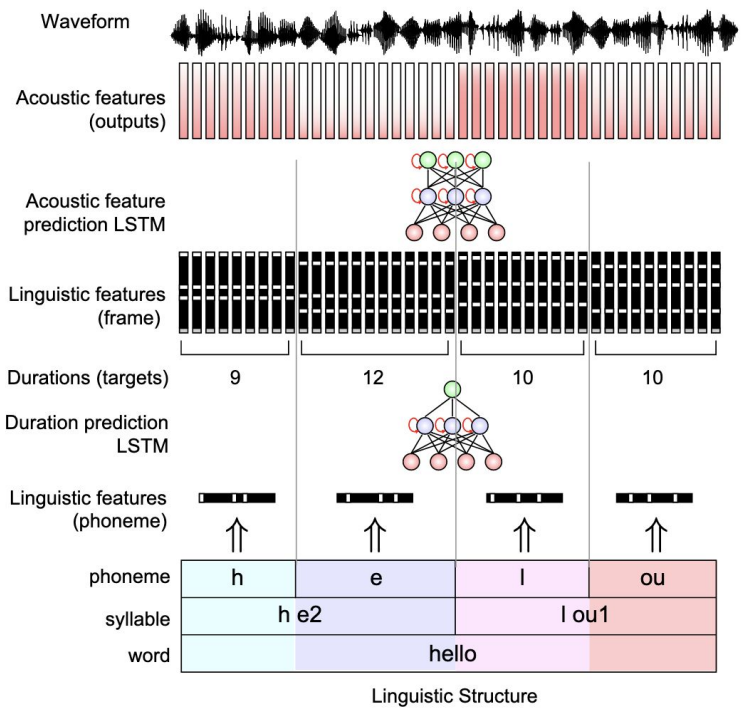
Source: Deutsches Museum in Munich

Early Work

- Formant Synthesis (60s-80s)
  - Waveform construction from components
- Diphone Synthesis (80s-90s)
  - Waveform by concatenation of small number of instances of speech
- Unit Selection (90s-00s)
  - Waveform by concatenation of very large number of instances of speech
- Statistical Parametric Synthesis (00s-2016)
  - Waveform construction from parametric models

Early Work - Statistical Parametric Models

- Independently trained modules
- Error accumulate across modules
- Extensive domain expertise, laborious to design

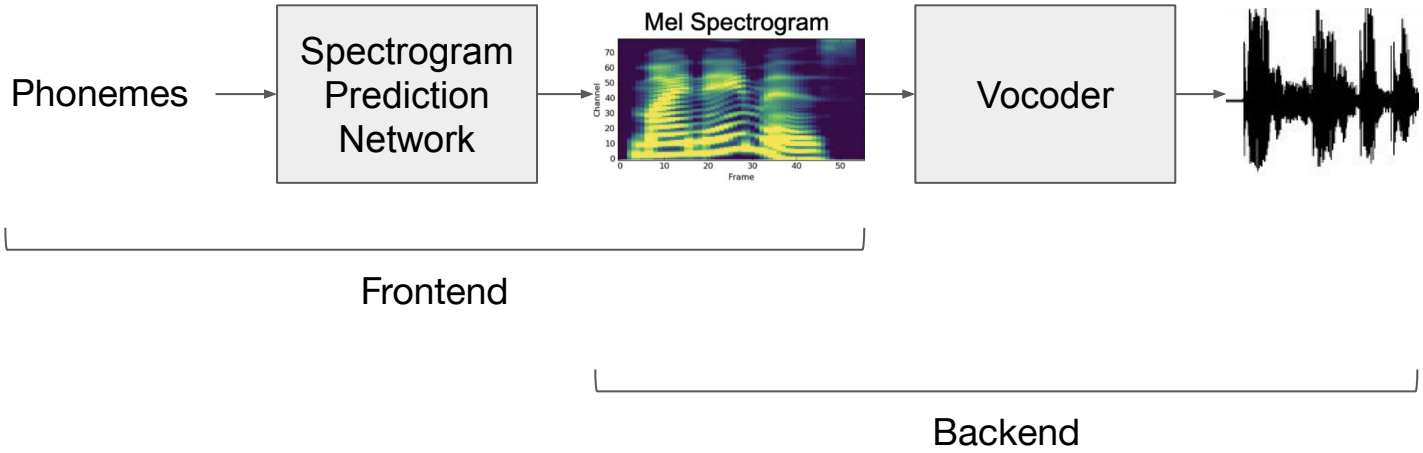


Source: Heiga Zen et al, 2016, [Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices]

Modern TTS Pipeline: Data-Driven

- Wang et al. (Sep 2016) - seq2seq attention + HMM alignment
- WaveNet (Sep 2016) - neural vocoder, acoustic model
- Char2Wav (Feb 2017) - end-to-end TTS
- DeepVoice (Feb 2017) - every TTS component becomes a neural net
- Tacotron (Mar 2017) - fully seq2seq, trained from scratch
- **Tacotron 2** (Dec 2017) - mel-spectrogram, WaveNet
- Transformer TTS (Sep 2018) - Transformer based
- ParaNet (May 2019) - non-autoregressive TTS, flow based
- FastSpeech (May 2019) - non-autoregressive TTS, Transformer TTS based
- **FastSpeech 2** (Jun 2020) - SOTA

Modern TTS Pipeline: 2-Step



Why two papers?

- Tacotron 2
  - Achieves MOS of 3.70
  - Completes the modern TTS architecture using seq2seq encoder-decoder architecture + neural vocoder
  - First time achieving MOS comparable to GT (3.92); makes a huge leap in quality by replacing statistical models
- FastSpeech 2
  - Achieves MOS of 3.83
  - Changes the paradigm of TTS architecture by using Transformer-based encoder and decoder, goes fully non-autoregressive
  - High quality, fast, takes less memory; makes industry scaling possible

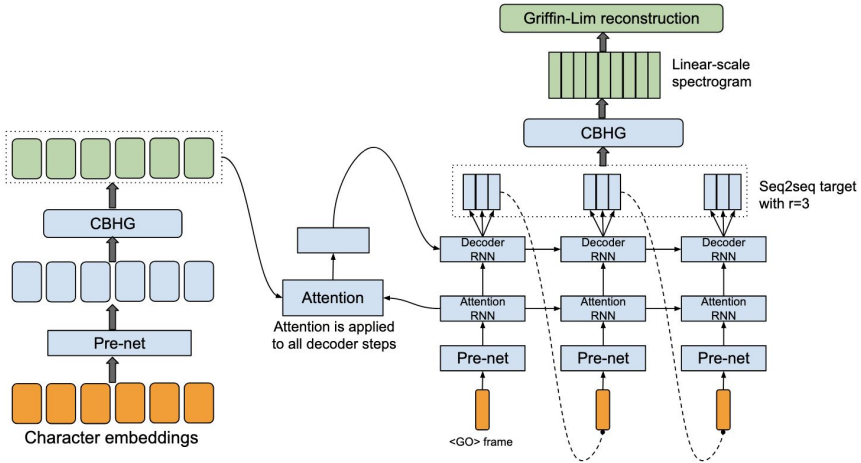
Neural Frontend

Tacotron: Towards End-to-End Speech Synthesis

Wang et al. (Google), Interspeech 2017

Overview

- An end-to-end generative TTS model that can be trained completely from scratch with <text, audio> pairs
- Achieves MOS of 3.82 on internal US English dataset (parametric system 3.69)
- Generates speech at frame level, so substantially faster than sample-level models
- Based on seq2seq with attention: encoder + attention decoder + post-net

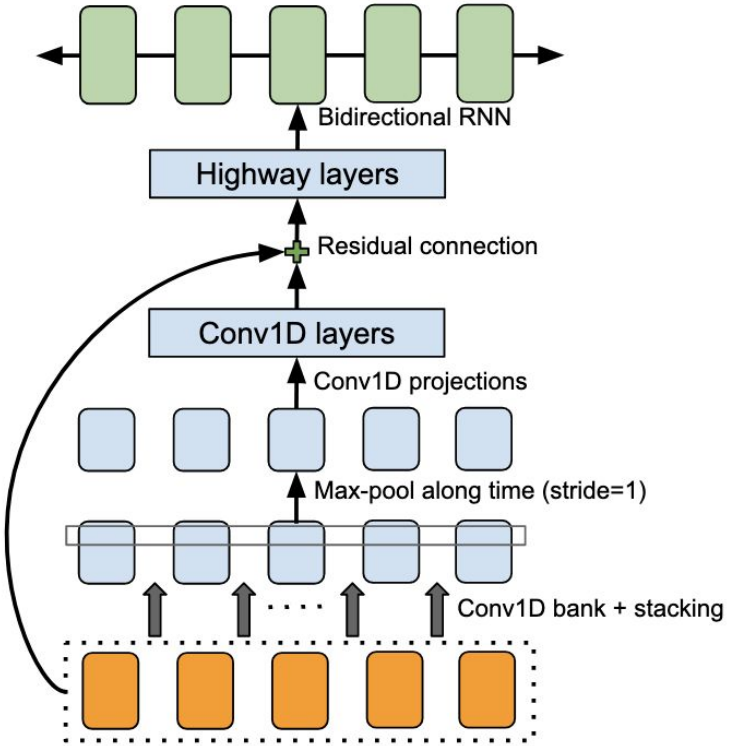


Encoder


Phonemes
1-hot Vector
Embedding
(256-D)
Pre-Net
(2 FC
bottleneck)
CBHG
Module
Encoder
Hidden

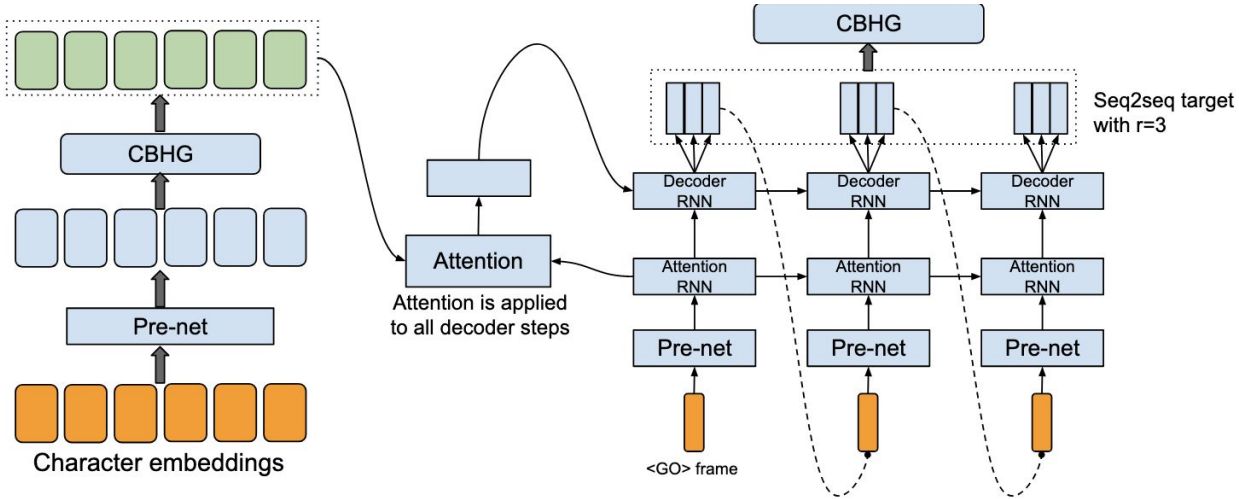
Encoder: CBHG Module

- Adapted from machine translation (Lee et al., 2016)
- 1D Convolution bank (local, contextual) with residual connection
- Highway network (Srivastava et al., 2015) for high-level features
- Bi-directional GRU RNN (sequential features)



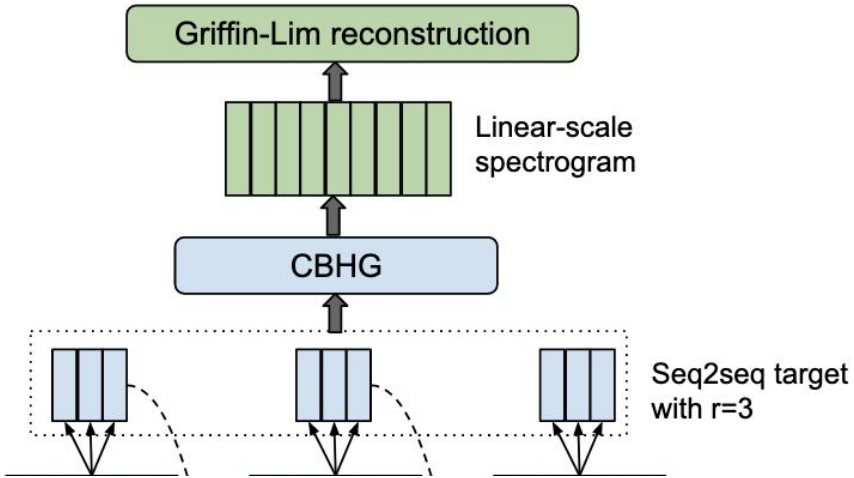
Decoder

- Content-based tanh attention decoder
  - Stateful recurrent layer produces the attention query at each decoder time step
  - Concatenate the context vector and the attention RNN cell as input for the decoder RNN
- Predicts 80 band mel-scale spectrogram
  - Predict multiple, non-overlapping output frames at each decoder step



Post-Processing Net

- CBHG predicts linear-scale spectrogram
- Griffin-Lim reconstruction + inverse STFT converts it to waveform



Problems with Tacotron

- Griffin-Lim produces characteristic artifacts and low audio quality
- Complex architecture: CBHG module, prenets, GRUs

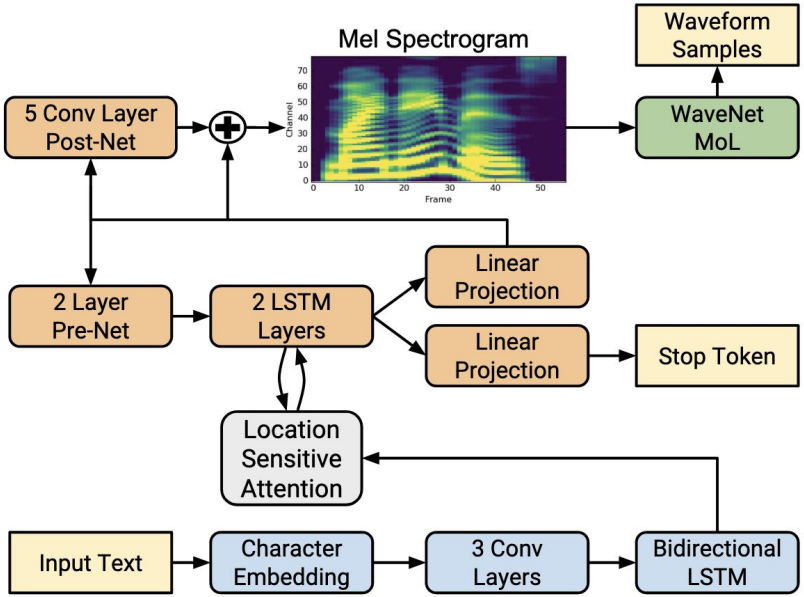
Paper 1

Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

Shen et al. (Google), ICASSP 2018

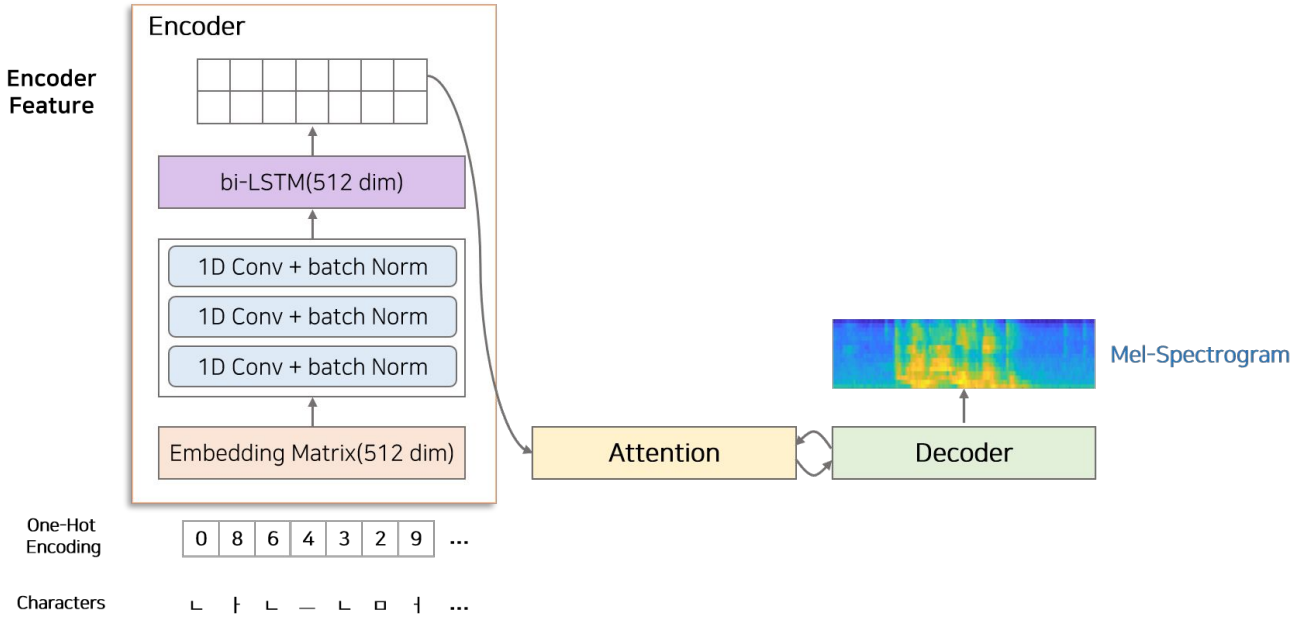
Overview

- Tacotron-like seq2seq network
- Modified WaveNet to generate time-domain waveform samples



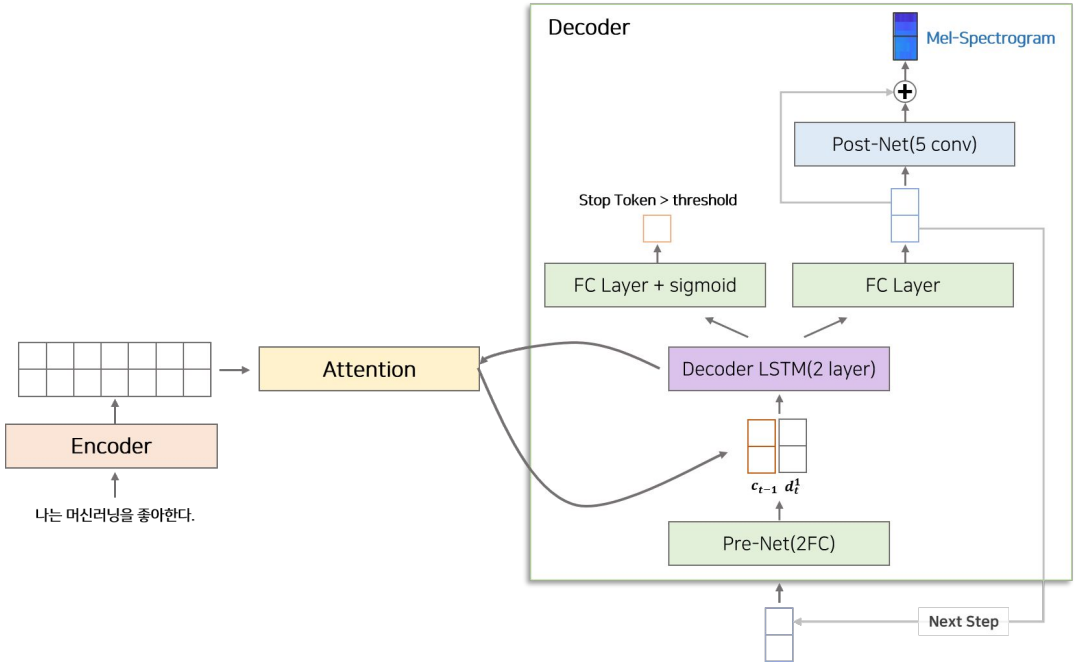
Phoneme Encoder

- Simplified building blocks



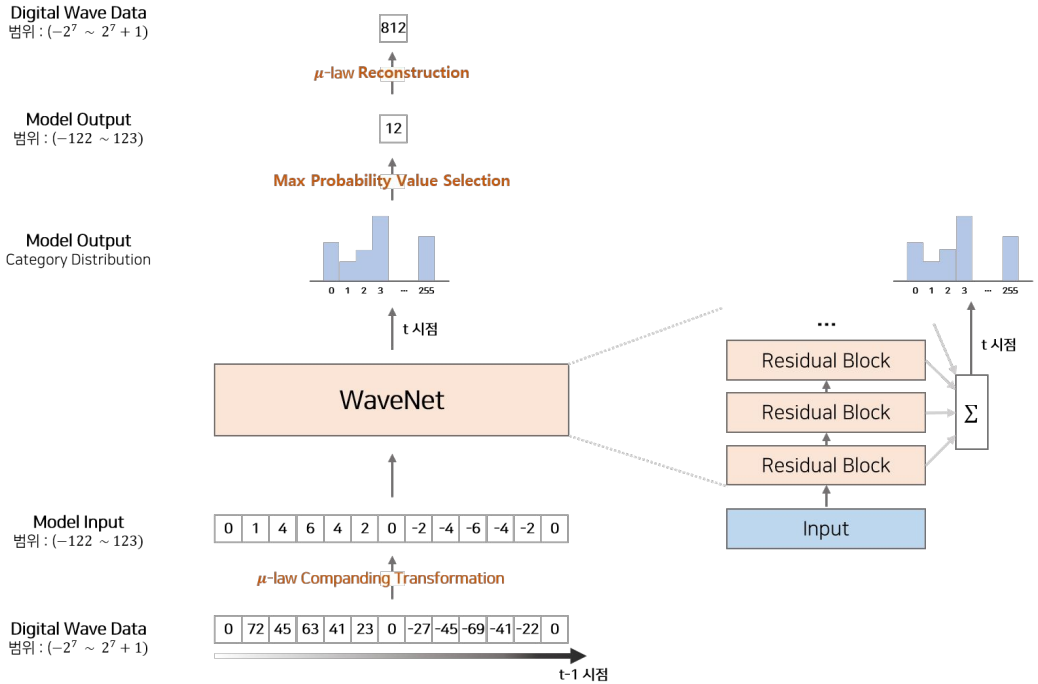
Attention-Based Decoder

- Location-sensitive attention
- Each decoder predicts single spectrogram frame



Revisit: WaveNet (DeepMind, 2016)

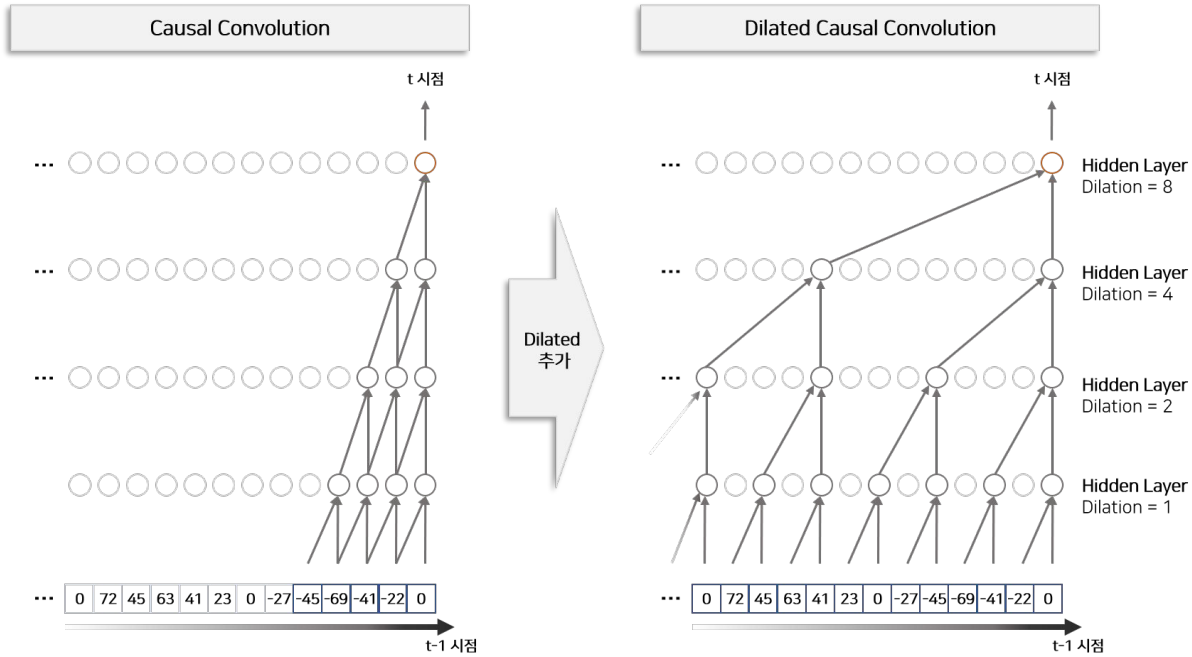
Softmax Distribution



Source: [joungheekim's blog]

Revisit: WaveNet (DeepMind, 2016)

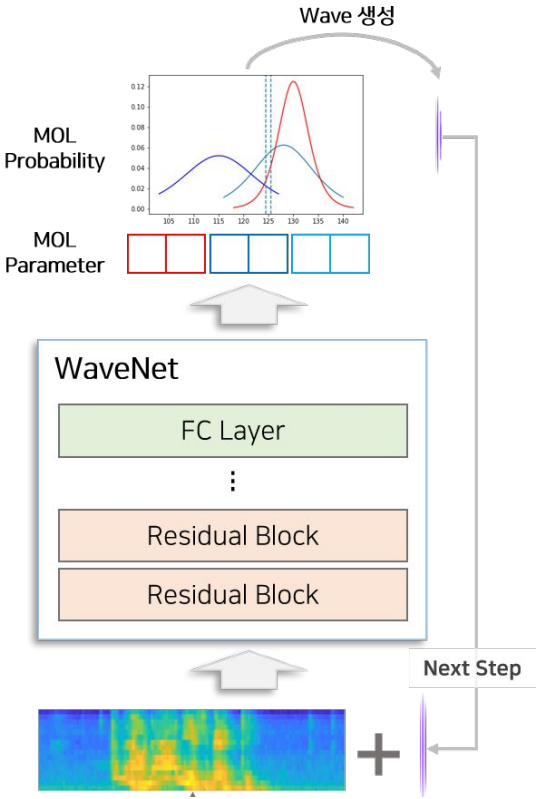
Dilated Causal Convolution



Source: [joungheekim's blog]

Neural Vocoder

- From WaveNet: dilated conv
- PixelCNN++ & Parallel WaveNet: use a 10-component MoL to generate 16-bit samples at 24 kHz
- Negative log likelihood training



Experiments

- Internal US English dataset of single female speaker (24.6 hours)
- https://google.github.io/tacotron/publications/tacotron2/

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Ablation Studies

- Linear Spectrogram vs. Mel-Spectrogram
- Griffin-Lim Reconstruction vs. WaveNet

System	MOS
Tacotron 2 (Linear + G-L)	3.944 ± 0.091
Tacotron 2 (Linear + WaveNet)	4.510 ± 0.054
Tacotron 2 (Mel + WaveNet)	$\mathbf{4.526 \pm 0.066}$

Simplifying WaveNet

- Model can thrive with shallow net with smaller receptive field
- Dilation is necessary for sufficient context

Total layers	Num cycles	Dilation cycle size	Receptive field (samples / ms)	MOS
30	3	10	6,139 / 255.8	4.526 ± 0.066
24	4	6	505 / 21.0	4.547 ± 0.056
12	2	6	253 / 10.5	4.481 ± 0.059
30	30	1	61 / 2.5	3.930 ± 0.076

FastSpeech: Fast, Robust and Controllable Text to Speech

Ren et al. (Zhejiang University & Microsoft Research Asia),
NeurIPS 2019

Issues with neural TTS

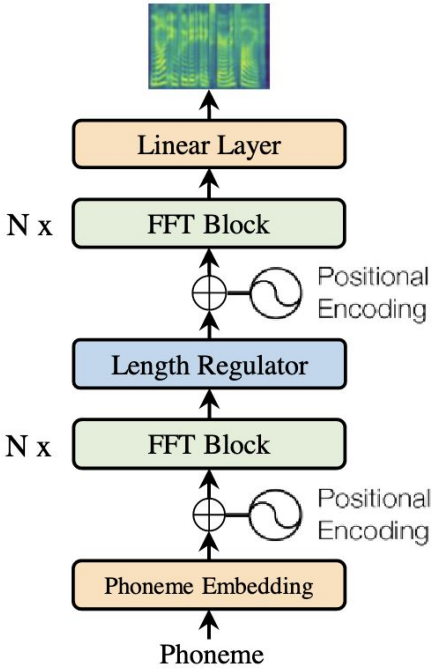
- Slow inference speed
- Usually not robust: some words are skipped or repeated
- Lack control over voice speed or prosody

Overview

- Feed-forward network based on Transformer to generate mel-spectrogram in parallel
- Predicts alignment map to expand phoneme embedding to match the length of target mel-spectrogram
- MOS comparable to autoregressive models, but faster: mel-spectrogram generation by 270x, e2e by 38x

Feed-Forward Transformer (FFT)

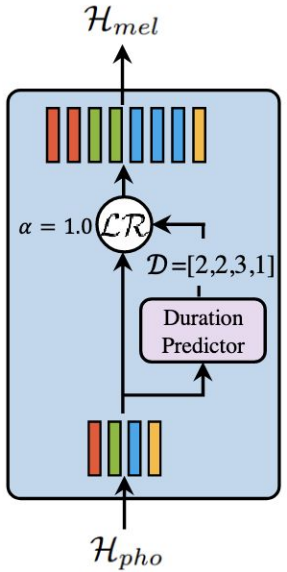
- N-stacked FFT blocks for encoding and decoding
- FFT block is based on self-attention and 1D convolution
  - Multi-head attention extracts cross-position information
  - 2-layer 1D conv (ReLU) to extract adjacent hidden states



(a) Feed-Forward Transformer (b) FFT Block

Length Regulator

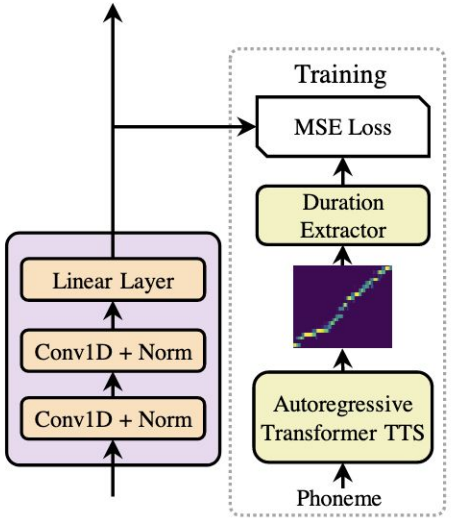
$$\mathcal{H}_{pho} = [h_1, h_2, ..., h_n],$$
$$\mathcal{D} = [d_1, d_2, ..., d_n],$$
$$\mathcal{H}_{mel} = \mathcal{LR}(\mathcal{H}_{pho}, \mathcal{D}, \alpha)$$



(c) Length Regulator

Duration Predictor

- Predicts duration from hidden phoneme representation
- Extracting the target alignment for training
  - Train Transformer TTS as teacher model
  - Extract alignment from a diagonal attention head, where duration of a phoneme is # mel-spectrograms attended to it
  - Accordingly, mel-spectrogram from teacher model as target mel-spectrogram



(d) Duration Predictor

Issues with FastSpeech

- Teacher-student training
  - Complicated
  - Target mel-spectrograms generated from the teacher model have information loss
  - Duration extracted from attention map is not accurate enough

Paper 2

FastSpeech 2: Fast and High-Quality End-to-End Text to Speech

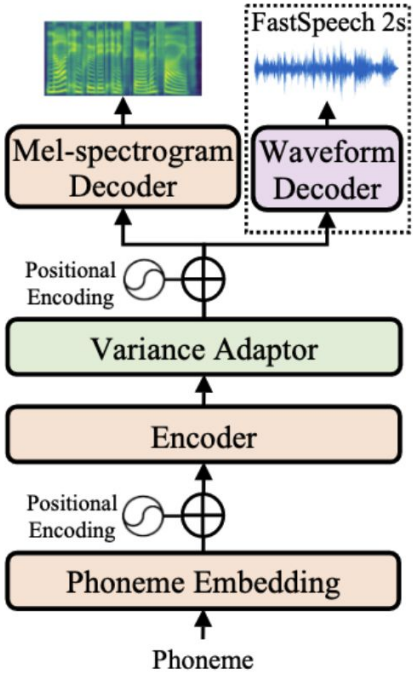
Ren et al. (Zhejiang University & Microsoft Research Asia),
ICLR 2021

Overview

- Directly trains the model with GT target
- Introduce more variation information of speech as conditional inputs
- Design FastSpeech 2s, the first E2E parallel TTS generation
- MOS of 3.83 (SOTA), 3x training speed-up from FS1

Improvements from FS1

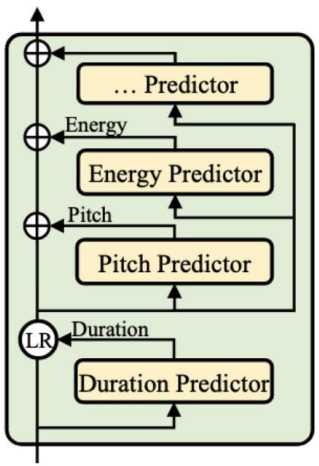
- Removed teacher-student distillation pipeline and directly use GT mel-spectrogram as target
- 3 types of variance adaptors for duration, pitch, and energy
  - GT phoneme duration from Forced Alignment (e.g. Montreal Forced Alignment)
  - Pitch and energy extracted from GT audio



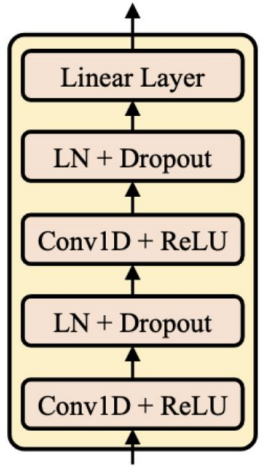
(a) FastSpeech 2

Variance Adaptor

- One-to-many mapping
  - Duration: how long the speech voice sounds
  - Pitch: conveys emotions and affects speech prosody
  - Energy: indicates frame-level magnitude of mel-spectrograms and directly affects volume/prosody
  - More variance adaptors can be added for finer control
- GT data during training with MSE loss, predicted values during inference



(b) Variance adaptor



(c) Duration/pitch/energy predictor

Experiments

- LJSpeech Dataset (US English, single female speaker, 24 hours)
- https://speechresearch.github.io/fastspeech2/

Method	MOS
GT	4.30 ± 0.07
GT (Mel + PWG)	3.92 ± 0.08
Tacotron 2 (Shen et al., 2018) (Mel + PWG)	3.70 ± 0.08
Transformer TTS (Li et al., 2019) (Mel + PWG)	3.72 ± 0.07
FastSpeech (Ren et al., 2019) (Mel + PWG)	3.68 ± 0.09
FastSpeech 2 (Mel + PWG)	3.83 ± 0.08
FastSpeech 2s	3.71 ± 0.09

(a) The MOS with 95% confidence intervals.

Method	CMOS
FastSpeech 2	0.000
FastSpeech	-0.885
Transformer TTS	-0.235

(b) CMOS comparison.

Experiments

- On 1 NVIDIA V100 GPU, batch size 48 for training and 1 for inference
- Speedups in training, inference
- Training time x3 than FS 1, inference speed x49 than Transformer TTS

Method	Training Time (h)	Inference Speed (RTF)	Inference Speedup
Transformer TTS (Li et al., 2019)	38.64	9.32×10^{-1}	/
FastSpeech (Ren et al., 2019)	53.12	1.92×10^{-2}	48.5×
FastSpeech 2	**17.02**	1.95×10^{-2}	47.8×
FastSpeech 2s	92.18	$\mathbf{1.80 \times 10^{-2}}$	**51.8**×

Duration Comparison

- Duration from teacher model and MFA

Method	Δ (ms)
Duration from teacher model	19.68
Duration from MFA	12.47

(a) Alignment accuracy comparison.

Setting	CMOS
FastSpeech + Duration from teacher	0
FastSpeech + Duration from MFA	+0.195

(b) CMOS comparison.

Table 5: The comparison of the duration from teacher model and MFA. Δ means the average of absolute boundary differences.

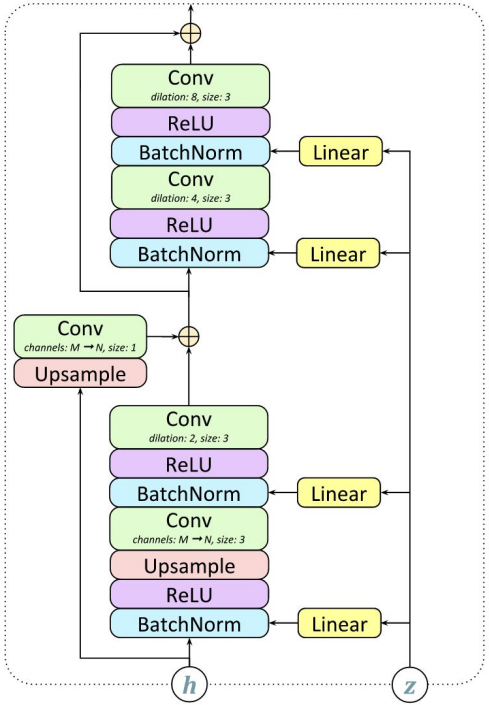
Ablation Studies

- Effectiveness of variance information

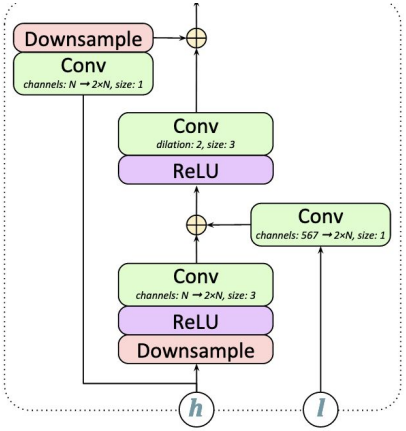
Setting	CMOS
FastSpeech 2	0
FastSpeech 2 - energy	-0.040
FastSpeech 2 - pitch	-0.245
FastSpeech 2 - pitch - energy	-0.370

And more frontends...

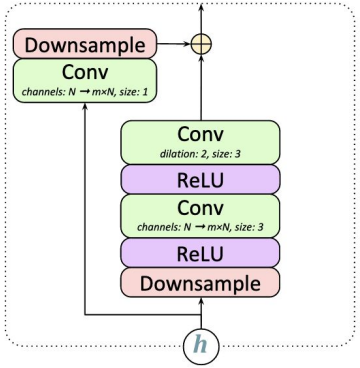
GAN-TTS (2019)



(a) GBlock

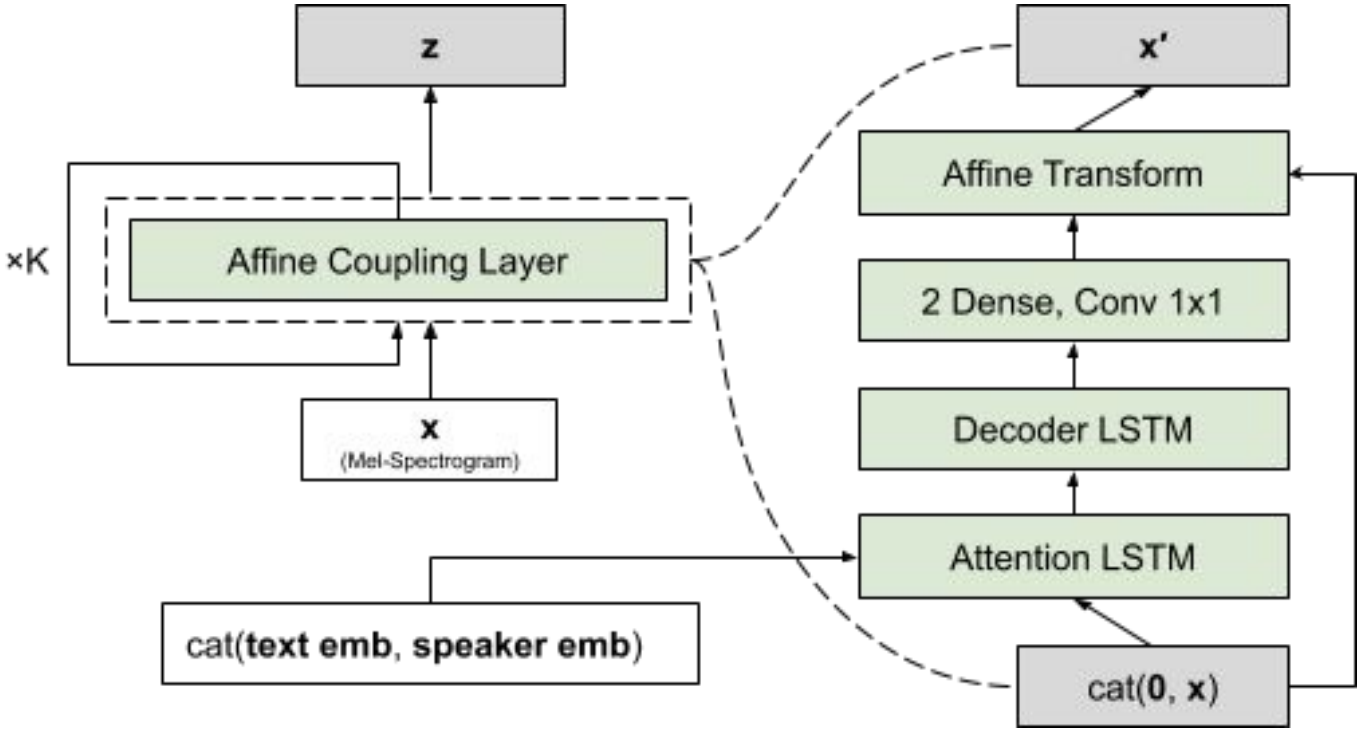


(b) Conditional DBlock

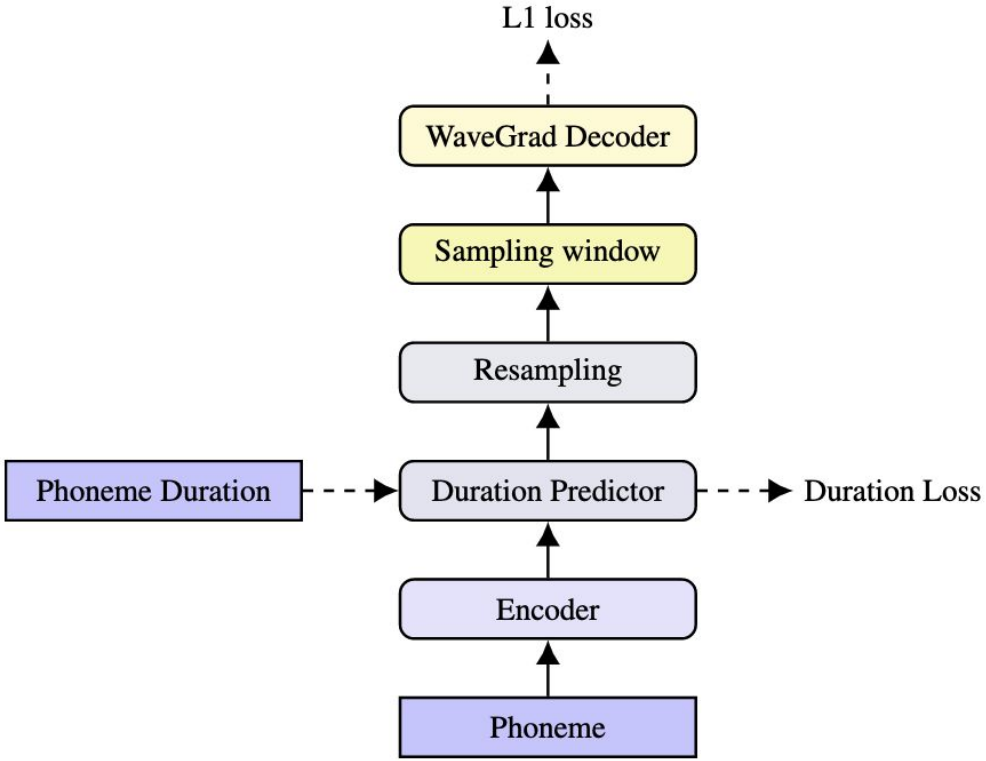


(c) DBlock

Flowtron (2020)



WaveGrad 2 (2021)



Summary

- Different encoding/decoding techniques
  - Seq2seq
  - Transformers
  - Gaussian upsampling
  - GAN
  - Flow models
- Different pipeline
  - <text, spectrogram>
  - <spectrogram, audio>
  - <text, audio>

Discussion

- How relatable are NLP and TTS?
- How are TTS models different from machine translation models?
- What are some obstacles and possible solutions in handling long utterances in TTS?

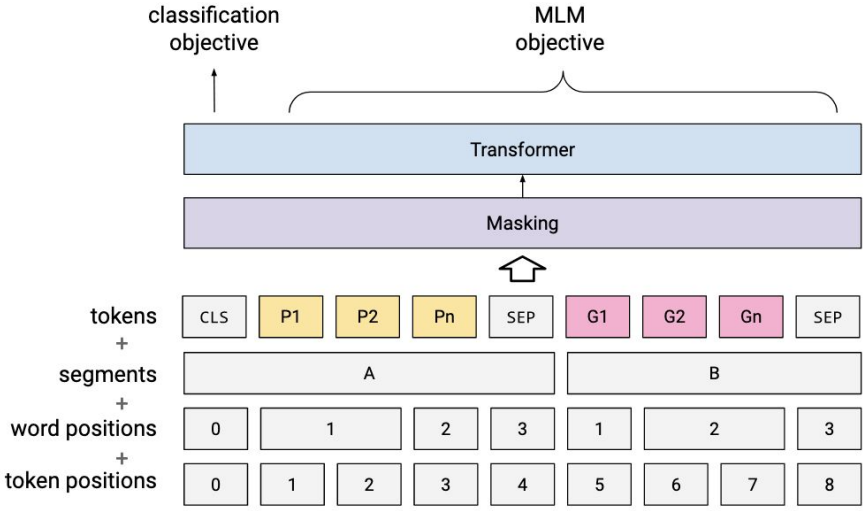
Prosody Modeling

Prosody

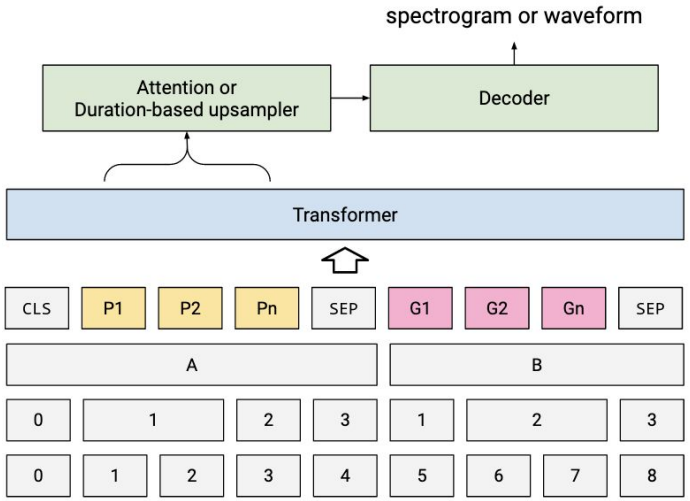
- Non-textual information of speech (e.g. rhythm, intonation)
- Important for natural speech
- Unsupervised learning problem
- Different granularity of control
  - Global vs. local
  - Coarse vs. fine
- Disentanglement from text, speakers, etc.

Language Modeling

PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS (Jia et al, Google Research, Interspeech 2021)



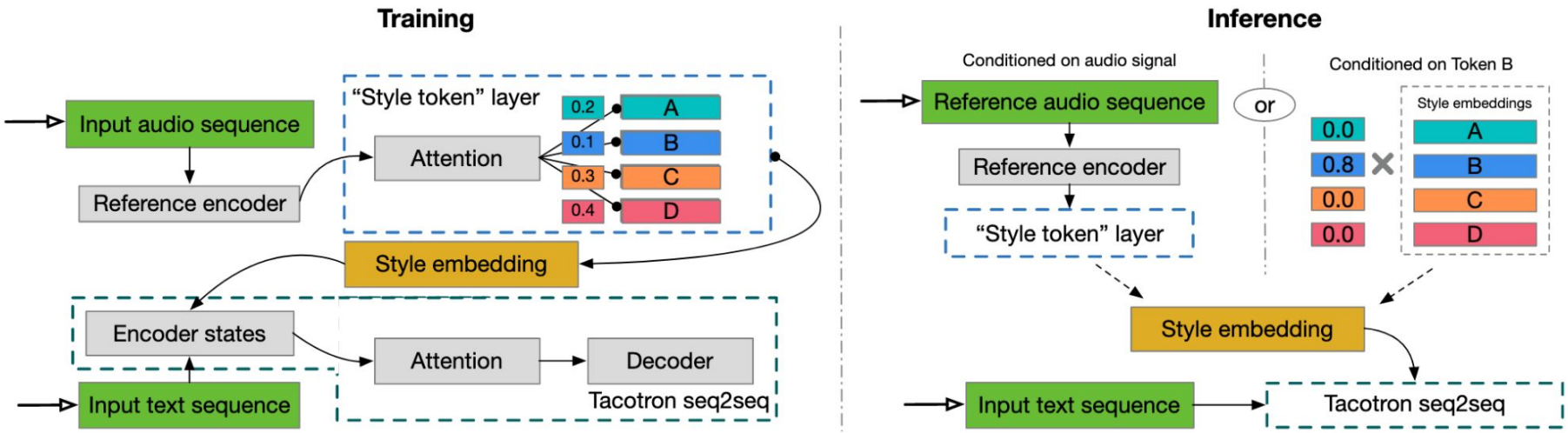
(a) PnG BERT and pre-training objectives.



(b) Using PnG BERT as the encoder for a neural TTS model.

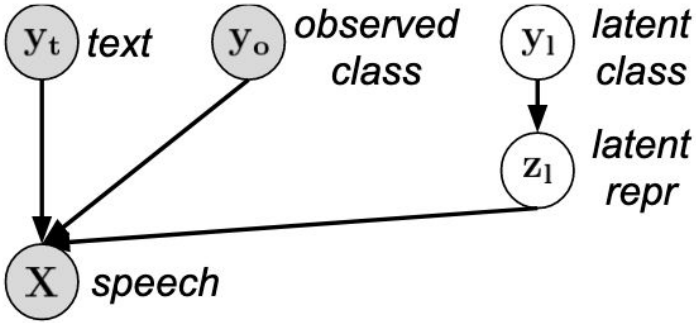
Style Modeling

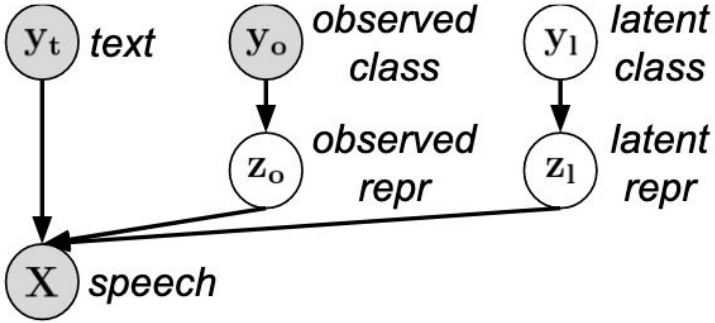
Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis (Wang et al, Google, 2018)



Style Modeling

Hierarchical Generative Modeling for Controllable Speech Synthesis (Hsu et al, MIT & Google, ICLR 2019)





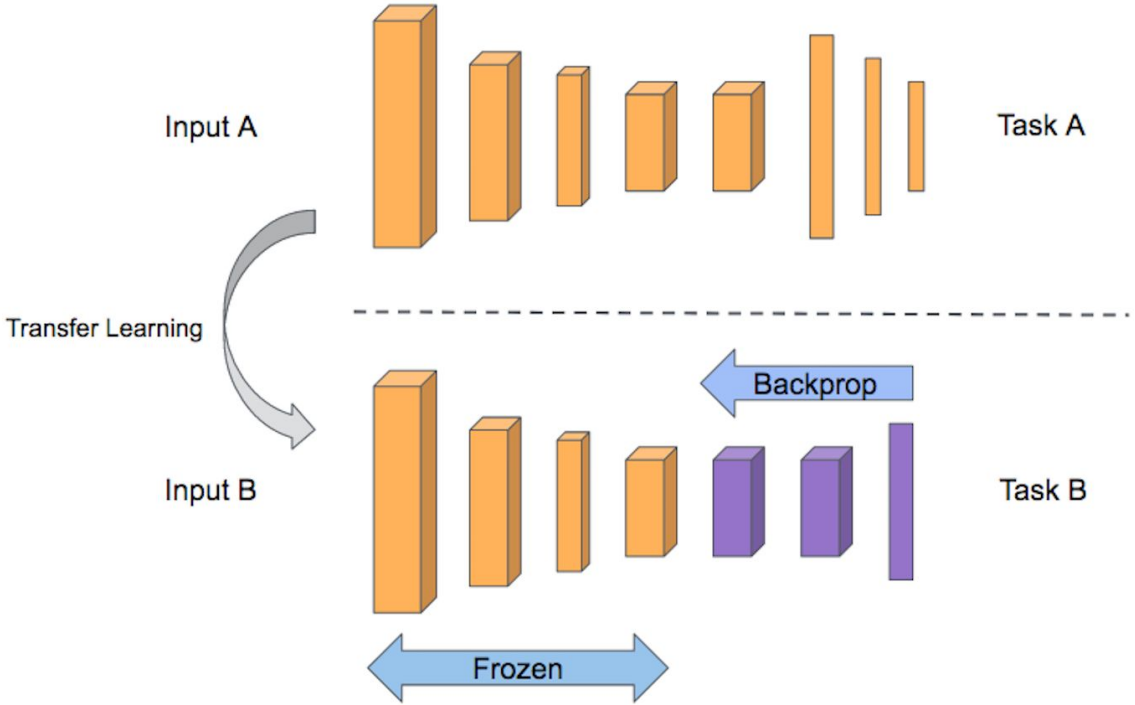
Discussion

- What potential issues do you find from unsupervised training methods?
- At what scenarios would prosody control/transfer be useful?
- How can we anchor useful styles in unsupervised learning?

Adaptive TTS

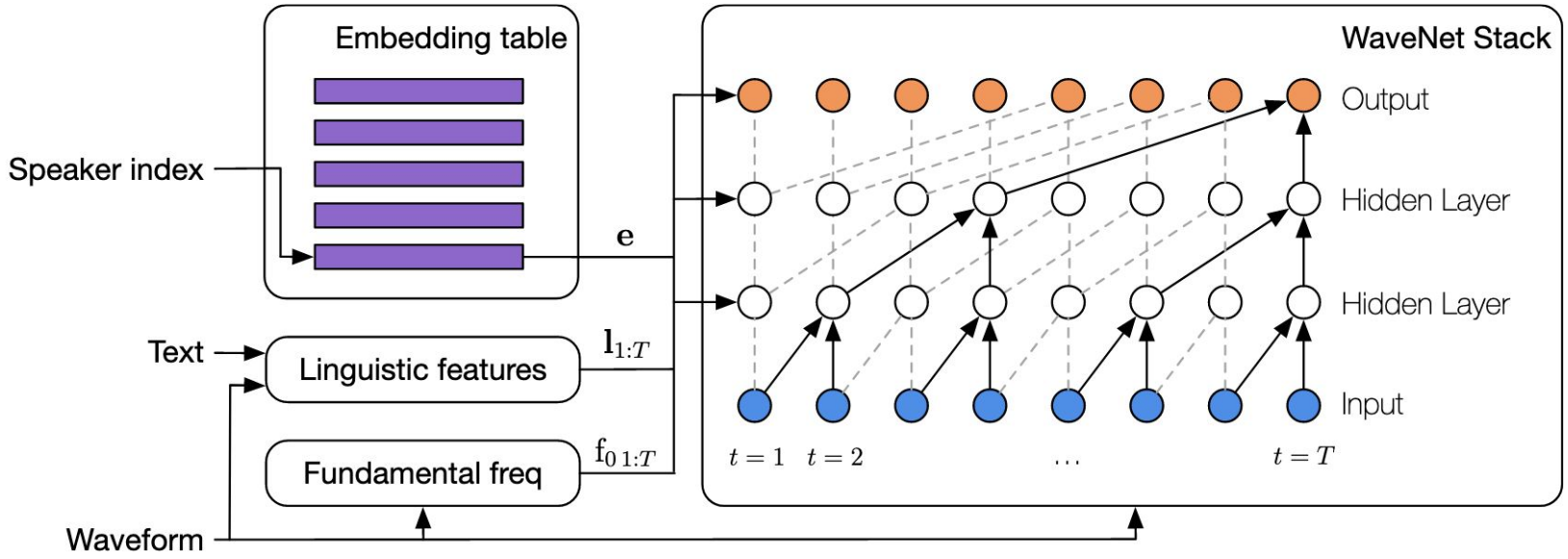
Why Adaptation?

- Quality improvement
- Scalability / data size
- Task optimization



Multi-Speaker Vocoder

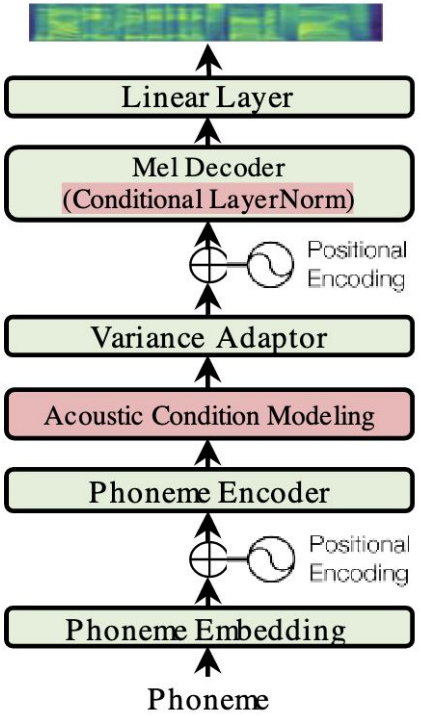
Sample Efficient Adaptive Text-to-Speech (Chen et al., DeepMind & Google, ICLR 2019)



Adaptation to New Voice

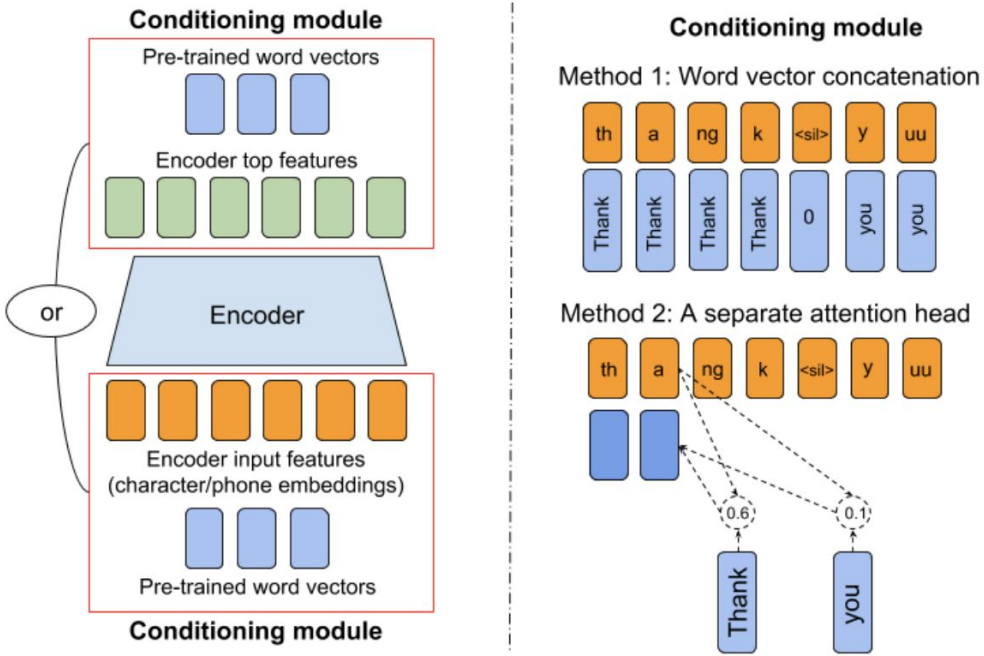
AdaSpeech: Adaptive Text to Speech for Custom Voice (Chen et al, Microsoft, ICLR 2021)

- Multiple granularity of acoustic conditioning
- Fine-tune conditional layer normalization



Semi-Supervision

Semi-Supervised Training for Improving Data Efficiency in End-to-End Speech Synthesis (Chung et al., MIT & Google, 2018)



Discussion: Deepfakes

Discussion

- How can we detect Deepfakes in terms of TTS? Can you find the difference between real speech and synthesized speech?
- Do you think a single multi-speaker model will outperform its single-speaker counterpart?
- What are some obstacles in scaling TTS models to different languages?
- What are some things you would have to take into account when creating a TTS dataset (record voice of voice actors)?

Conclusion

Conclusion

- Difficult task to solve
  - Complex pipelines
  - Huge models
  - Slow training and inference
- Various modeling objectives
  - Frontend, backend
  - Prosody
  - Multi-speaker, multi-lingual
  - Adaptation / task optimization
- Industry driven research area
  - Led by teams in Google, Microsoft, Amazon (maybe Apple)
  - Ideas open-source, industry scaling as difficult engineering problem

Future Work

- Improved vocoder architecture for better speech quality
- Intuitive and global control of prosody
- Towards end-to-end, multi-speaker, multilingual TTS model

Tools & Demo

- https://ttstool.com/
- https://www.ispeech.org/
- Google Cloud APIs
- Live FastSpeech 2 Demo

```
(chae) cl2322@tangra:/data/lily/cl2322/FastSpeech2$ CUDA_VISI
BLE_DEVICES=1 python synthesize.py --text "The forms of print
ed letters should be beautiful, and that their arrangement on
 the page should be reasonable and a help to the shapeliness
of the letters themselves." --restore_step 900000 --mode sing
le -p config/LJSpeech/preprocess.yaml -m config/LJSpeech/mode
l.yaml -t config/LJSpeech/train.yaml
Removing weight norm...
Raw Text Sequence: The forms of printed letters should be bea
utiful, and that their arrangement on the page should be reas
onable and a help to the shapeliness of the letters themselve
s
Phoneme Sequence: {DH AH0 F AO1 R M Z AH0 V P R IH1 N AH0 D L
 EH1 T ER0 Z SH UH1 D B IY0 B Y UW1 T AH0 F AH0 L sp AE1 N D
DH AE1 T DH EH1 R ER0 EY1 N JH M AH0 N T AA1 N DH AH0 P EY1 J
H SH UH1 D B IY0 R IY1 Z AH0 N AH0 B AH0 L AE1 N D AH0 HH EH1
 L P T AH0 DH AH0 SH EY1 P L IY0 N AH0 S AH0 V DH AH0 L EH1 T
 ER0 Z DH AH0 M S EH1 L V Z}
(chae) cl2322@tangra:/data/lily/cl2322/FastSpeech2$
```