神经语音合成前沿

秦涛 微软亚洲研究院

MLA 2020

http://research.microsoft.com/~taoqin

alexa

Hi, I'm Cortana.

Google

Siri

Concatenative TTS

How does it work?

- a very large database of short speech fragments are recorded from a single speaker
- speech fragments are recombined to form complete utterances

Limitations: difficult to modify the voice

- switching to a different speaker
- altering the emphasis or emotion

without recording a whole new database

Parametric TTS

How does it work?

- Using a parametric model
- All the information required to generate the speech is stored in the parameters of the model
- The contents and characteristics of the speech can be controlled via the inputs to the model

Limitations: less natural than concatenative TTS

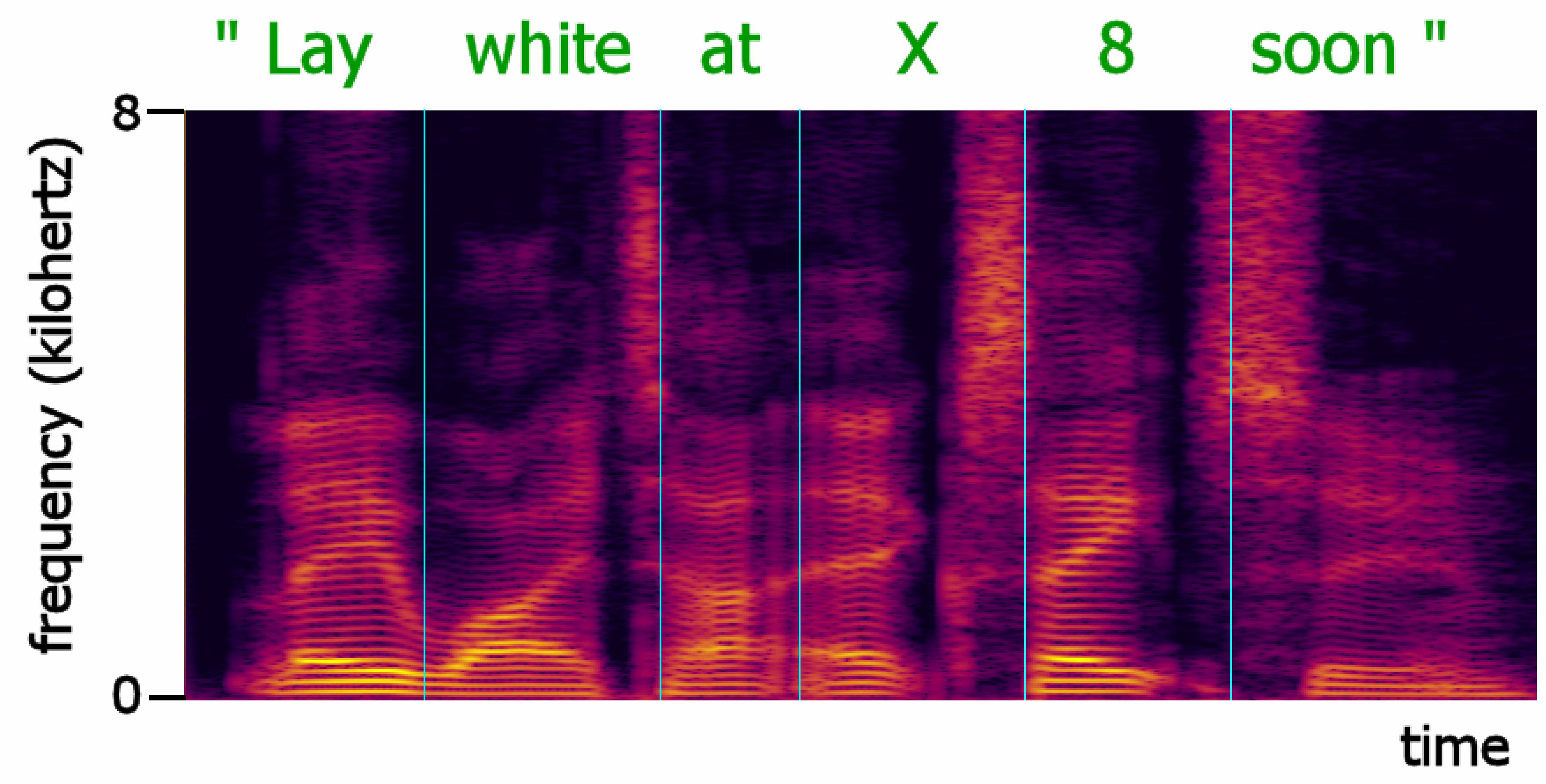
Examples

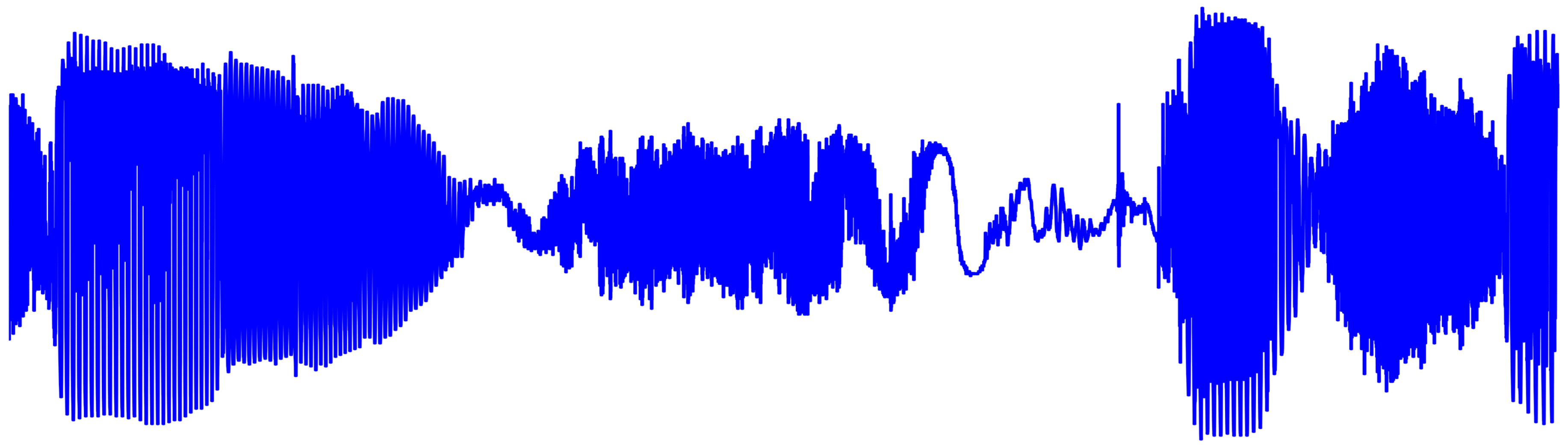
Concatenative	Parametric	Neural

Credit: Xin Wang @ NII

Components of text-to-speech system

- Front end
  - Normalization: Converting non-spoken tokens (numbers, dates, etc) to spoken words, such as "1901" to "nineteen oh one" or "5/12" to "may twelfth".
  - Tagging: Labeling words by their part of speech, pause, stress, emotion, etc.
  - Phoneme conversion: Converting words to a phonetic represe
- Acoustic model
  - Converting the phonemes into a high-level representation of spectrograms, F0, spectral envelope, LSP or LPC coefficients, e
- Vocoder
  - Converting the high-level representation into a final audio waveform.





Overview of current (neural) algorithms

Target	Sub-types	Models
Acoustic modeling: Text ➔ acoustic features	Autoregressive generation	Tacotron, Deep Voice 1/2/3, Transformer TTS, …
	Parallel generation	FastSpeech 1/2, ParaNet, …
Vocoder: Acoustic features ➔ waveform	Non-neural models	Griffin-Lim, WORLD, …
	Neural models	WaveNet, Parallel WaveNet, WaveRNN, WaveGlow, WaveFlow, SampleRNN, LPCNet, MelGAN…
End to end: Text ➔ waveform	Autoregressive generation	Tacotron 2, Char2Wav, ClariNet, …
	Parallel generation	FastSpeech 2S, …

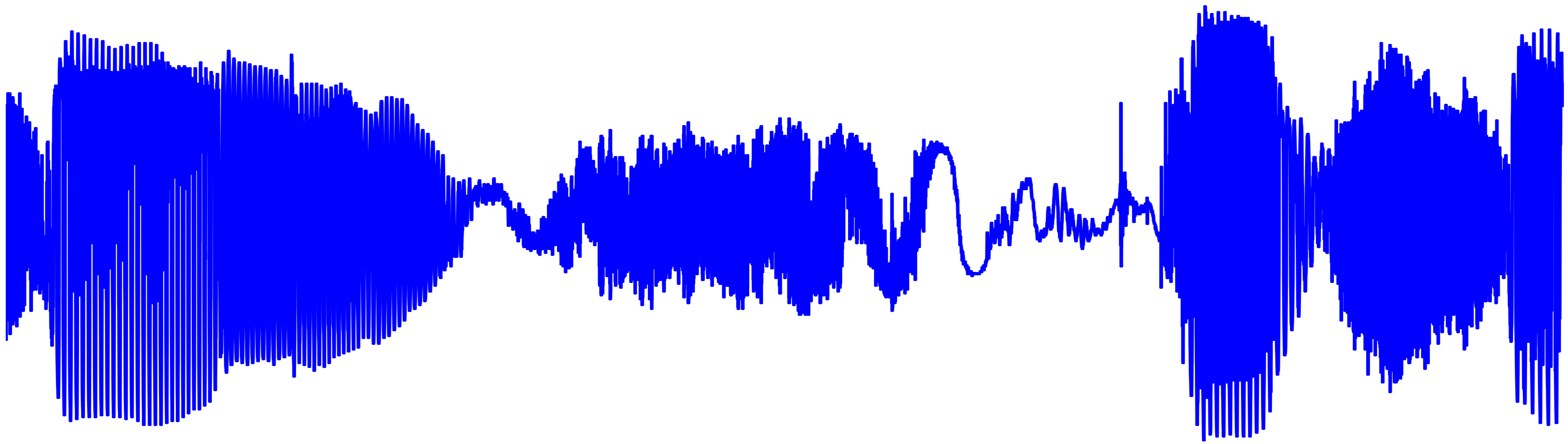
Outline

1. WaveNet: a convolutional vocoder
2. Autoregressive neural acoustic models
   - Deep Voice 3: a convolutional acoustic model
   - Tacotron 2: an LSTM-based acoustic model
   - Transformer TTS: a Transformer-based acoustic model
3. Non-autoregressive neural acoustic models
   1. FastSpeech: a Transformer-based acoustic model
   2. FastSpeech 2/2S: improving FastSpeech
4. Future directions

1. WaveNet: a convolutional vocoder

Google DeepMind, 2016

Autoregressive model



$$p(x) = \Pi_{t=1}^{T} p(x_t | x_1, x_2, \dots, x_{t-1})$$

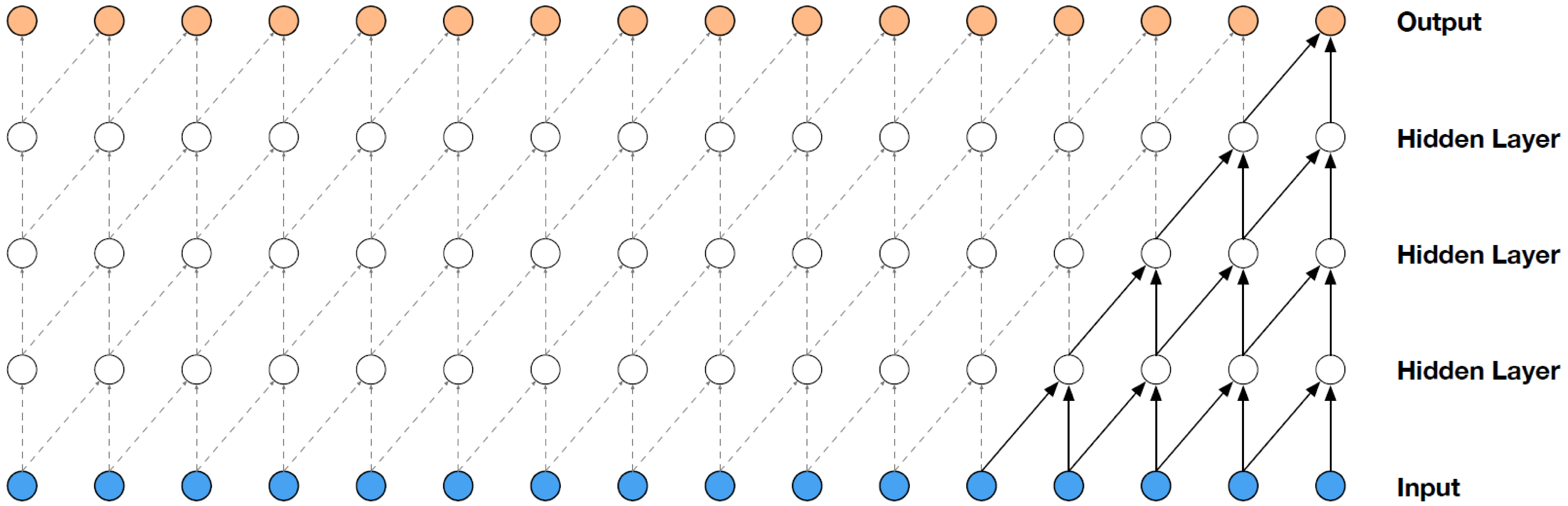
Modeling raw audios

$$p(x_t|x_1, x_2, \dots, x_{t-1})$$

- Raw audio is typically stored as a sequence of 16-bit integer values (one per timestep)
- 65536-class classification is computational costly
- Solution: μ-law transformation + 256 quantization

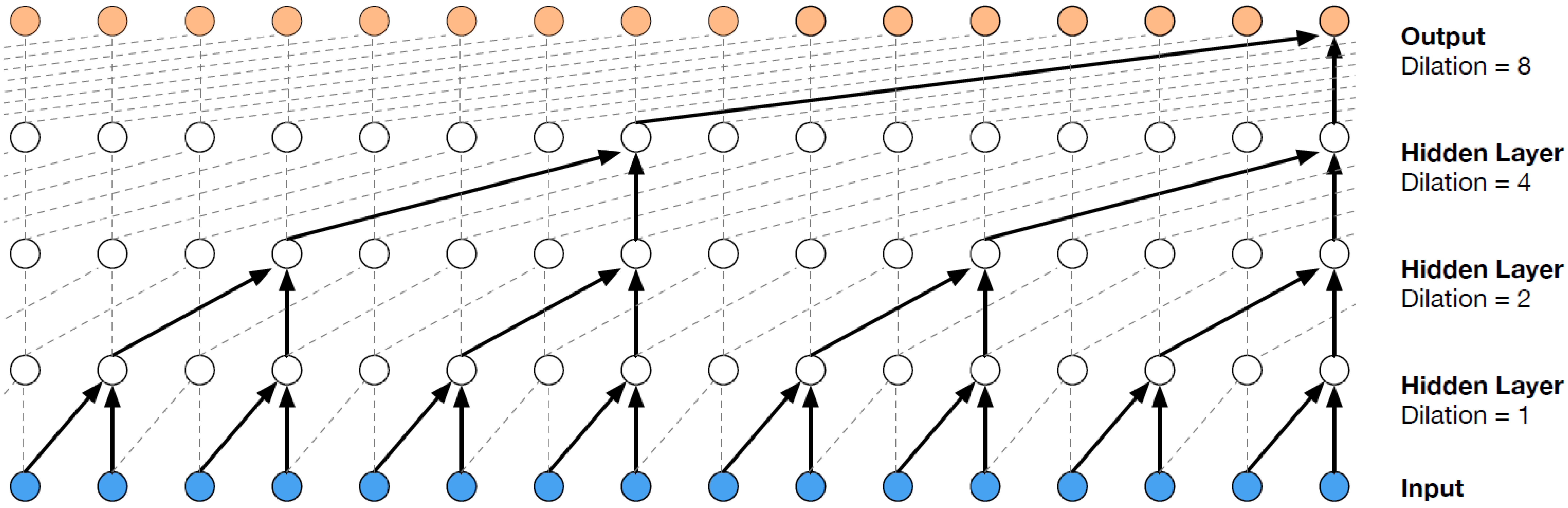
$$f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

Casual convolution



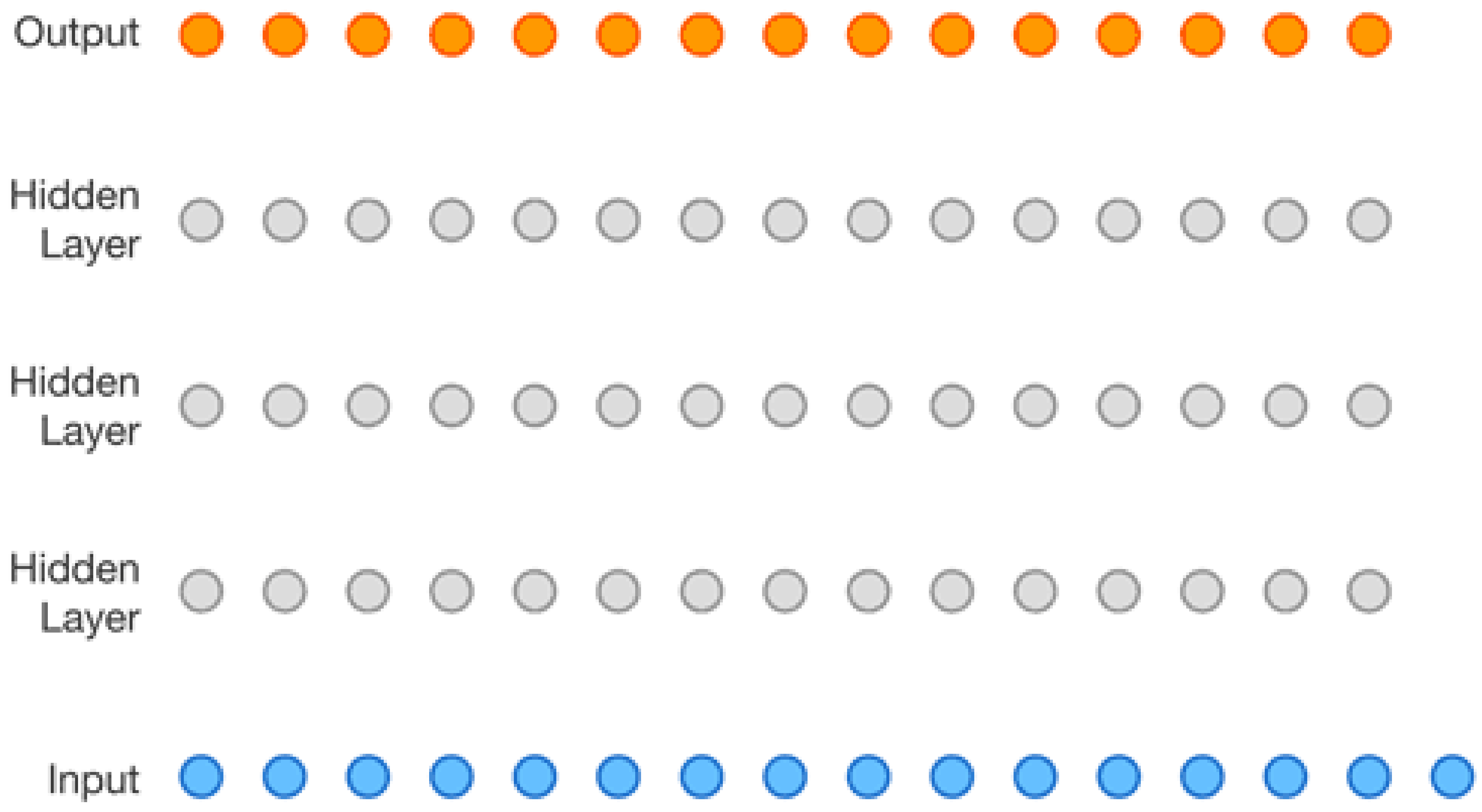
Challenge: cannot model long-term dependency!

Dilated causal convolution



1, 2, 4, … , 512;   1, 2, 4, … , 512;   1, 2, 4, …, 512

WaveNet: inference



WaveNet for Text to Speech

- Input: linguistic features
  - Derived from input texts
  - Linguistic features include phone, syllable, word, phrase, and utterance-level features (e.g. phone identities, syllable stress, the number of syllables in a word, and position of the current syllable in a phrase) with additional frame position and phone duration features
- Input: F0
  - Logarithmic fundamental frequency (log F0)
- Need external models to predict
  - log F0 values
  - phone durations

WaveNet results

- Mean opinion score (MOS)
  - 1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent

	Subjective 5-scale MOS in naturalness	
Speech samples	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	**4.21** ± 0.081	**4.08** ± 0.085
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

2.1. Deep Voice 3: a convolutional acoustic model

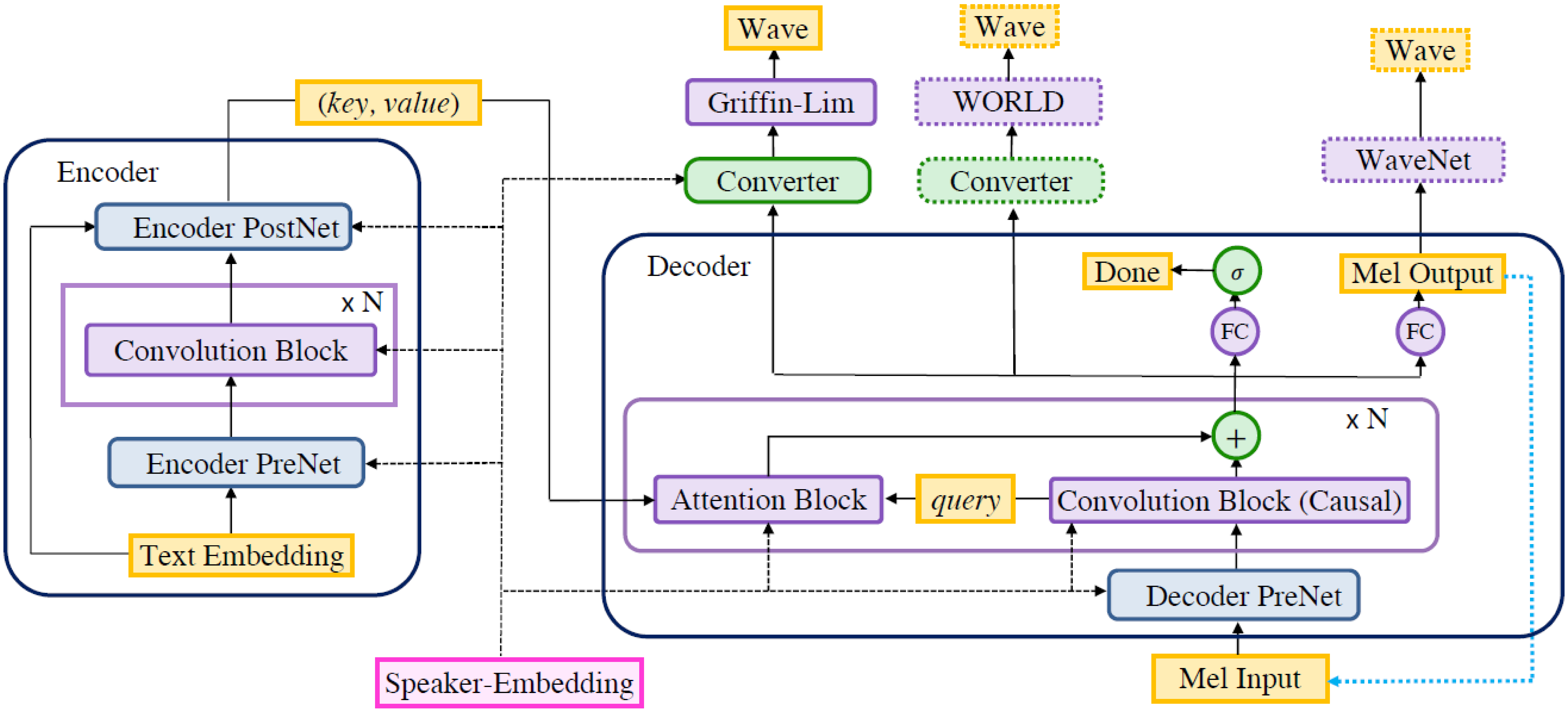
Baidu, ICLR 2018

Deep Voice 3 vs. 1/2

- Deep Voice 1 & 2 retain the traditional structure of TTS pipelines
  - Separating grapheme-to-phoneme conversion, duration and frequency prediction, and waveform synthesis.

- Deep Voice 3 employs a more compact architecture
  - Can converting a variety of textual features (e.g. characters, phonemes, stresses) into a variety of vocoder parameters, e.g. mel spectrograms, linear-scale log magnitude spectrograms, fundamental frequency, spectral envelope, and aperiodicity parameters
  - These vocoder parameters can be used as inputs for audio waveform synthesis models.

Overall architecture



Text preprocessing (front end)

- Uppercase all characters in the input text
- Remove all intermediate punctuation marks
- End every utterance with a period or question mark
- Replace spaces between words with special separator characters which indicate the duration of pauses
  - “Either way, you should shoot very slowly,” ➔ “Either way%you should shoot/very slowly%.”
  - % represents a long pause and / a short pause

Character/phoneme inputs

- Common practice:
  - Use a dictionary maps words to their phonemes, or
  - Directly convert characters (including punctuation and spacing) to acoustic features and learn an implicit grapheme-to-phoneme model
- Deep Voice 3: Mix character-and-phoneme representations
  - Out-of-vocabulary words are input as characters
  - In training, every word is replaced with its phoneme representation with some fixed probability at each training iteration
  - Improves pronunciation accuracy and minimizes attention errors, especially for utterances longer than those seen during training
  - Allow correcting mispronunciations in a phoneme dictionary

Results

Model	Mean Opinion Score (MOS)
Deep Voice 3 (Griffin-Lim)	3.62 ± 0.31
Deep Voice 3 (WORLD)	3.63 ± 0.27
Deep Voice 3 (WaveNet)	3.78 ± 0.30
Tacotron (WaveNet)	3.78 ± 0.34
Deep Voice 2 (WaveNet)	2.74 ± 0.35

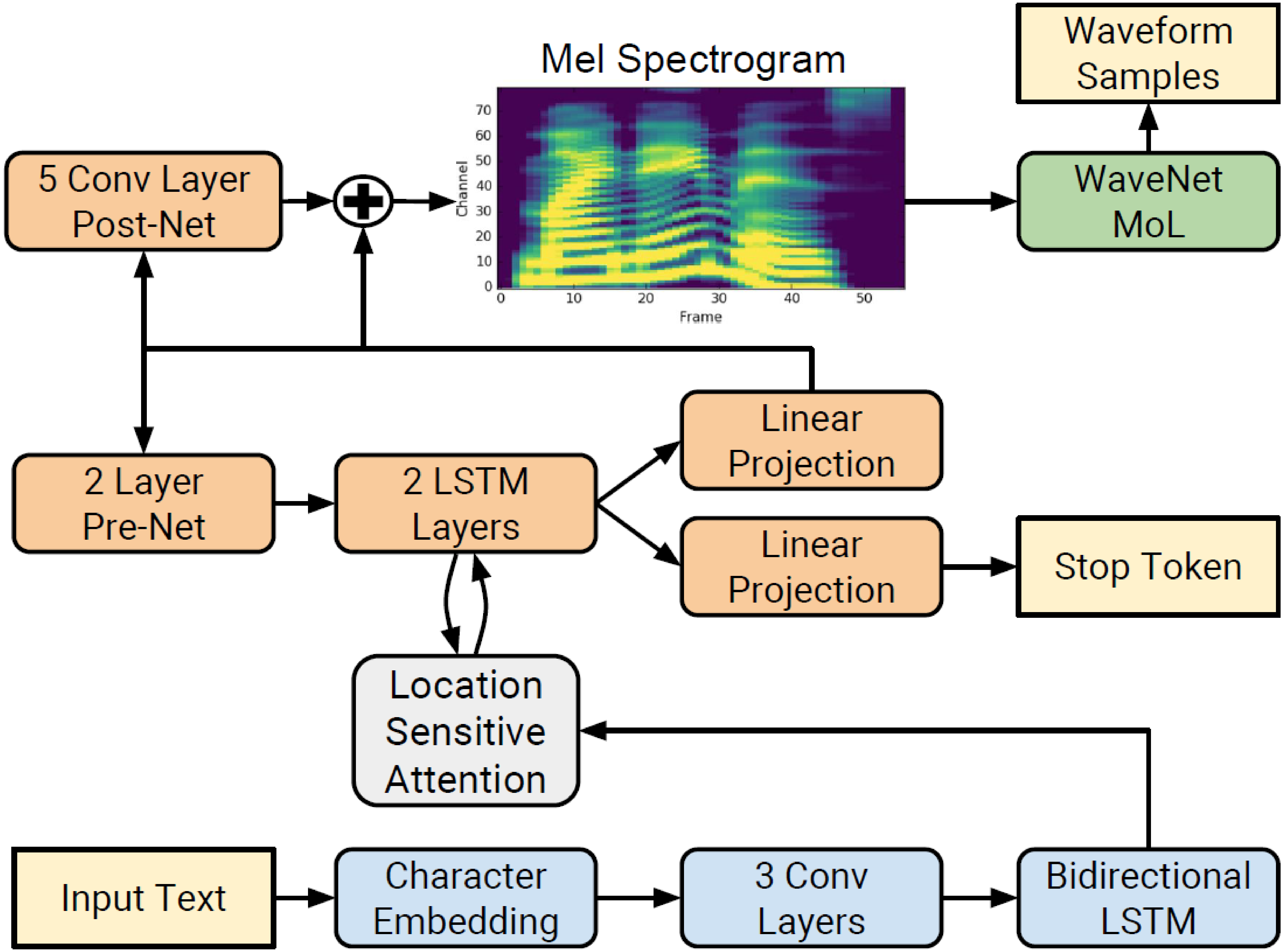
2.2. Tacotron 2: a LSTM-based acoustic model

Google, ICASSP 2018

Tacotron 2 vs. Tacotron

- Tacotron: a LSTM based acoustic model
  - From text to acoustic features, e.g., magnitude spectrograms
  - Rely on a separate vocoder for waveform synthesis
- Tacotron 2: end to end text to speech
  - From text directly to waveform
  - Combine Tacotron-style acoustic model and a modified WaveNet vocoder
  - The acoustic model in 2 is much simpler than Tacotron

Architecture



Results

- Achieves state-of-the-art sound quality close to that of natural human speech

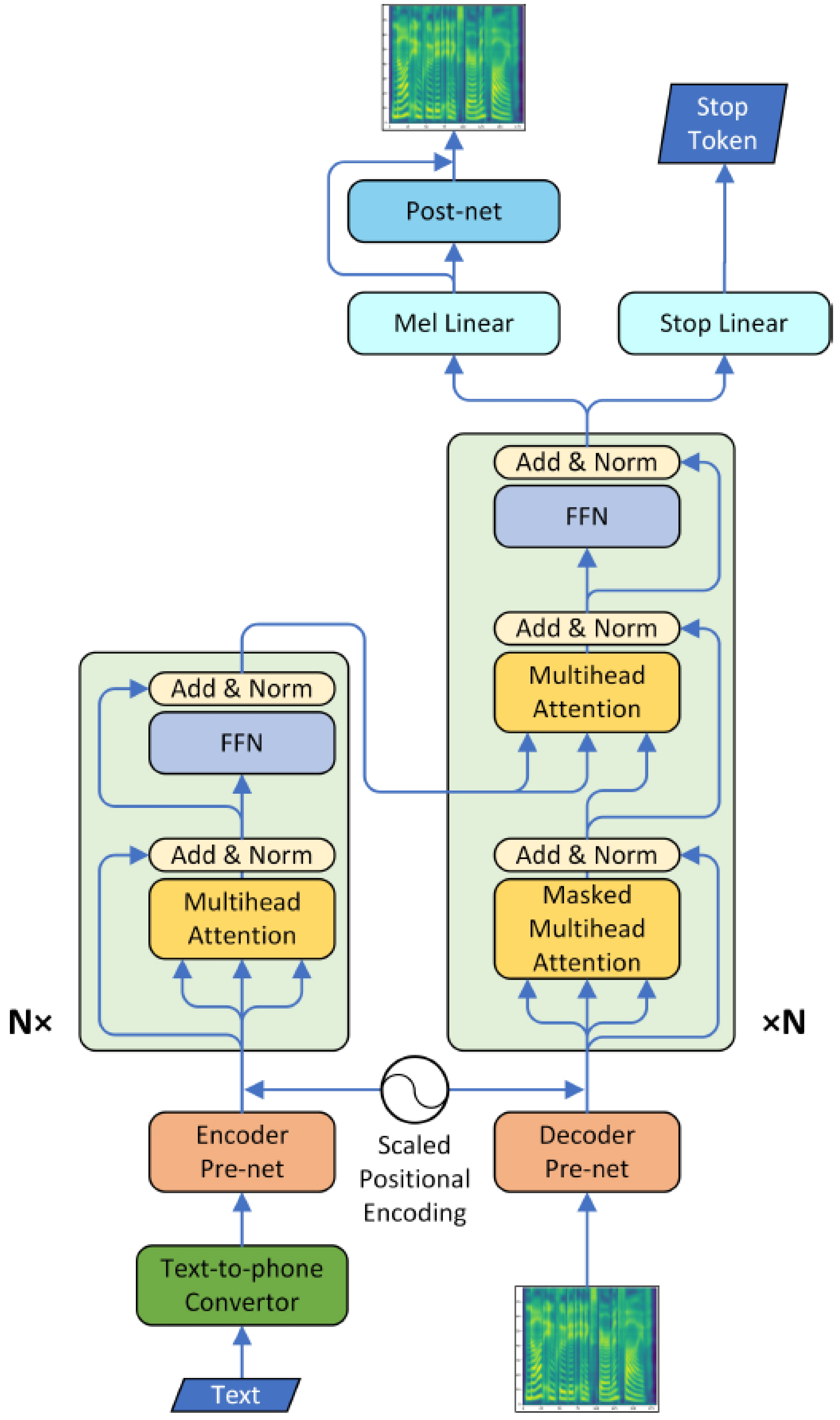
System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

2.2. Transformer TTS

MSRA, AAAI 2019

Architecture

- Follow standard Transformer for machine translation
- Some changes for TTS
  - Rule based text-to-phoneme convertor
  - Scaled positional encoding
  - Encoder and decoder pre-nets



Results

- Training: ~4 times faster than Tacotron 2

System	MOS	CMOS
Tacotron2	4.39 ± 0.05	0
Our Model	4.39 ± 0.05	**0.048**
Ground Truth	4.44 ± 0.05	-

3.1. FastSpeech: Fast, Robust and Controllable Text to Speech

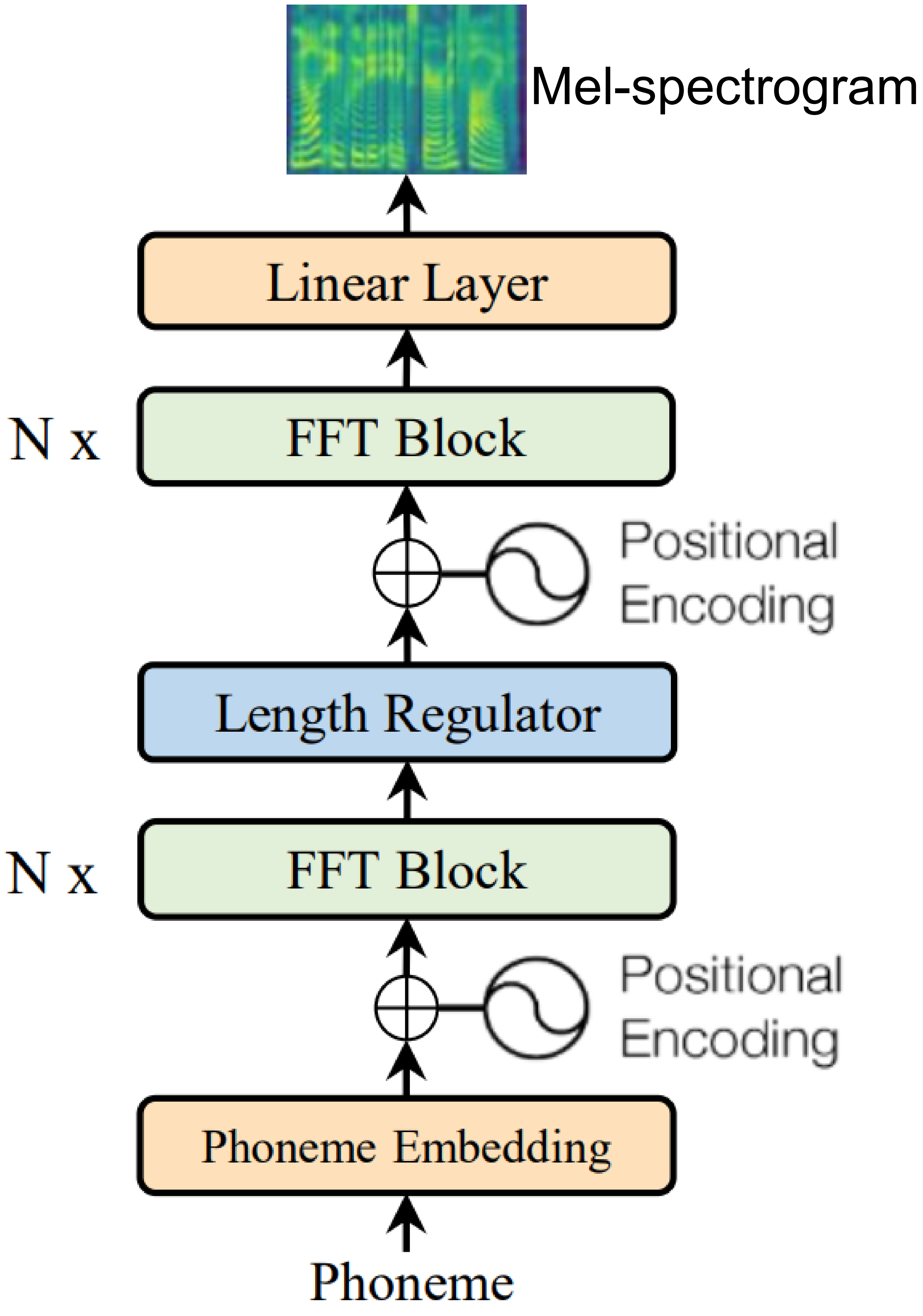
Our work, NeurIPS 2019

Motivation

- Limitations of end-to-end neural TTS
  - **Slow inference speed**: autoregressive mel-spectrogram generation is slow for long sequence;
  - **Not robust**: words skipping and repeating;
  - Lack of controllability

You can call me directly at 4257037344 or my cell 4254447474 or send me a meeting request with all the appropriate information.

FastSpeech architecture

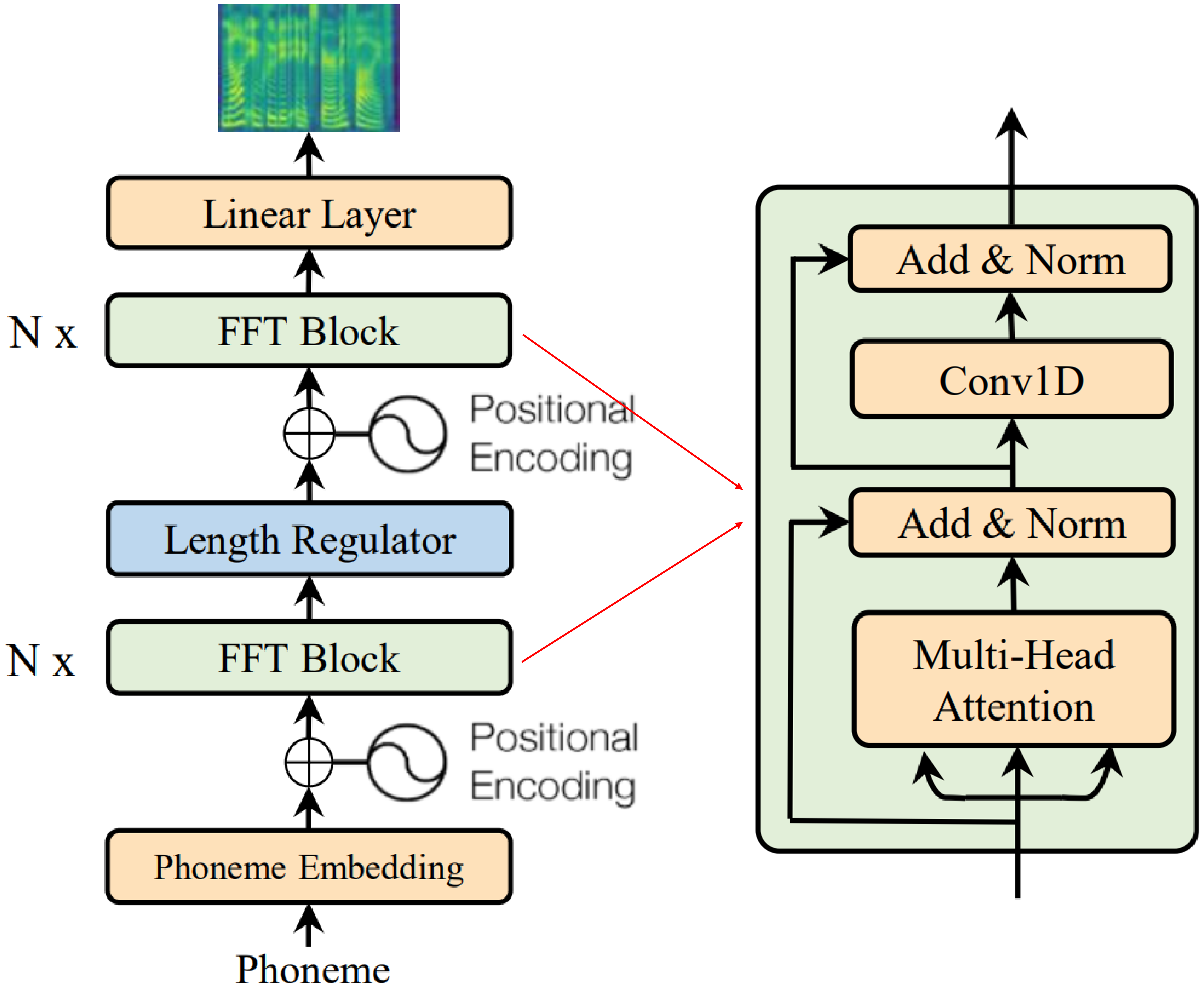


(vocoder)

- Phoneme ----> Mel-spectrogram -----> Voice

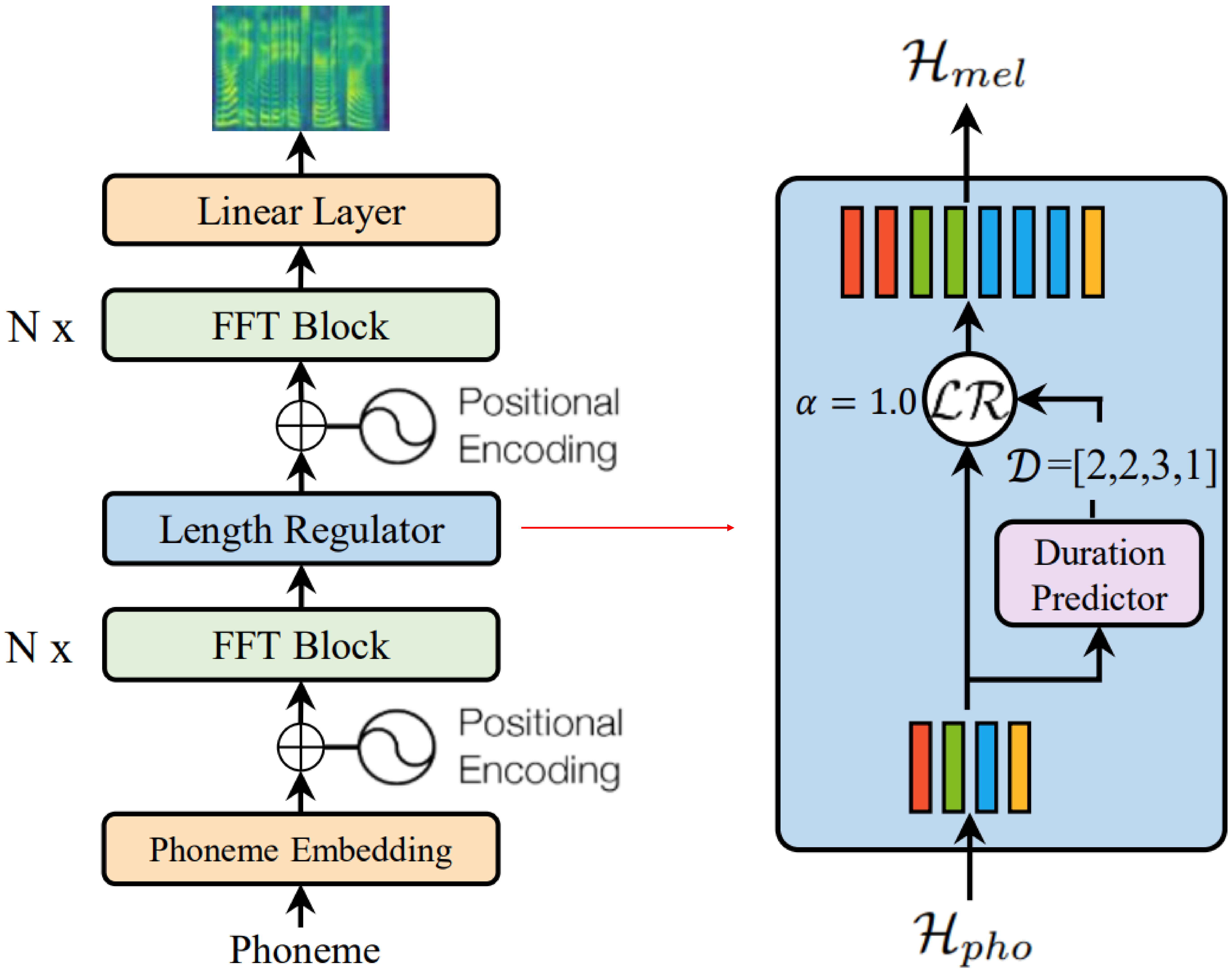
- Feed-forward transformer: generate mel-spectrogram in parallel both in training and inference (speedup)
- Remove the attention mechanism between text and speech (robustness)
- Length Regulator: bridge the length mismatch between phoneme and mel sequence (controllability)

FFT block



- FFT (Feed-Forward Transformer) block: basic block from Transformer, stack N layers.
- Replace dense connection with 1D convolution in speech problem.
- Share the same model structure between the phoneme side and mel side.

Length Regulator



speech

1.0x speed

s p iy ch $\xleftarrow{\alpha=1}$ 2 2 3 1

s s p p iy iy iy ch

α = 0.5 (1 1 1.5 0.5 → 1 1 2 1)

2.0x speed

s p iy ch ← 2 2 3 1

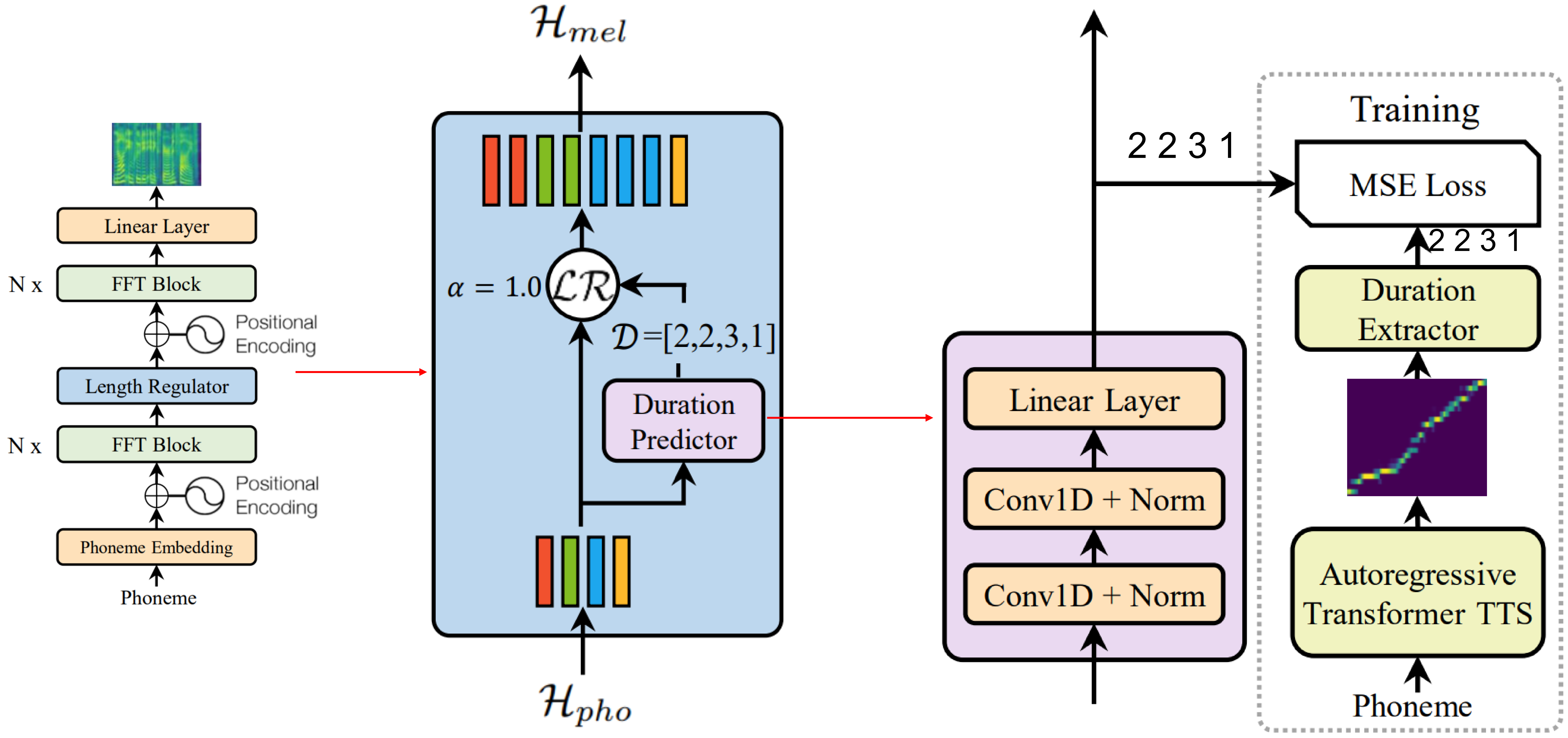
s p iy iy ch

0.5x speed

s p iy ch $\xleftarrow{\alpha=2}$ (4 4 6 2) 2 2 3 1

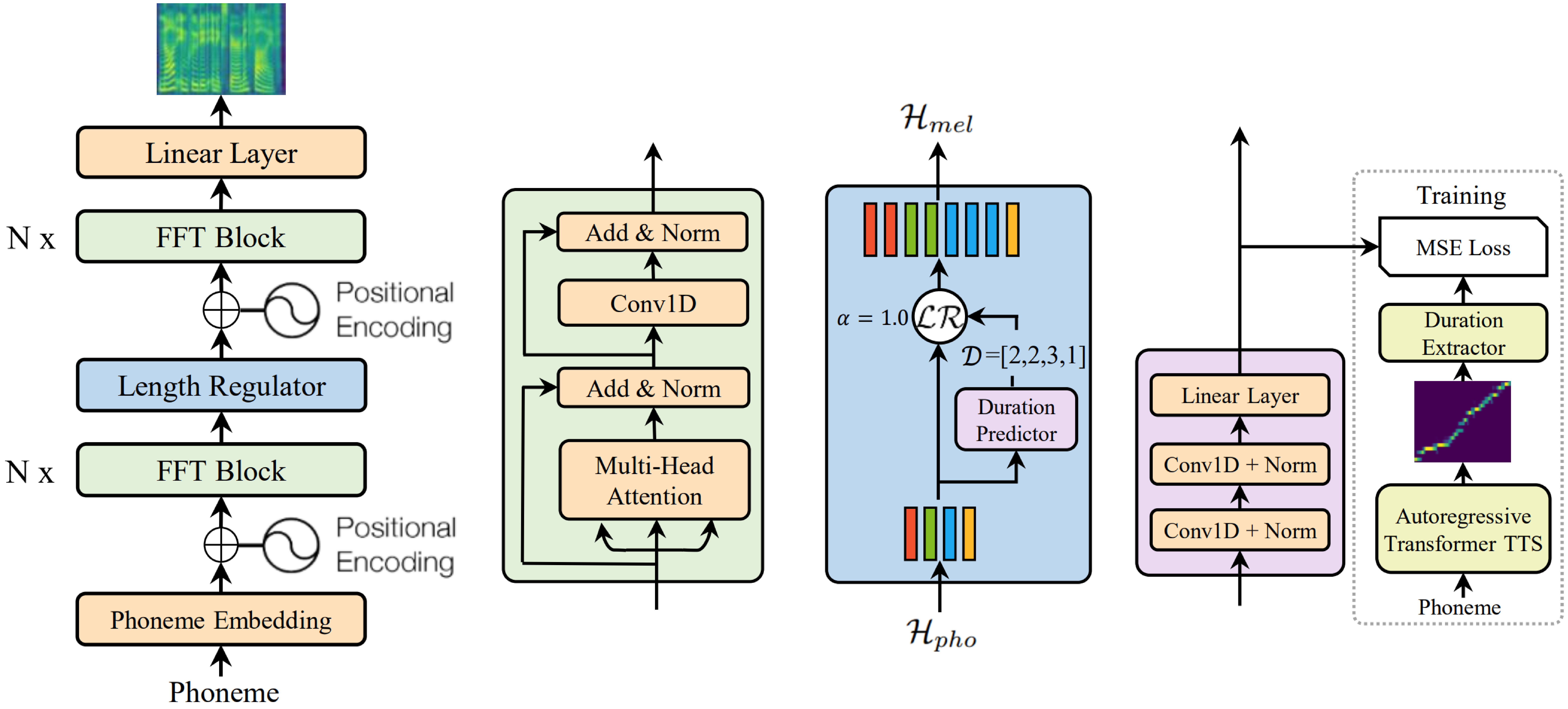
s s s s p p p p iy iy iy iy iy iy ch ch

Duration Predictor



- How to get the label to train the duration predictor?
- Extract duration based on the attention alignments from the autoregressive teacher

Detailed architecture



(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Inference speedup

Microsoft

Method	Latency (s)	Speedup
Transformer TTS [13] (Mel)	6.735 ± 3.969	/
FastSpeech (Mel)	0.025 ± 0.005	$269.40\times$
Transformer TTS [13] (Mel + WaveGlow)	6.895 ± 3.969	/
FastSpeech (Mel + WaveGlow)	0.180 ± 0.078	$38.30\times$

270x speedup for mel-spectrogram generation!

38x speedup for voice synthesis!

Robustness

Method	Repeats	Skips	Error Sentences	Error Rate
Transformer TTS	7	15	17	34%
FastSpeech	0	0	0	0%

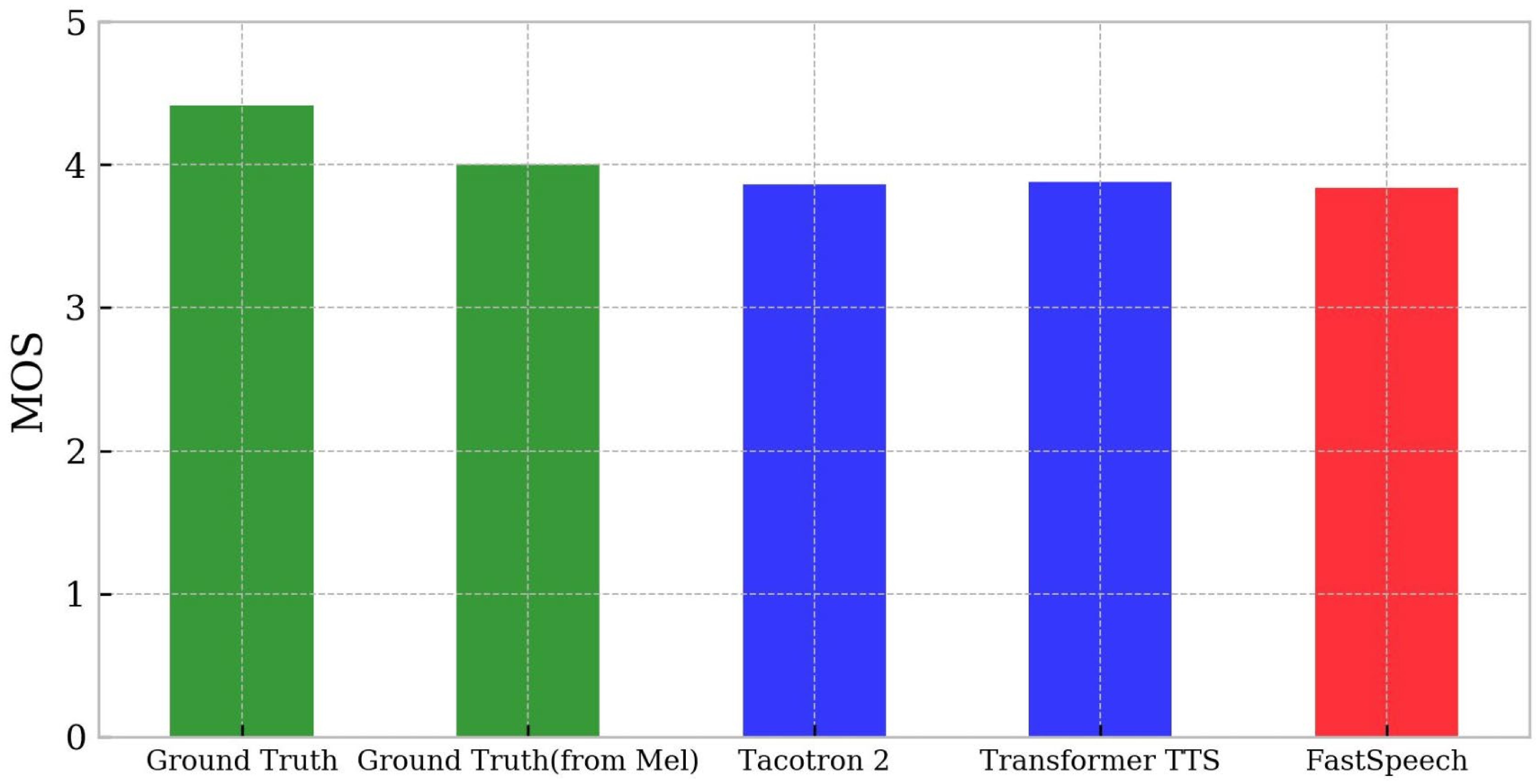
Test on 50 extremely hard sentences provided by TTS team
FastSpeech has no repeating, skipping and error sentences

Transformer TTS FastSpeech

You can call me directly at 4257037344 or my cell 4254447474 or send me a meeting request with all the appropriate information.

Http0XX , Http1XX , Http2XX , Http3XX

Speech synthesis quality



FastSpeech achieves comparable voice quality with Tacotron2 and Transformer TTS, and is close to ground-truth recordings.

https://speechresearch.github.io/fastspeech/

Microsoft

Impact of FastSpeech

- FastSpeech is **extremely fast and high-quality,** with **270x** speedup on mel-spec generation, **38x** speedup on audio generation!
- FastSpeech is widely supported by the community: ESPNet, Baidu, Nvidia, Mozilla

- FastSpeech is the backbone of Azure Speech Service (TTS)
- Supports over 50 languages and locales

https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech

3.2. FastSpeech 2/2S: improving FastSpeech

Our work, under submission

FastSpeech 2 vs. FastSpeech

- The problem in FastSpeech
  - Training pipeline complicated: two-stage teacher-student distillation
  - Target is not good: the target mels distilled from teacher suffer from information loss
  - Duration is not accurate: the duration extracted from teacher is not accurate enough

- Improvements in FastSpeech 2
  - Simplify training pipeline: remove teacher-student distillation
  - Use ground-truth speech as target: avoid information loss
  - Improve duration & Introduce more variance information: ease the one-to-many mapping problem

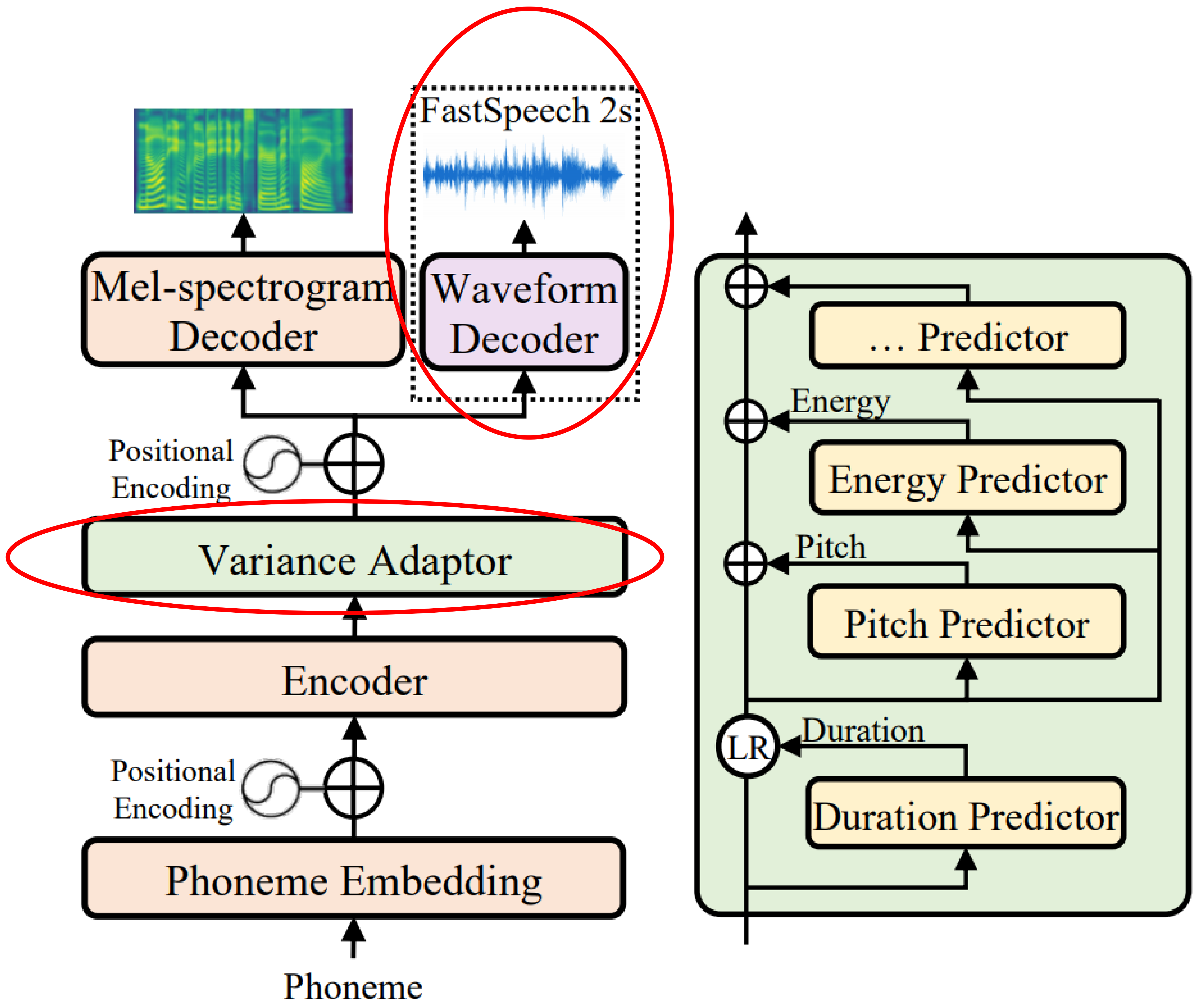
Text

↓

multiple speech variations

(duration, pitch, sound volume, speaker, style, emotion, etc)

FastSpeech 2



(a) FastSpeech 2 (b) Variance adaptor

- Variance adaptor: use variance predictor to predict duration, pitch, energy, etc.
- FastSpeech 2 improves FastSpeech with
  - more simplified training pipeline

    3x training speed up
  - higher voice quality

    0.26 CMOS gain
  - maintain the advantages of **fast, robust and even more controllable** synthesis in FastSpeech
- FastSpeech 2s
  - a fully end-to-end text to wave neural model
  - comparable (high) quality with FastSpeech 2

4. Future directions

Future directions

- Low resource TTS: learning from very limited paired data
  - E.g., 10/20 utterances
- Noisy TTS: learning from noisy speech
  - Previous works need high-quality speech recorded in professional studios
  - Can we train a good model from mobile recorded speech?
- Emotional TTS: synthesize emotional speech

- Singing voice synthesis
- Music composition

Speech related research at my group

1. HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis, arXiv 2020.
2. PopMAG: Pop Music Accompaniment Generation. Multimedia 2020.
3. DualLip: A System for Joint Lip Reading and Generation. Multimedia 2020.
4. FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech. arXiv 2020.
5. XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System, INTERSPEECH 2020.
6. MultiSpeech: Multi-Speaker Text to Speech with Transformer. INTERSPEECH 2020.
7. LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition. KDD 2020.
8. DeepSinger: Singing Voice Synthesis with Data Mined From the Web. KDD 2020.
9. SimulSpeech: End-to-End Simultaneous Speech to Text Translation. ACL 2020.
10. FastSpeech: Fast, Robust and Controllable Text to Speech, NeurIPS 2019.
11. Token-Level Ensemble Distillation for Grapheme-to-Phoneme Conversion, InterSpeech 2019.
12. …

Deep and Reinforcement Learning Group @ MSRA

Machine translation
Healthcare
Speech/music synthesis/recognition
Game playing
Logistics

Generative models
Graph neural networks
Transfer learning
Sequence modeling
Causal learning

Deep learning
Reinforcement learning

We're hiring researchers!

If you are passionate about machine learning research, especially deep learning and reinforcement learning , welcome to join us!!

Contact: taoqin@Microsoft.com
http://research.Microsoft.com/~taoqin

Thanks!

http://research.microsoft.com/~taoqin