The Mobvoi TTS System for BC2019

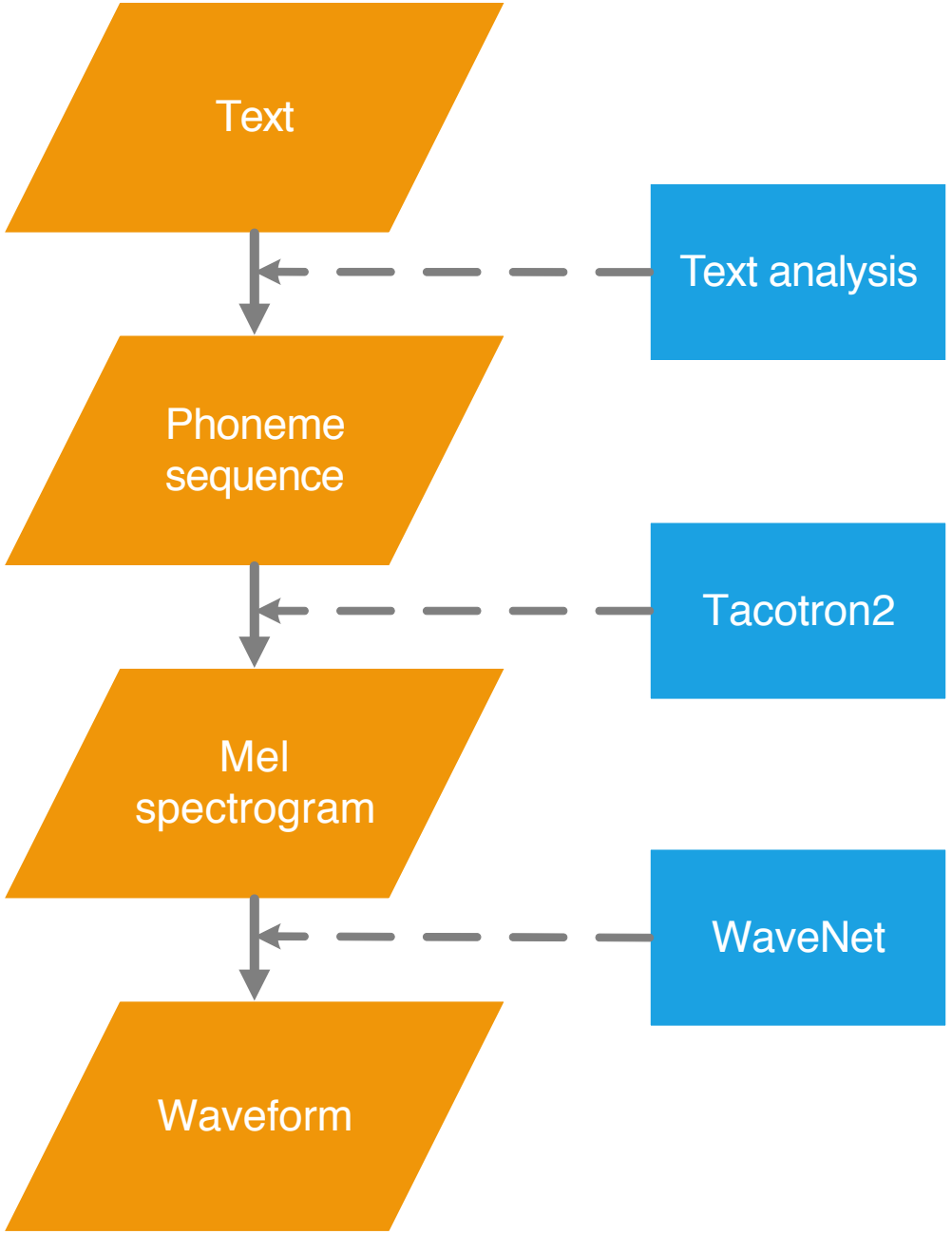
Bing Liu, Yunlin Chen, Hao Yin,
Yongqiang Li, Xin Lei, Lei Xie

Presenter: Shan Yang

Outline

- System
- Data processing
- Front-end
- Back-end: Tacotron2
- Vocoder: WaveNet
- Results

System


Text
Text analysis
Phoneme sequence
Tacotron2
Mel spectrogram
WaveNet
Waveform

Data processing

- Source data: 480 audio files, 48 kHz, 8 hours, MP3 format
- Data processing:
  - Convert to WAV format
  - downsample: 48 kHz -> 16 kHz
  - Segmentation: <=10s per segment
  - Control silence at the beginning and end
  - Energy-based normalization
  - Clean-up text
- Final data: 4187 audio files, 16k Hz, WAV format

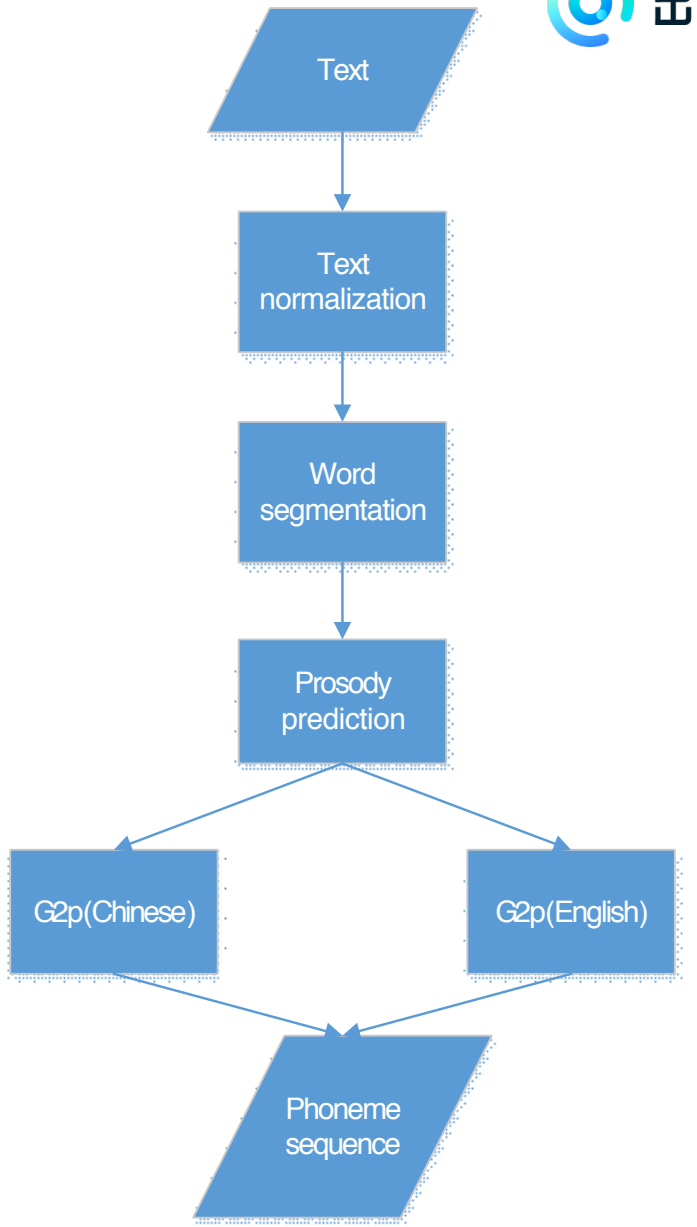
Front-end

Text：

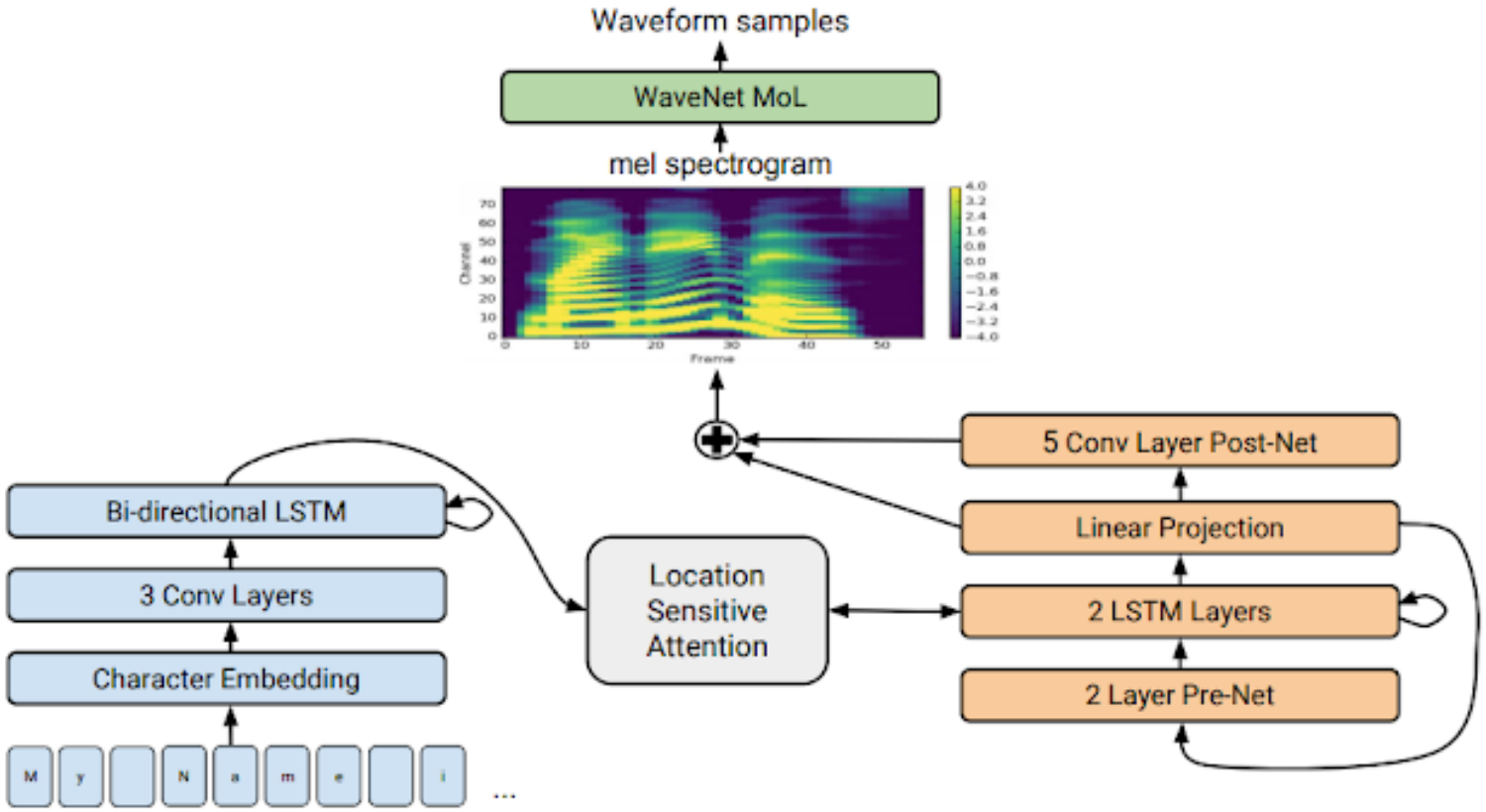
今晚八点,在得到App的直播间,吴军老师会做一场直播。

Phoneme sequence:

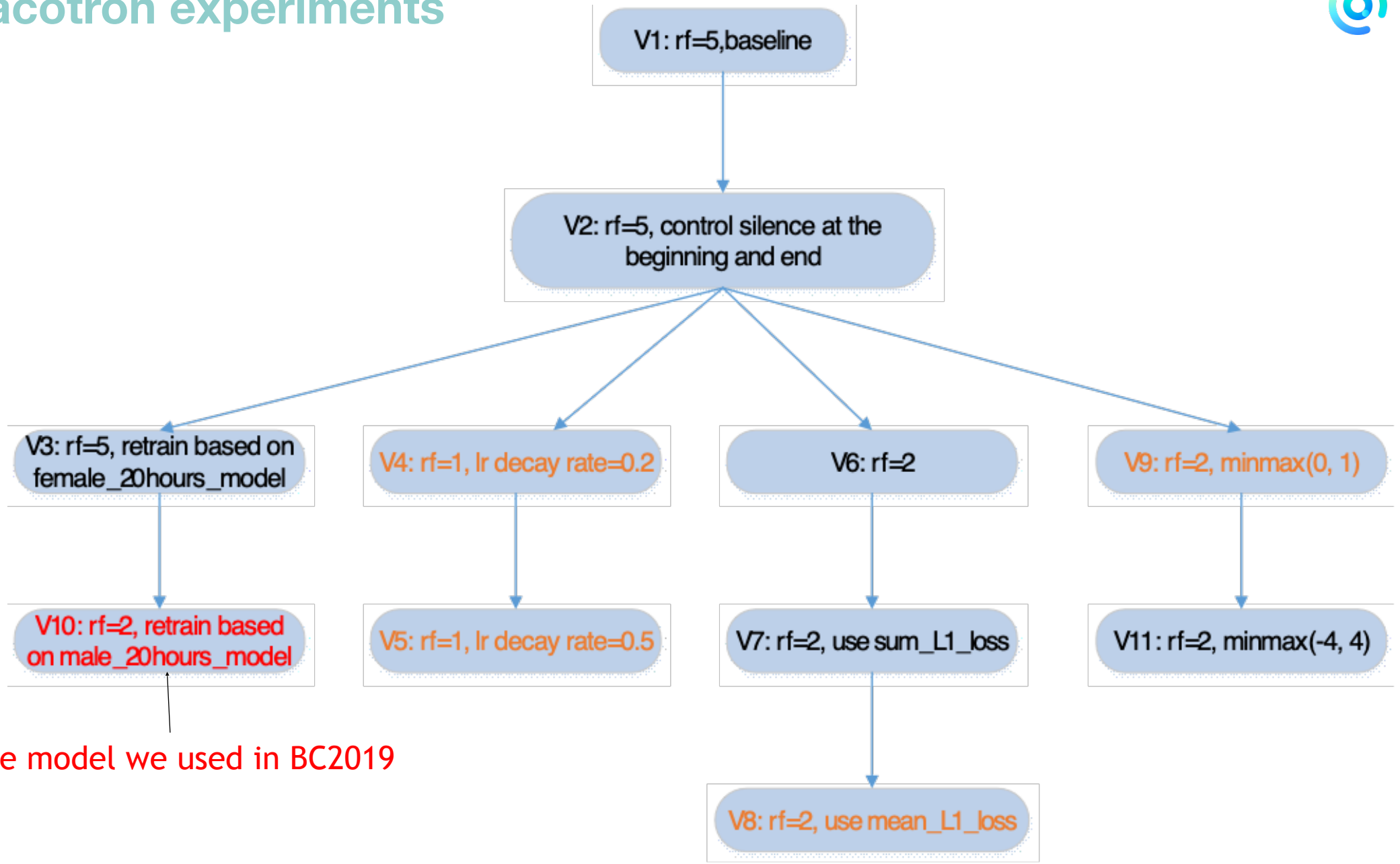
j in 1 w an 3 #1 b a 1 d ian 3 #3 z ai 4 #1 d e 2 d ao 4 #1 / EY 1 . P IY 1 . P IY 1 / d e 5 #1 zh i 2 b o 1 j ian 1 #3 w u 2 j vn 1 #1 l ao 3 sh i 1 #2 h ui 4 z uo 4 #1 y i 1 ch ang 3 #1 zh I 2 b o 1 SIL



Back-end: Tacotron2



Tacotron experiments



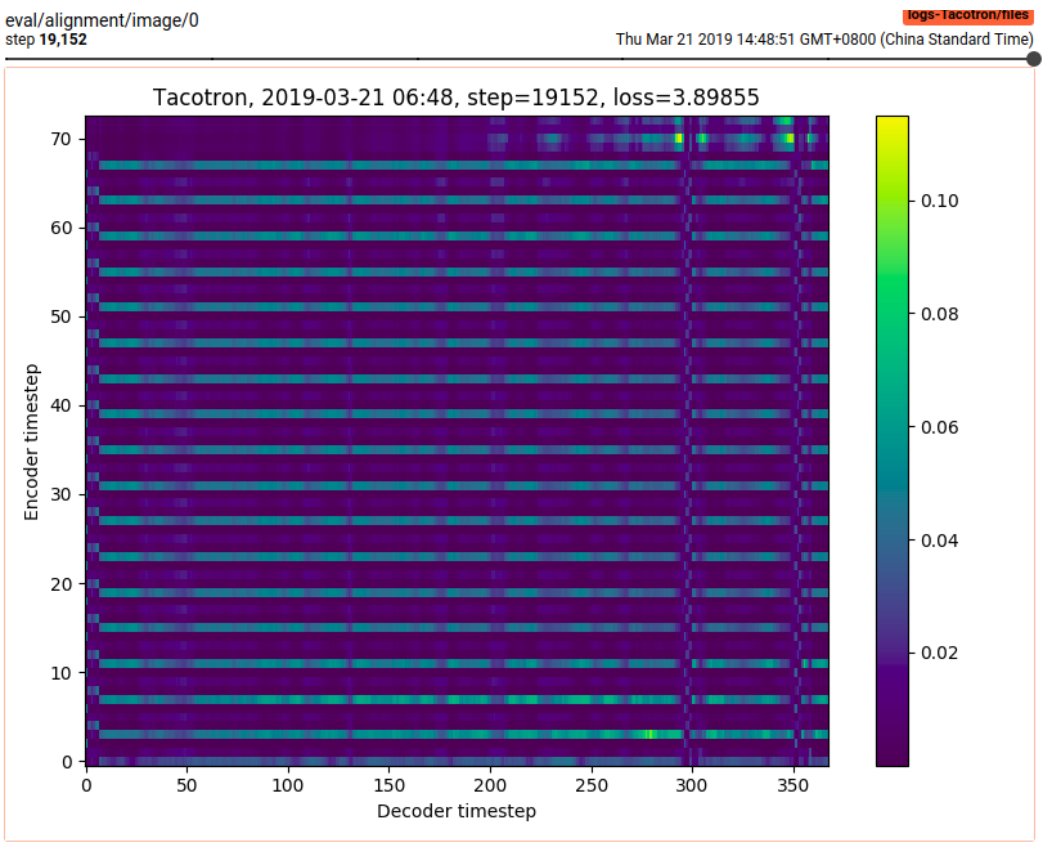
The model we used in BC2019

Tacotron experiments

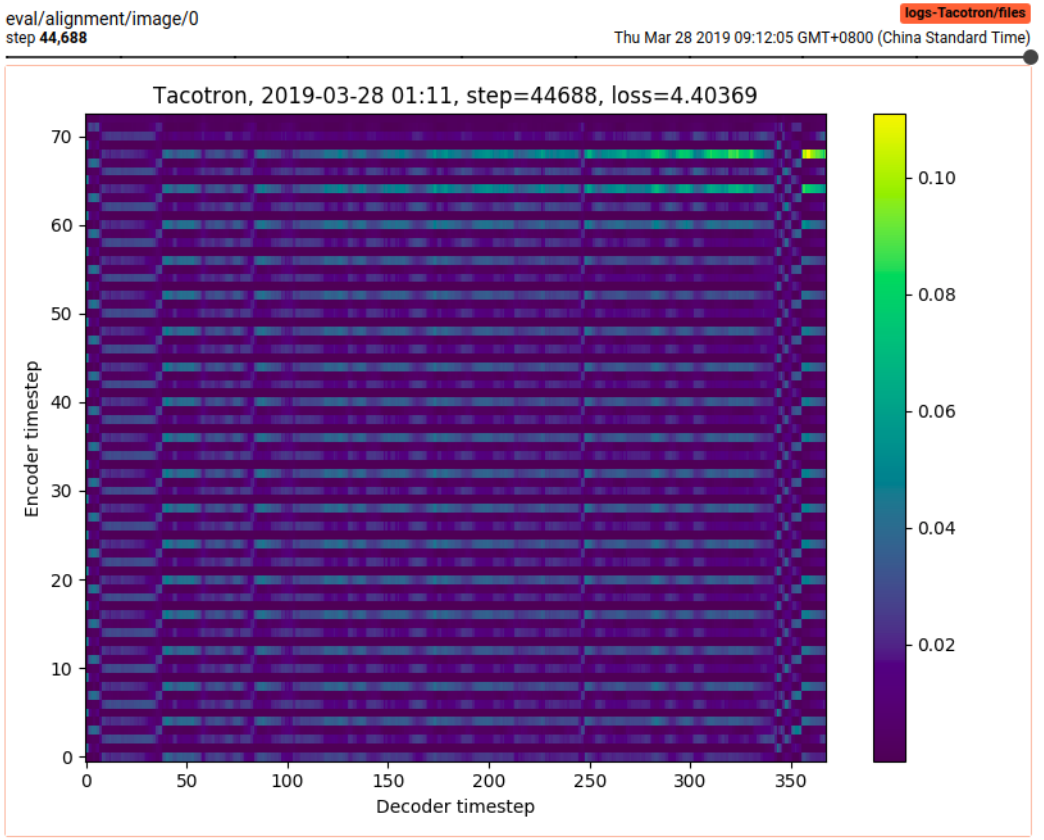
control silence at the beginning and end	V1:V2	To control silence at the beginning and end is better
Adaptation	V2:V3, V6:V10	Adaptation is better, great improvement
Reduce factor (rf)	V2(rf=5):V6(rf=2)	rf=2 is better
	V4, V5 (rf=1)	No alignment
L1 loss	V6:V7(sum L1 loss)	So close
	V8(mean L1 loss)	No alignment
Minmax normalization	V6:V11(minmax(-4,4))	Minmax is better
	V9(minmax(0,1))	No alignment

Tacotron experiments

- No alignment



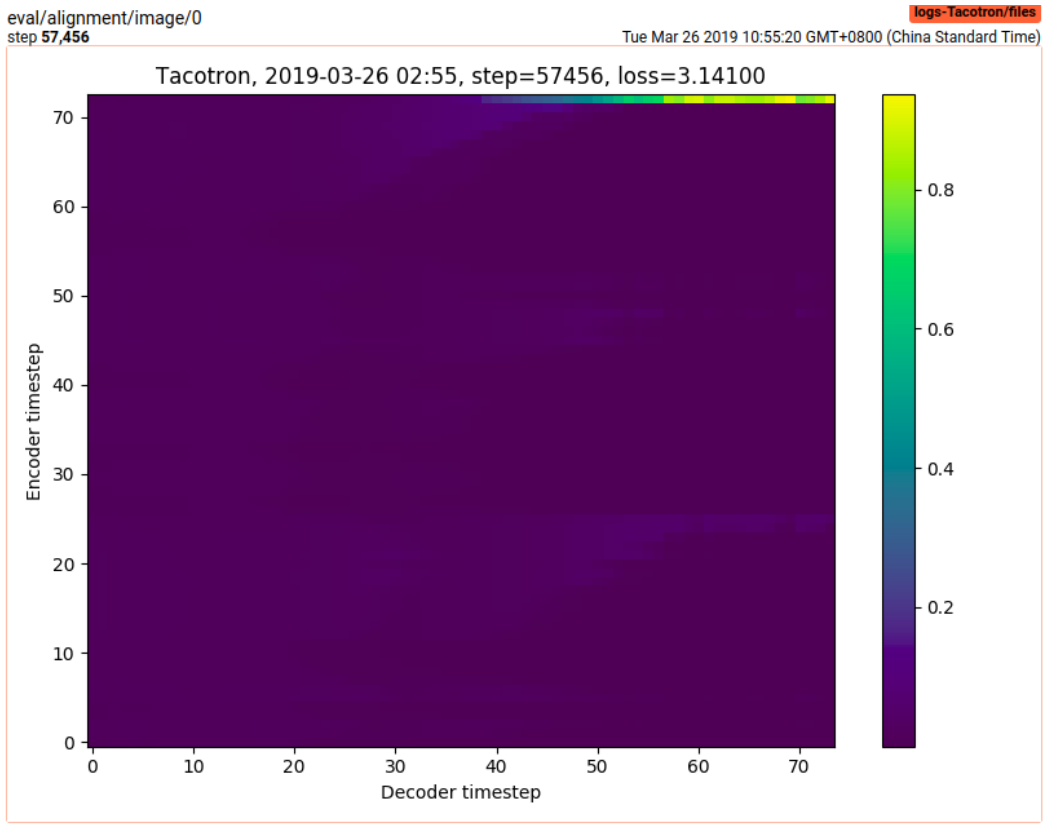
V4: rf=1,
lr_decay_rate=0.2



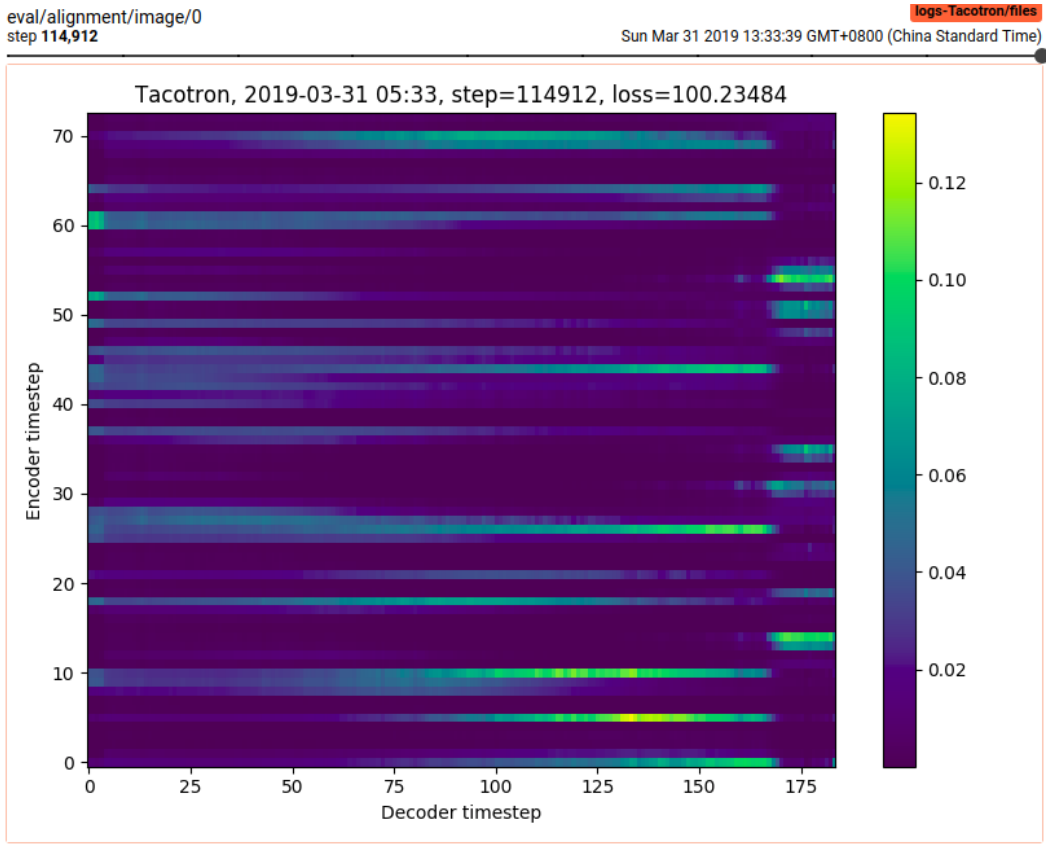
V5: rf=1,
lr_decay_rate=0.5

Tacotron experiments

- No alignment



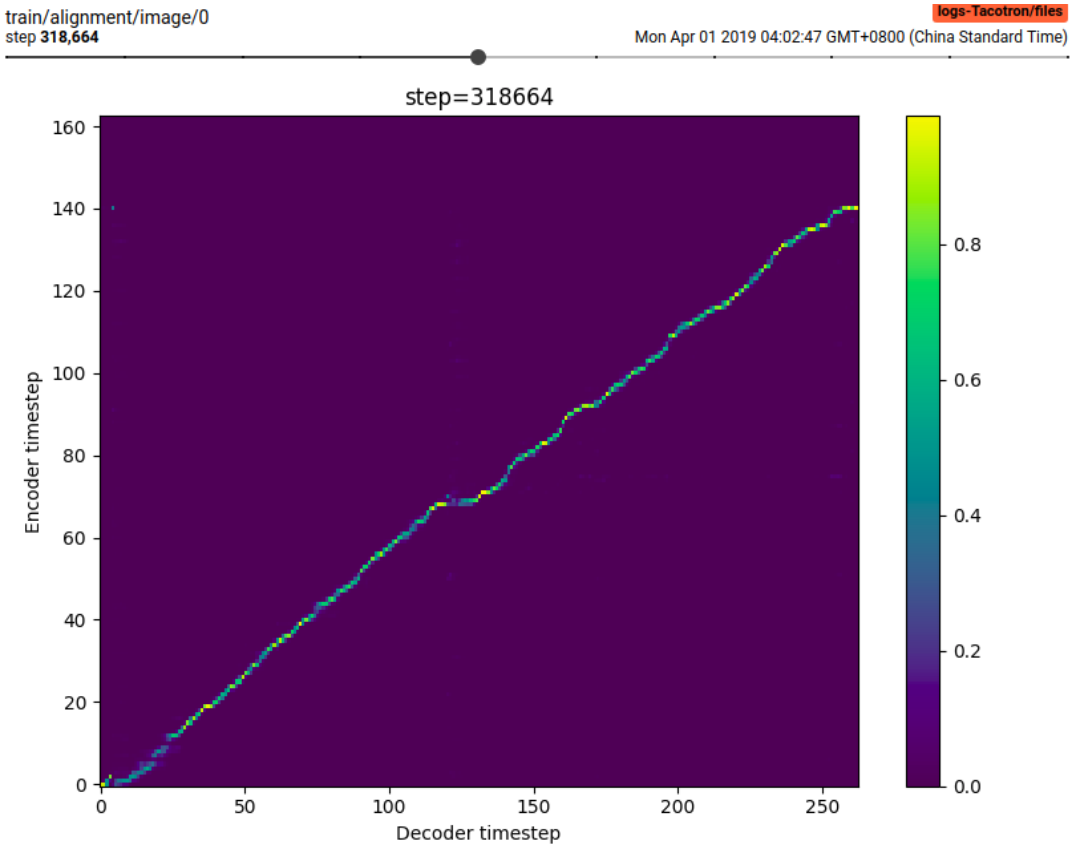
V8: rf=2, use mean L1 loss



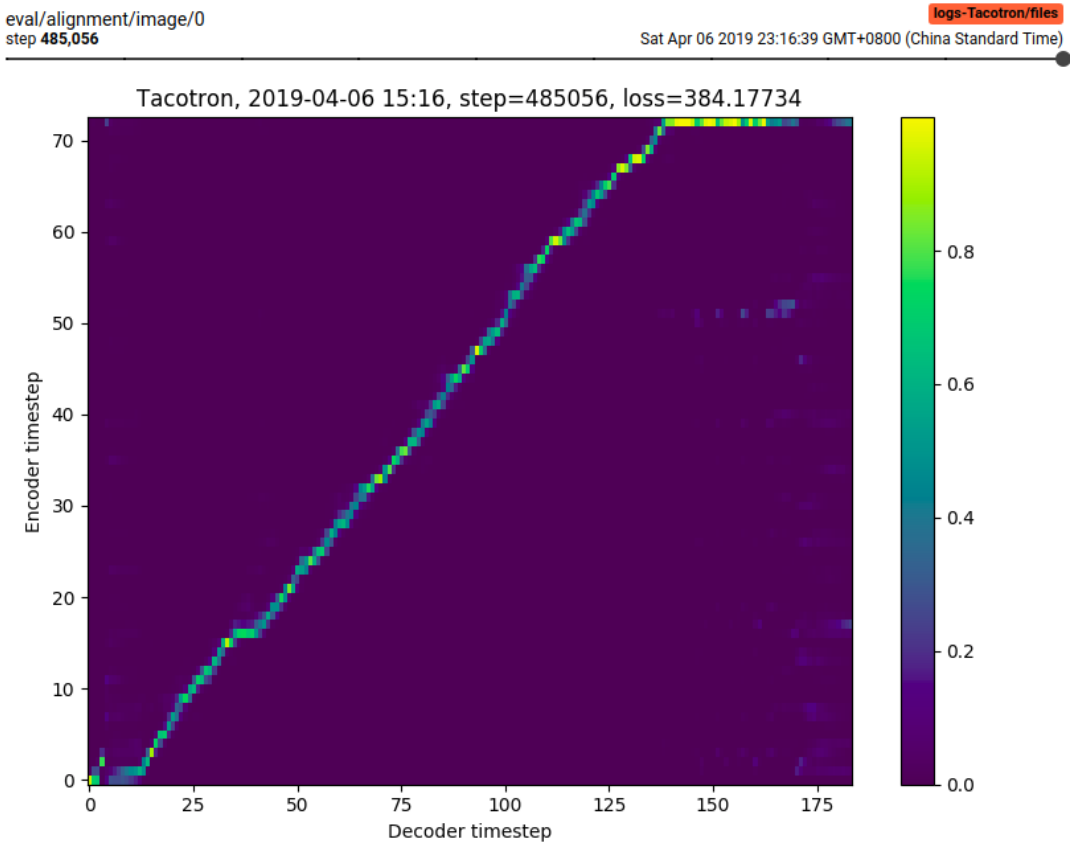
V9: rf=2, minmax(0, 1)

Tacotron experiments

- Normal alignment

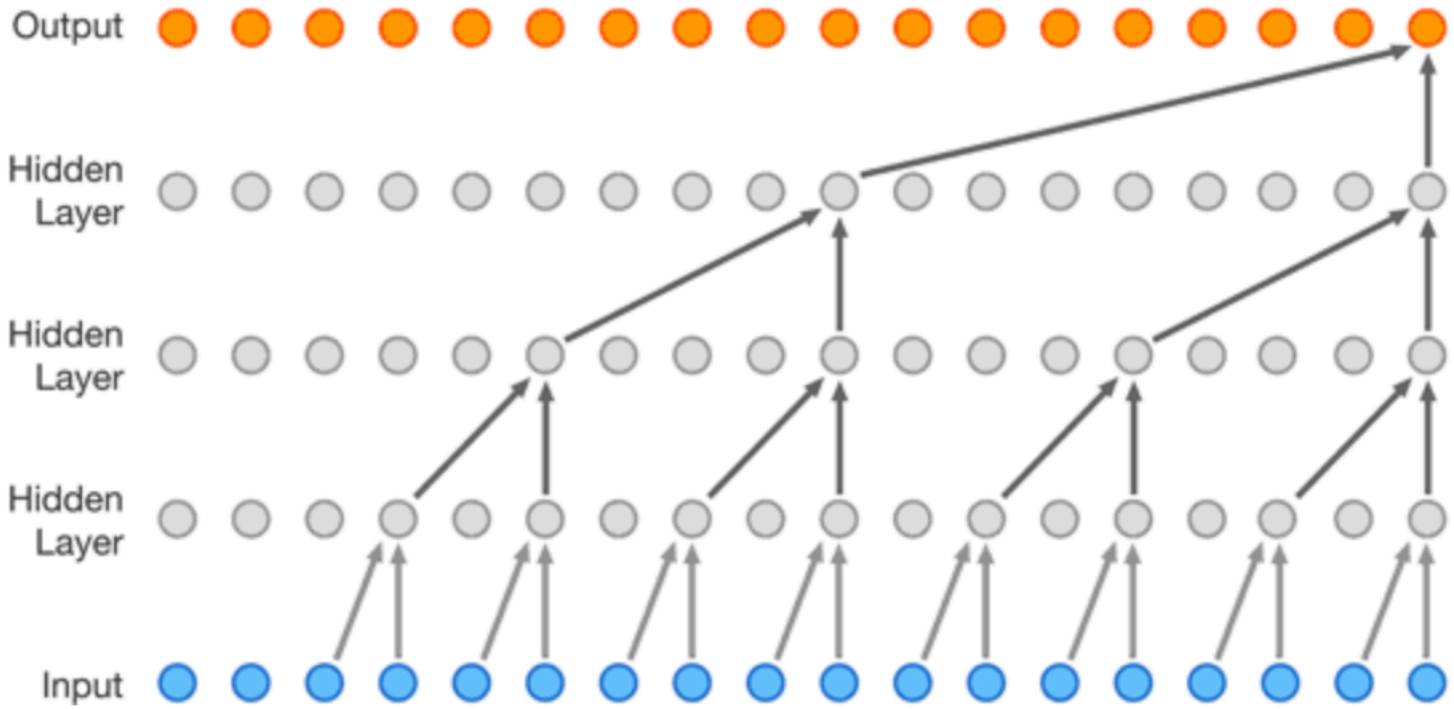


V10: rf=2, retrain, **training** set



V10: rf=2, retrain, **valid** set

Vocoder: Wavenet



1. 20 dilated convolution layers, grouped into 2 dilation cycles
2. Predict 10-component mixture of logistic distributions (MoL)

WaveNet experiments

- WaveNet experiments:
  - **V1**: Baseline
    - Conditioning in ground truth mel spectrogram
    - Stop_step: 1,000,000
  - **V2**: GTA(taco v10) model:
    - Conditioning in mel spectrogram generated from tacotron v10 GTA inference
    - Stop_step: 2,000,000, retrain based on v1

WaveNet experiments

- Ground Truth Alignment(GTA) training

Training	Synthesis	
	Predicted	Ground truth
Predicted	4.526 ± 0.066	4.449 ± 0.060
Ground truth	4.362 ± 0.066	4.522 ± 0.055

Table 2. Comparison of evaluated MOS for our system when WaveNet trained on predicted/ground truth mel spectrograms are made to synthesize from predicted/ground truth mel spectrograms.

Results

Original Wave	
WaveNet_1M(Ground Truth feature)	
WaveNet_1M(taco_v10)	

WaveNet_retrain_2M (taco_v10)	Random test	
	BC2019 test	

Results

出门问问

- To improve
  - **Rhotic accent**: some are somewhat blurred.
  - **Interrogative sentences**: Insufficient tone.
  - **Rhythm**: some sentences are read faster, and some sentences are unreasonably paused (too long or too short).
  - **English**: english adaptive training with less data.
  - **Digital**: missing or blurring of consecutive identical numbers.
  - **Missing**: miss word at the end of the sentence.
  - **Ancient poetry**: the rhythm of ancient poetry needs to be improved.

出门问问

Thanks