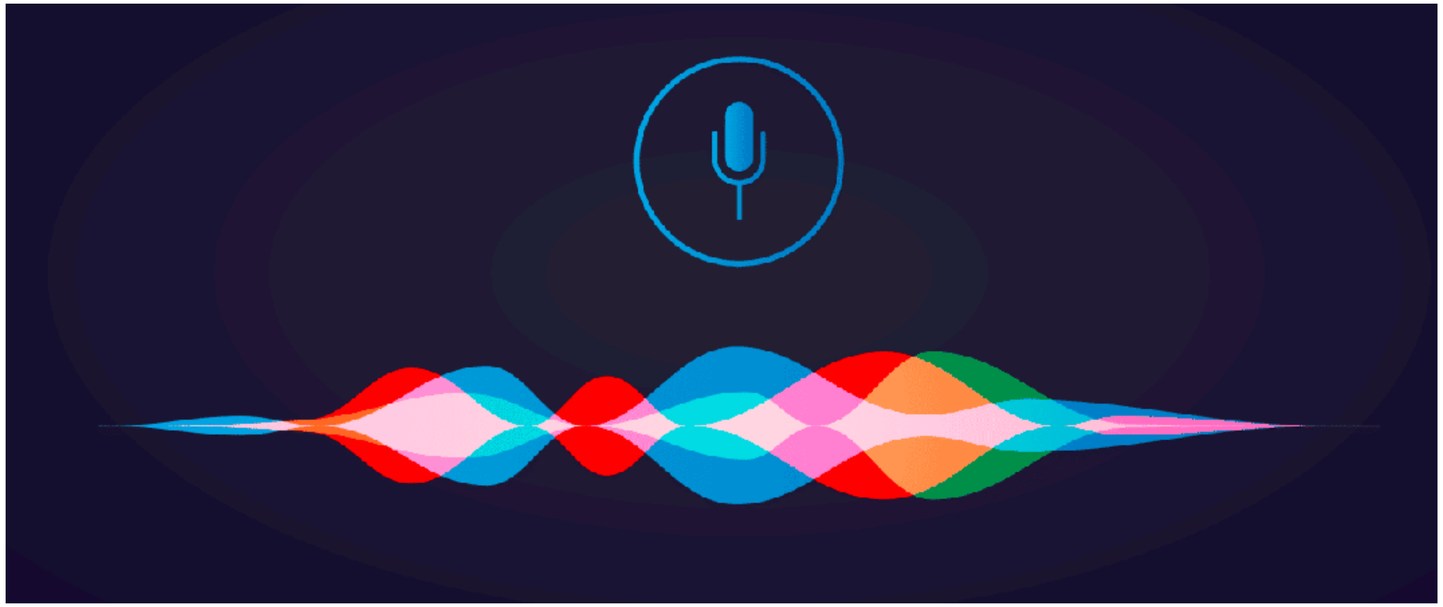
Lecture 18: Text-to-Speech Synthesis

Zhizheng Wu

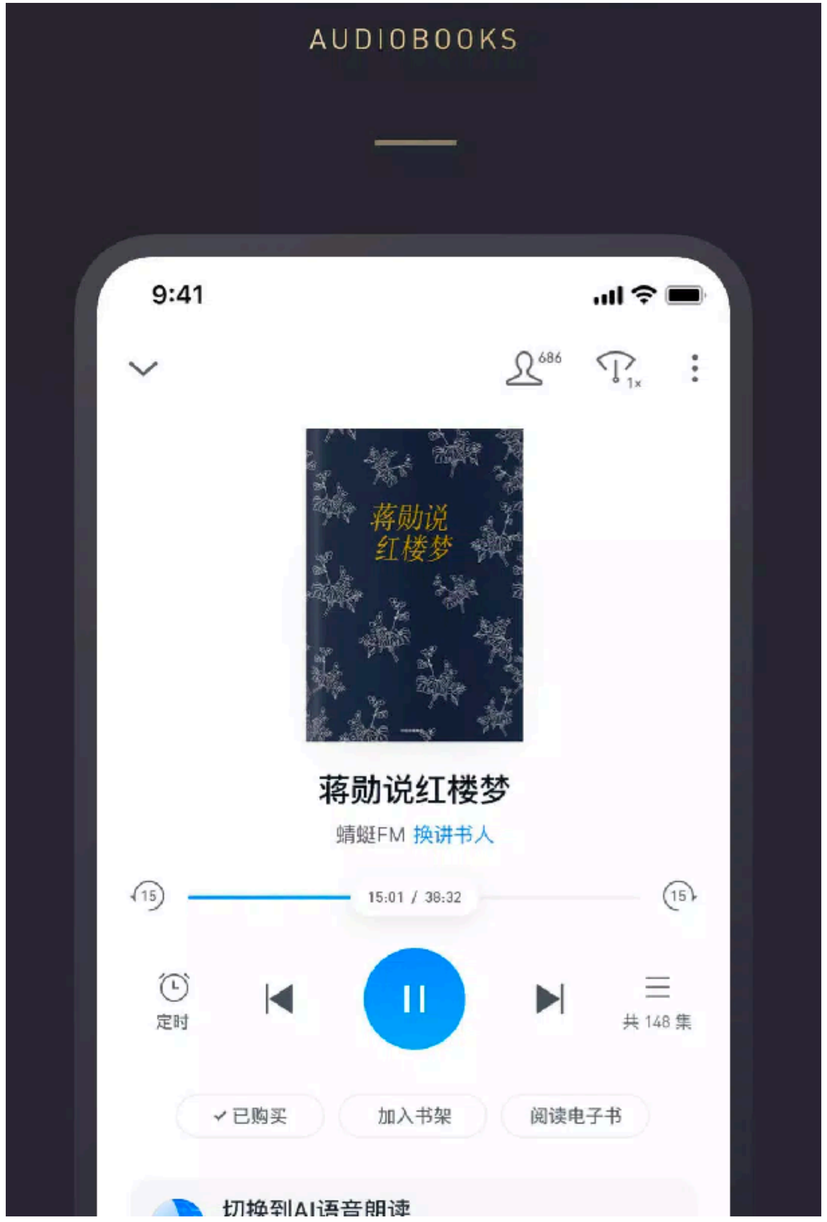
Agenda

- Applications
- Overview of text to speech
- Frontend
- Acoustic model
- Waveform generator
- Tools and readings

Applications



Applications

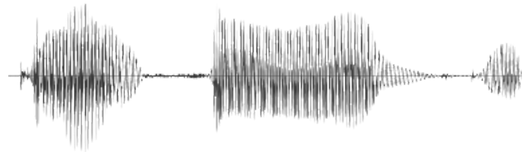


Applications

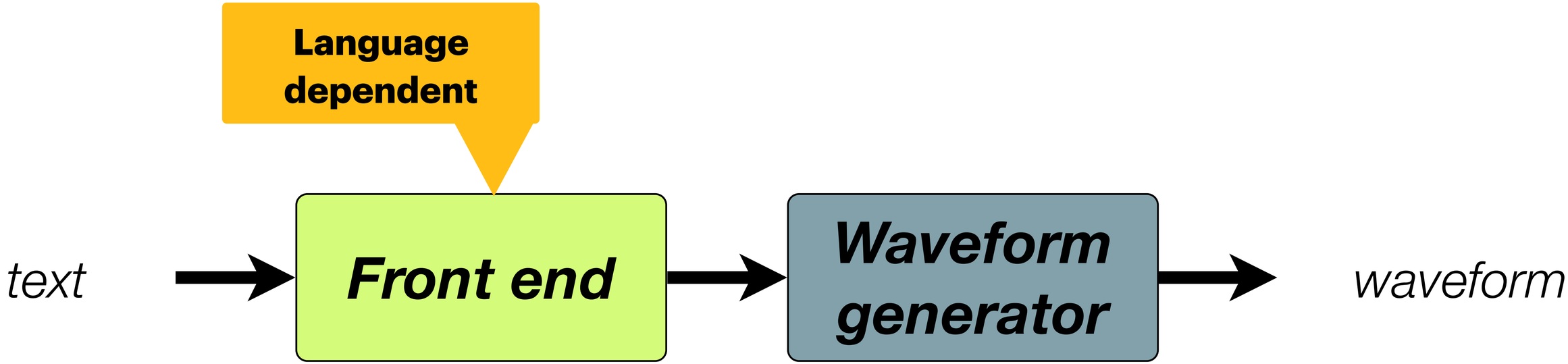
The end-to-end problem we want to solve

text → ***Text-to-Speech*** → *waveform*

`Author of the…`

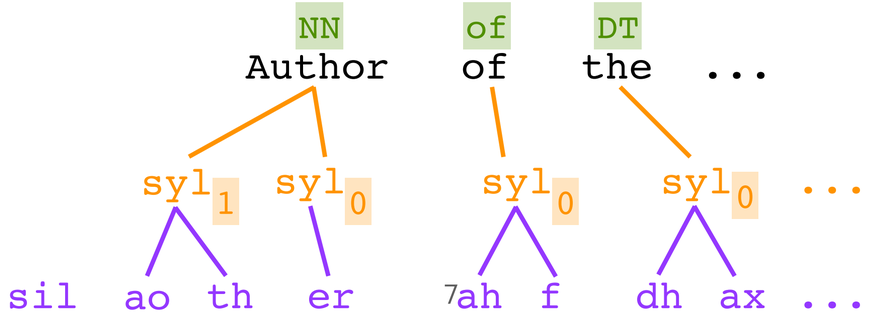


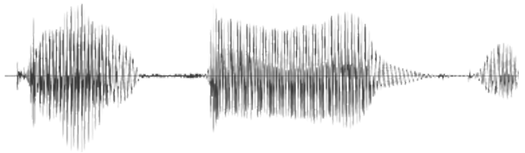
The two-stage pipeline



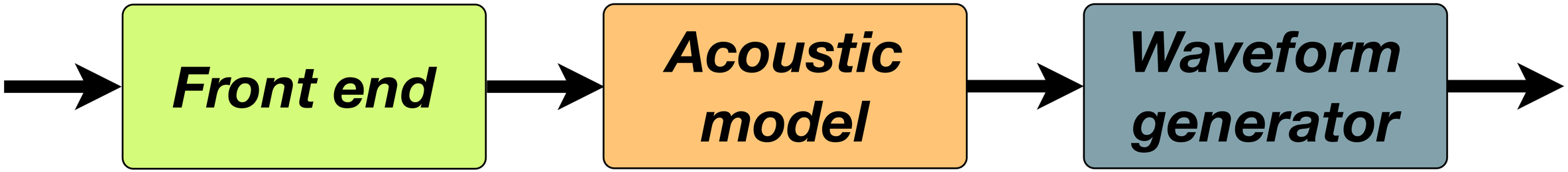
Author of the…

linguistic specification





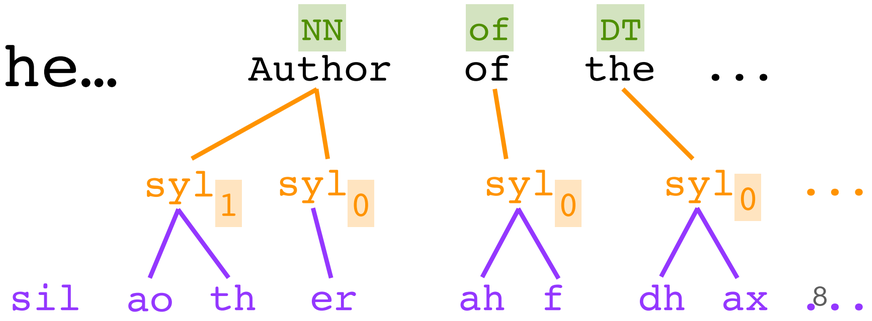
The three-stage pipeline



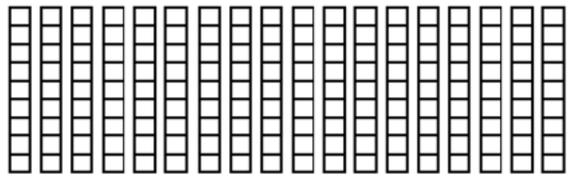
text

Author of the…

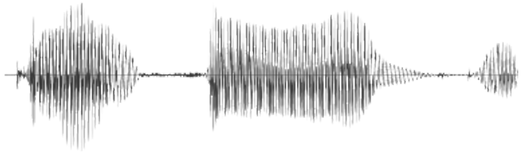
linguistic specification



acoustic features

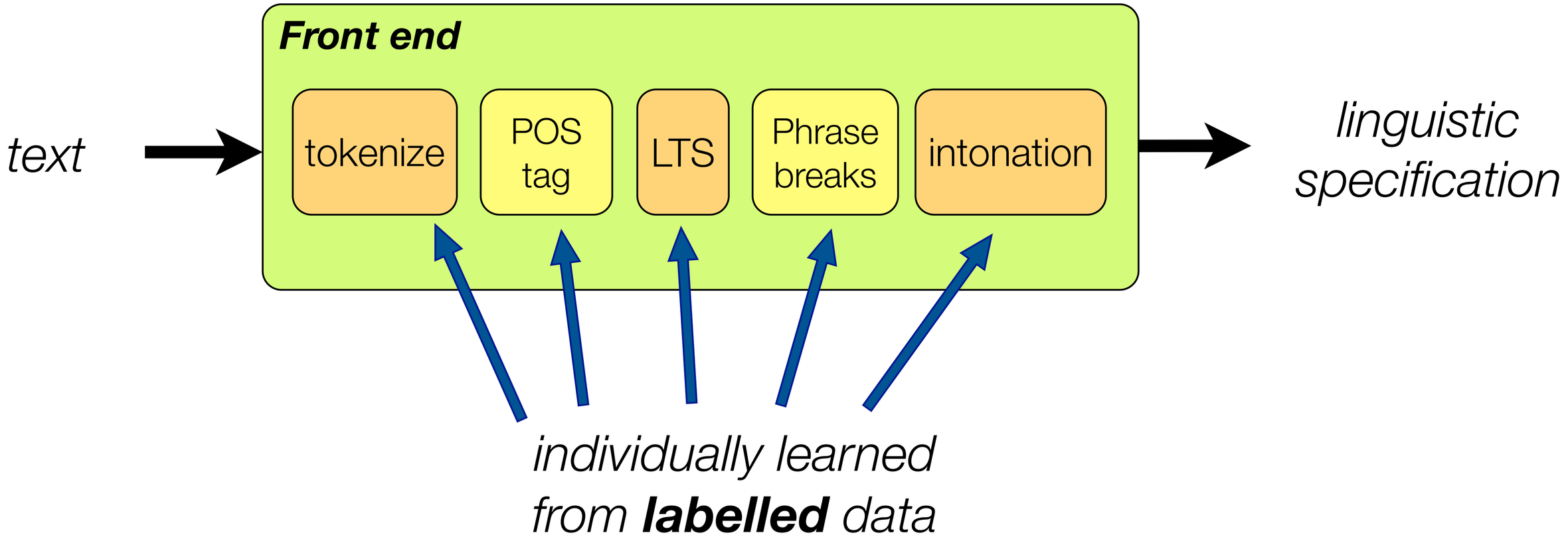


waveform



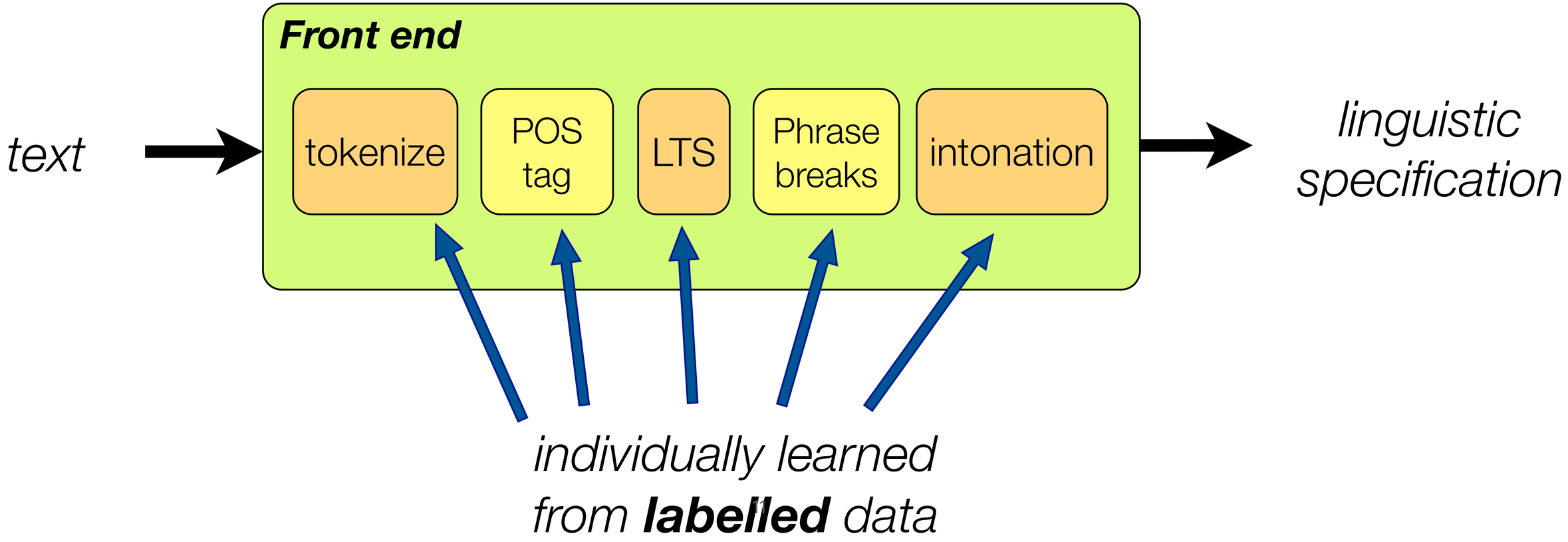
Front end

Front end


Front end
text
tokenize
POS tag
LTS
Phrase breaks
intonation
linguistic specification
individually learned from **labelled** data

Classic front end

- A chain of **processes**
- Each process is performed by a **model**
- These models are independently trained in a **supervised** fashion on annotated data



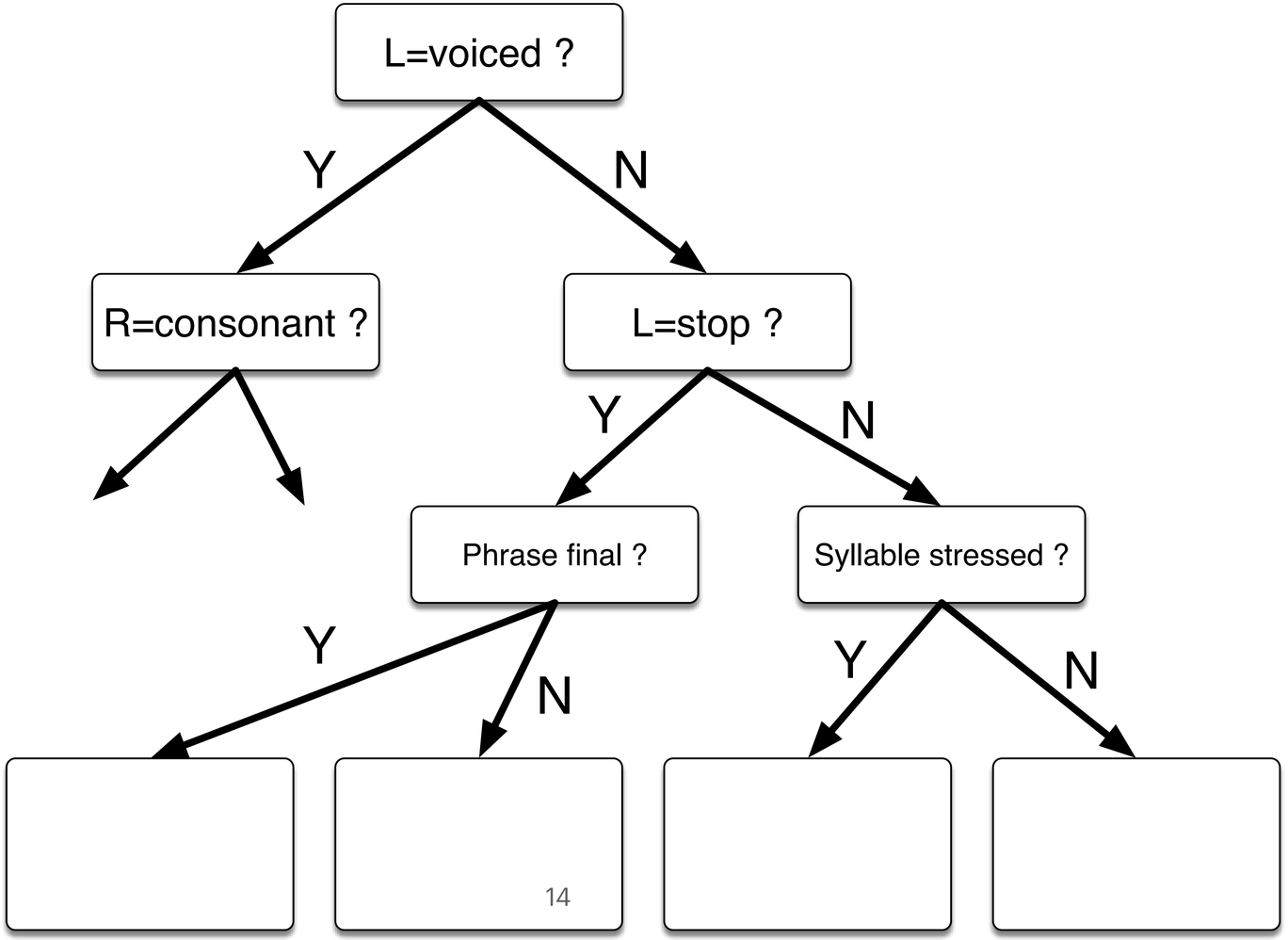
Neural front end

- Learn by a neural network

Acoustic model

Acoustic model - Decision tree

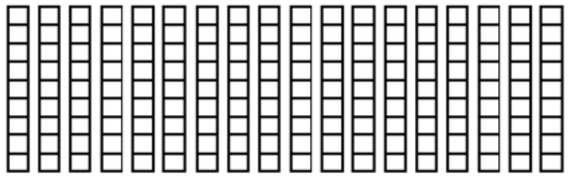
- Decision tree to group HMM states, which model acoustic feature distribution

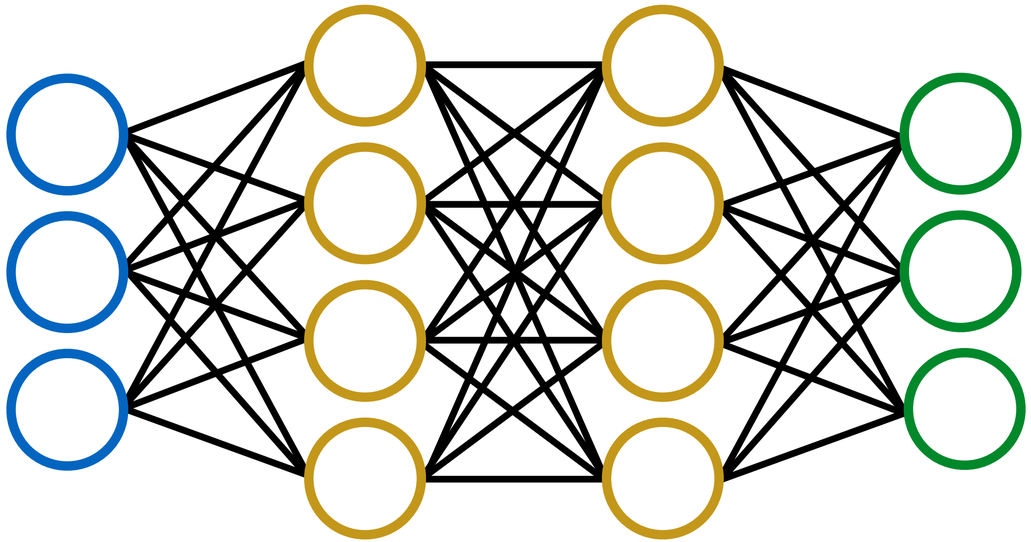


Acoustic model: DNN

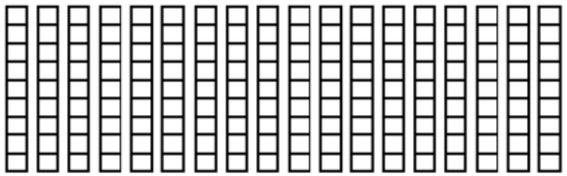
- Feedforward neural network

input
Linguistic features





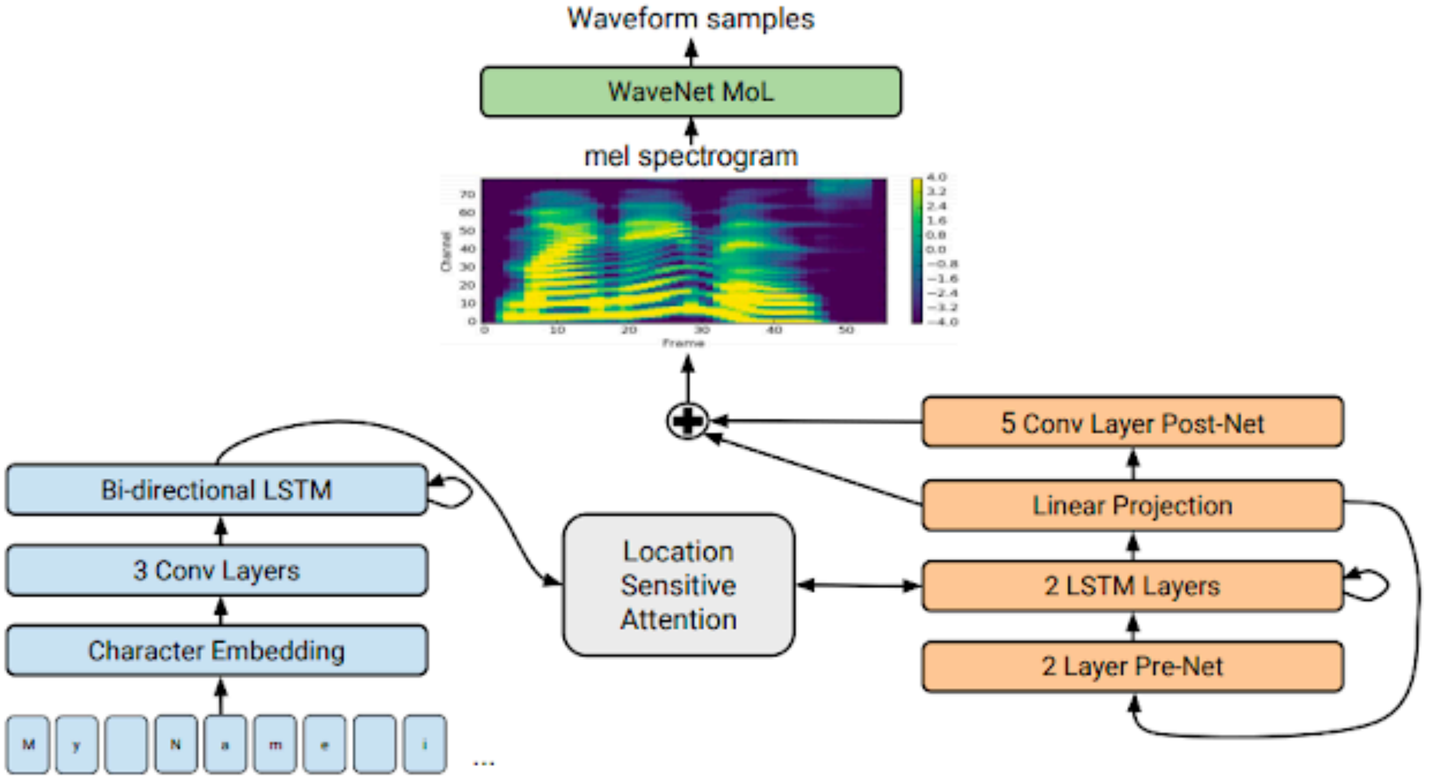
output
acoustic features



Watts, Oliver, Gustav Eje Henter, Thomas Merritt, Zhizheng Wu, and Simon King. "From HMMs to DNNs: where do the improvements come from?." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5505-5509. IEEE, 2016.

Acoustic model - RNN based

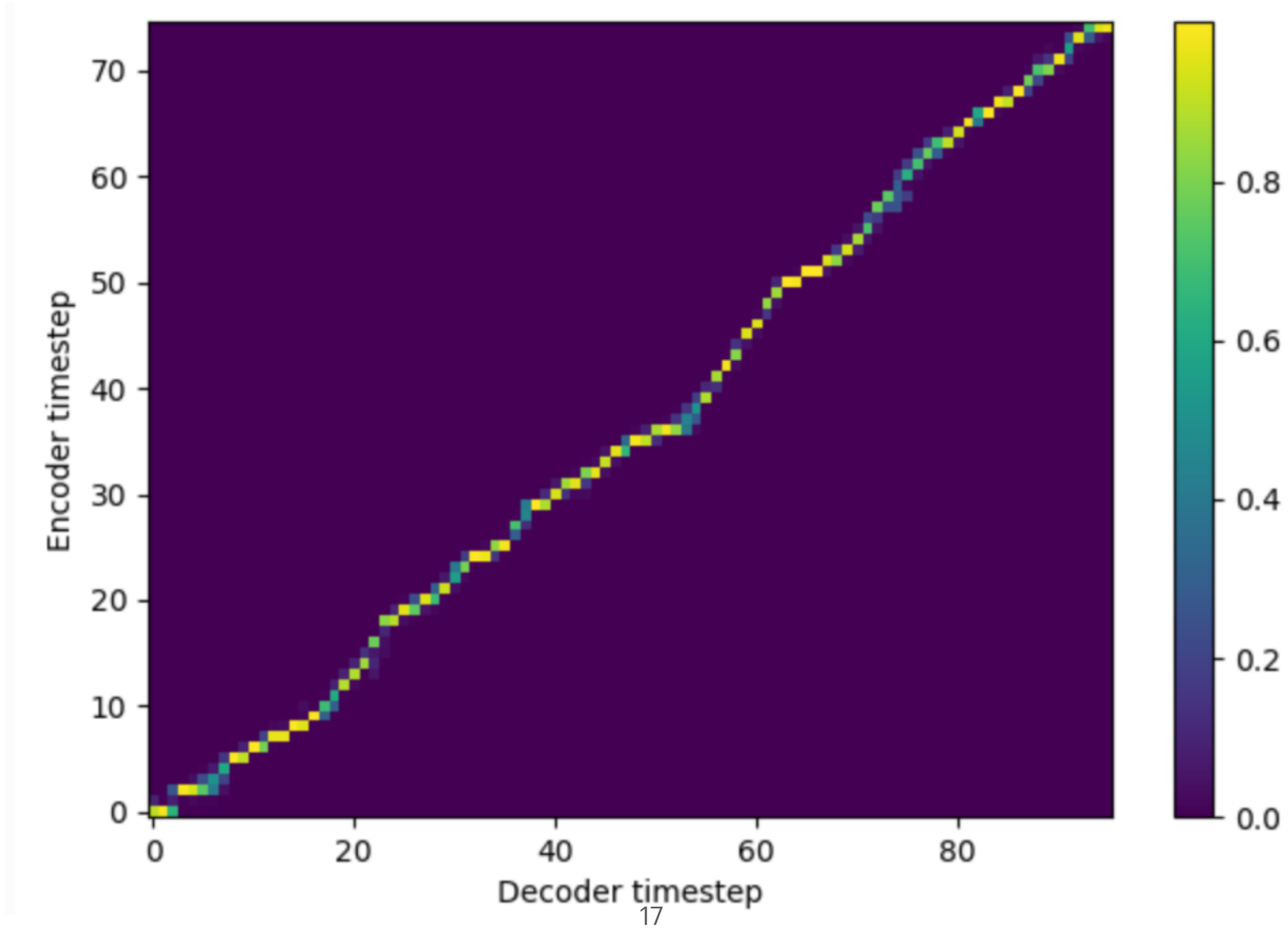
- Tacotron2: A sequence-to-sequence model based on Recurrent Neural Networks



Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

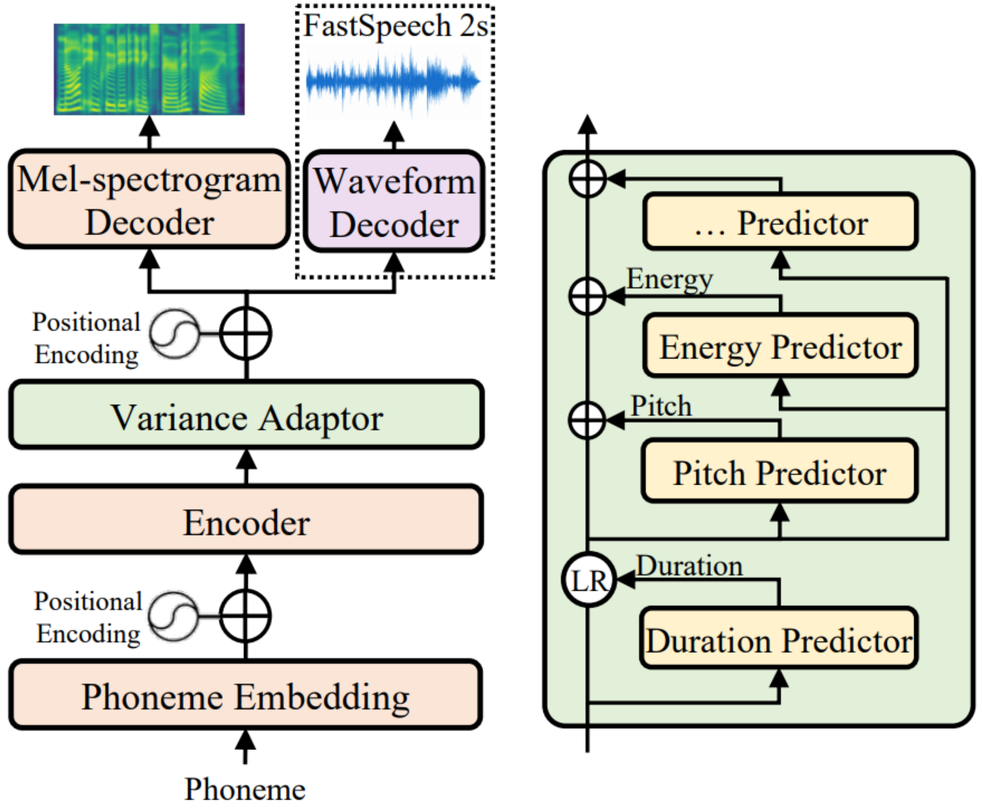
Acoustic model - RNN based

- Attention



Acoustic model - Transformer based

- FastSpeech2: parallel generation and not depending on the location attention

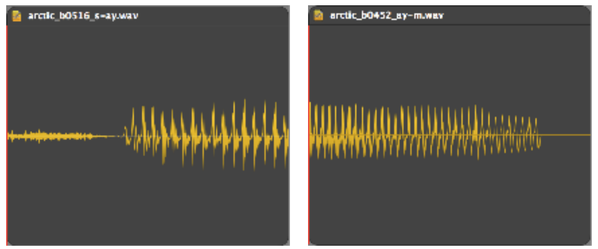


(a) FastSpeech 2 (b) Variance adaptor

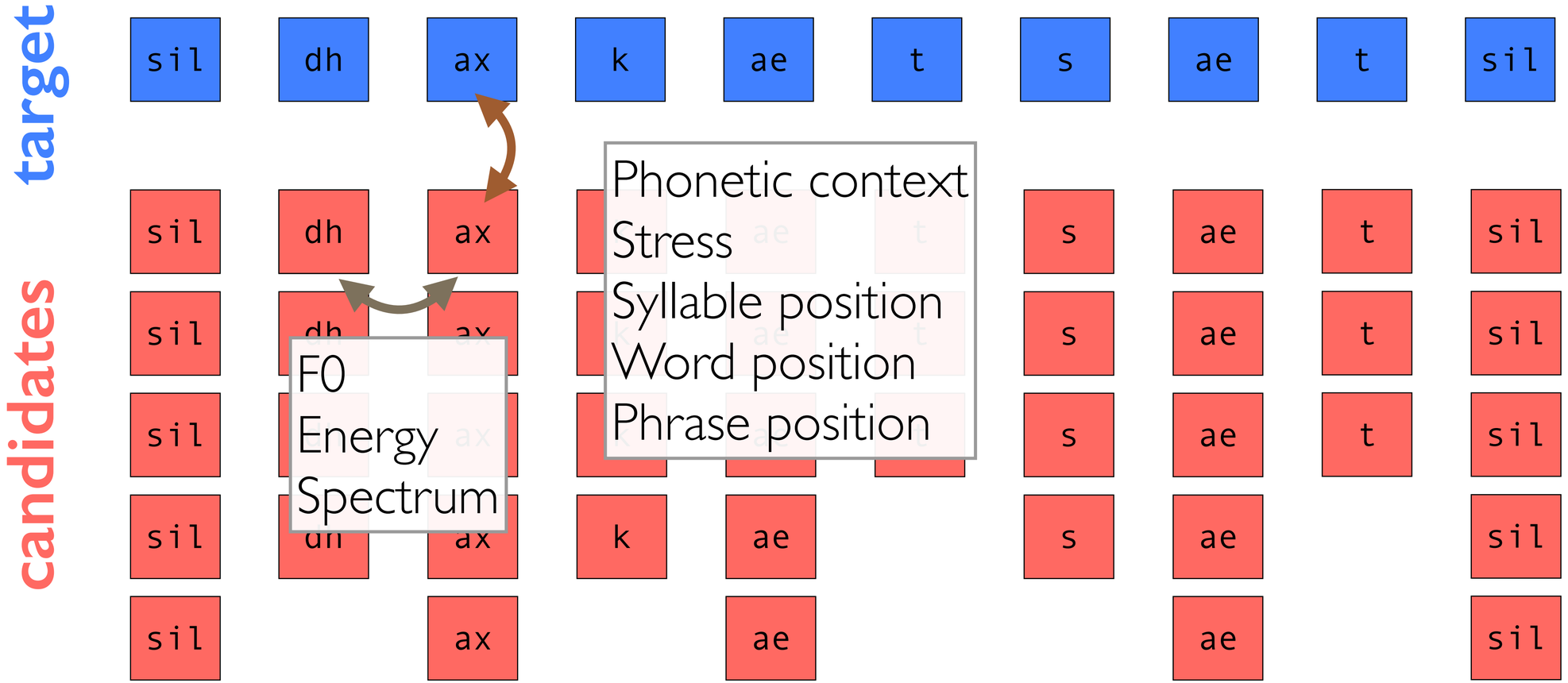
Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. "Fastspeech 2: Fast and high-quality end-to-end text to speech." *arXiv preprint arXiv:2006.04558* (2020).

Waveform generator

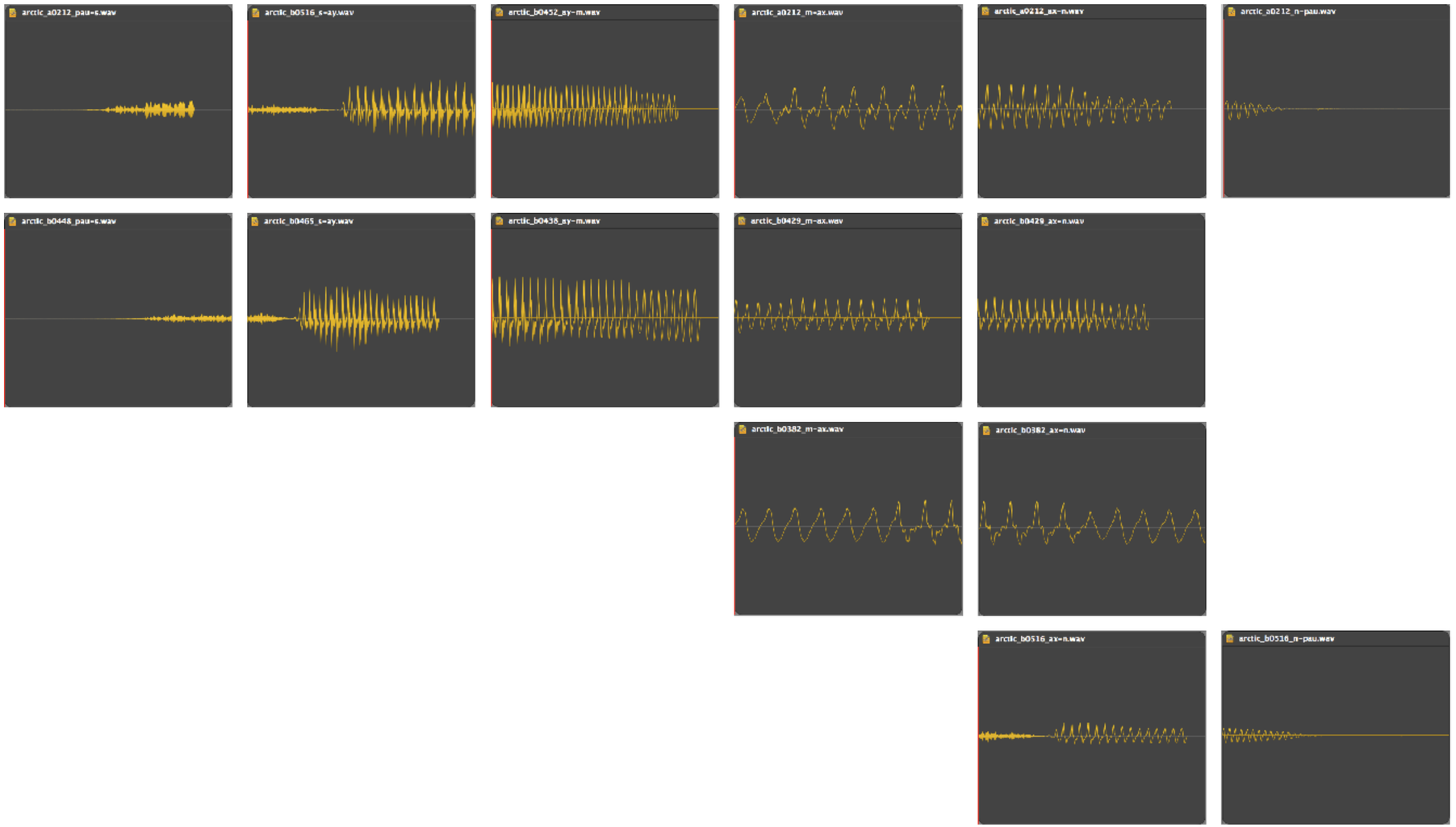
Waveform generator: Waveform concatenation



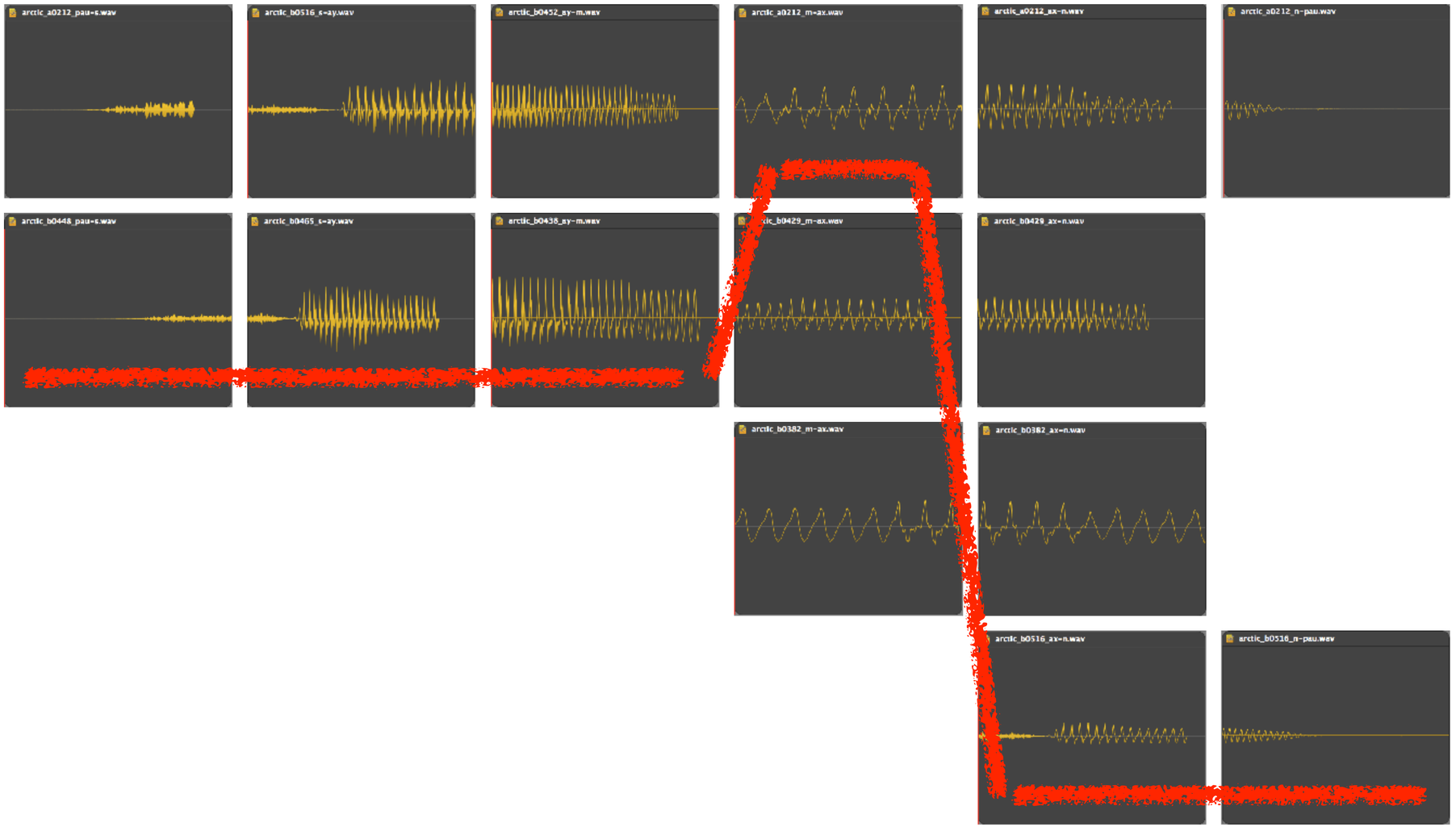
Classical unit selection (drawn here with phone units) - target and join costs



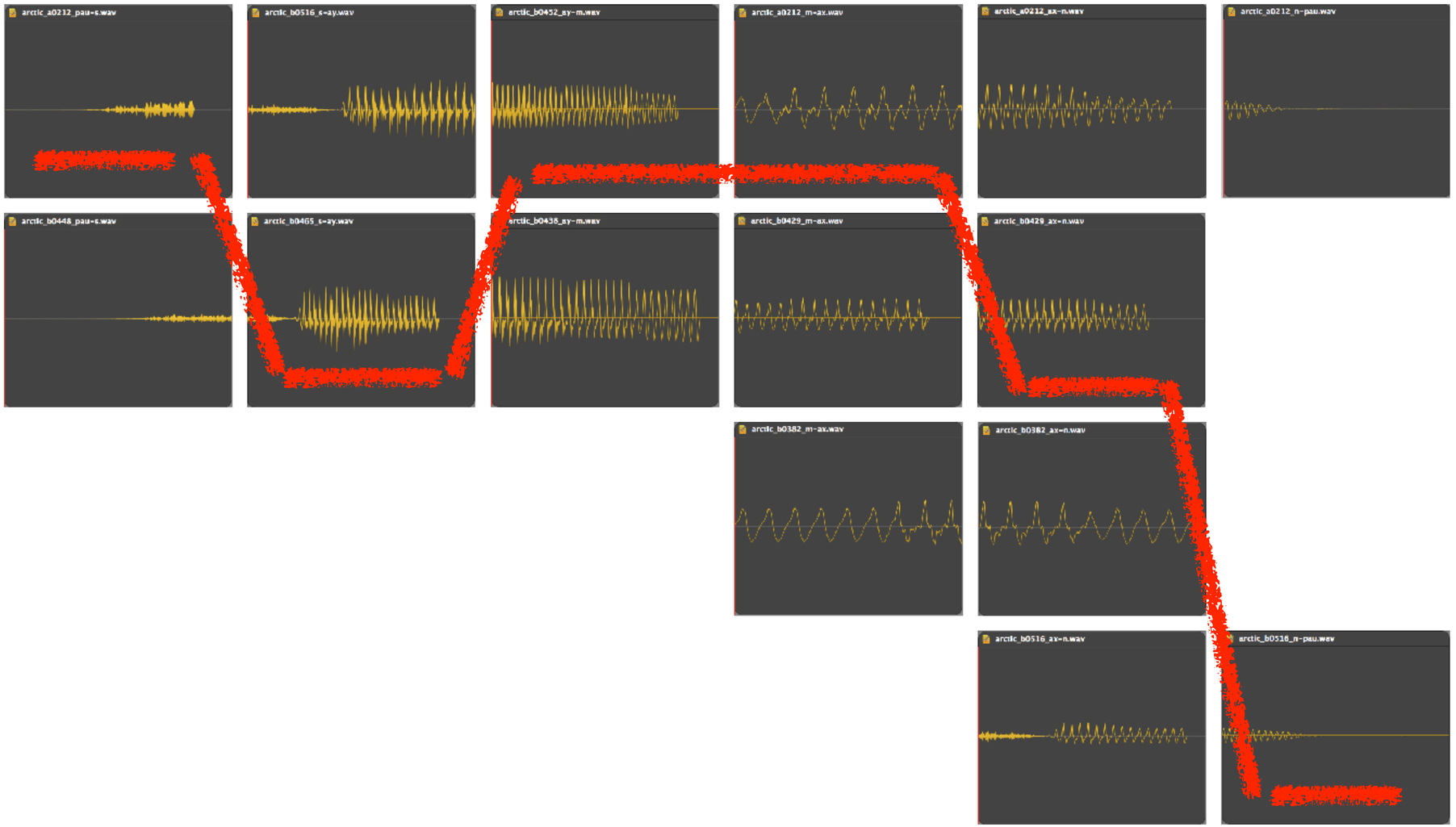
“Simon”


arctic_a0212_pau-s.wav
arctic_b0516_s-ay.wav
arctic_b0452_ay-m.wav
arctic_a0212_m-ax.wav
arctic_a0212_ax-n.wav
arctic_a0212_n-pau.wav
arctic_b0448_pau-s.wav
arctic_b0465_s-ay.wav
arctic_b0438_ay-m.wav
arctic_b0429_m-ax.wav
arctic_b0429_ax-n.wav
arctic_b0382_m-ax.wav
arctic_b0382_ax-n.wav
arctic_b0516_ax-n.wav
arctic_b0516_n-pau.wav

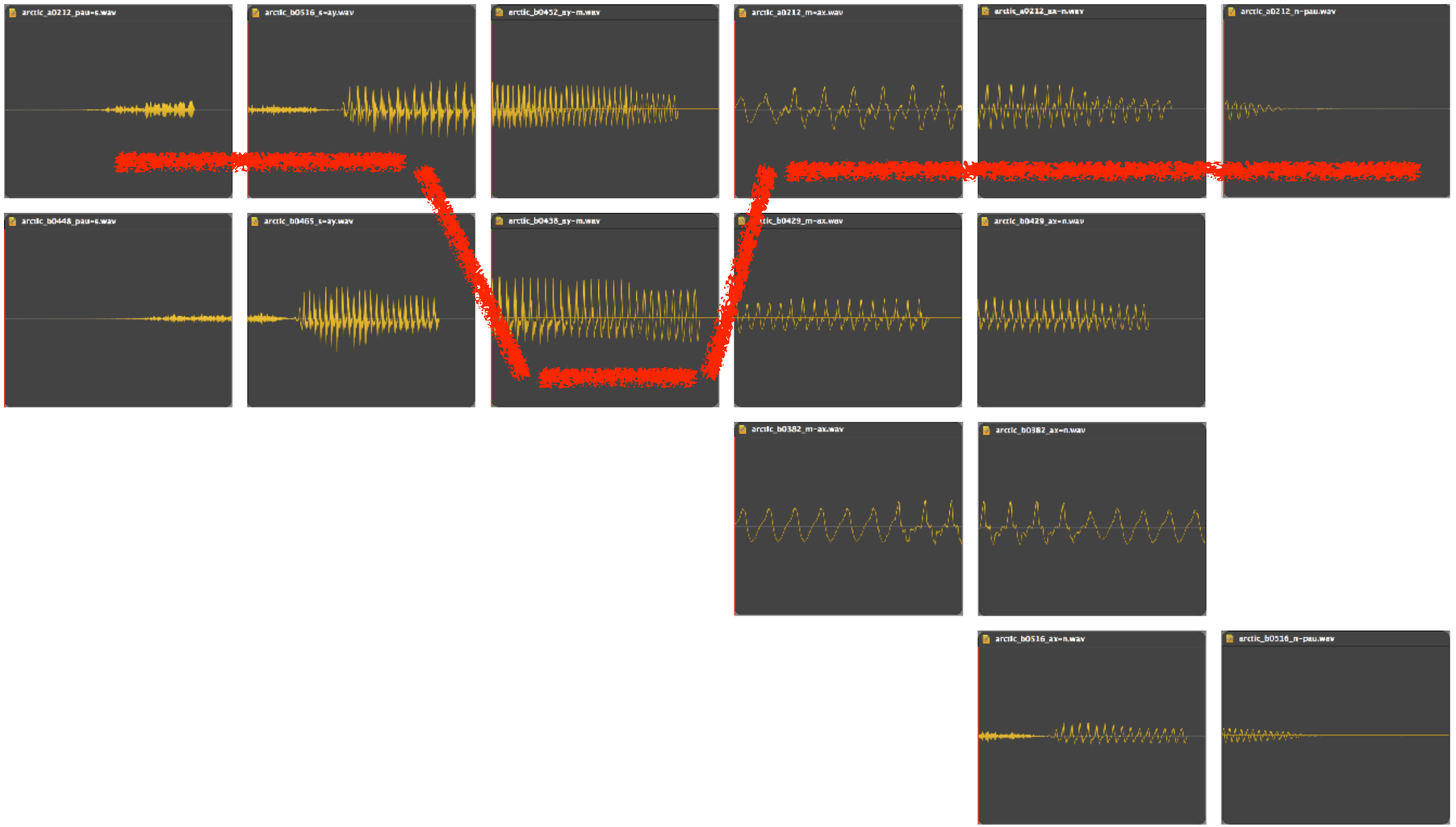
“Simon”


arctic_a0212_pau-s.wav
arctic_b0516_s-ay.wav
arctic_b0452_ay-m.wav
arctic_a0212_m-ax.wav
arctic_a0212_ax-n.wav
arctic_a0212_n-pau.wav
arctic_b0448_pau-s.wav
arctic_b0465_s-ay.wav
arctic_b0438_ay-m.wav
arctic_b0429_m-ax.wav
arctic_b0429_ax-n.wav
arctic_b0382_m-ax.wav
arctic_b0382_ax-n.wav
arctic_b0516_ax-n.wav
arctic_b0516_n-pau.wav

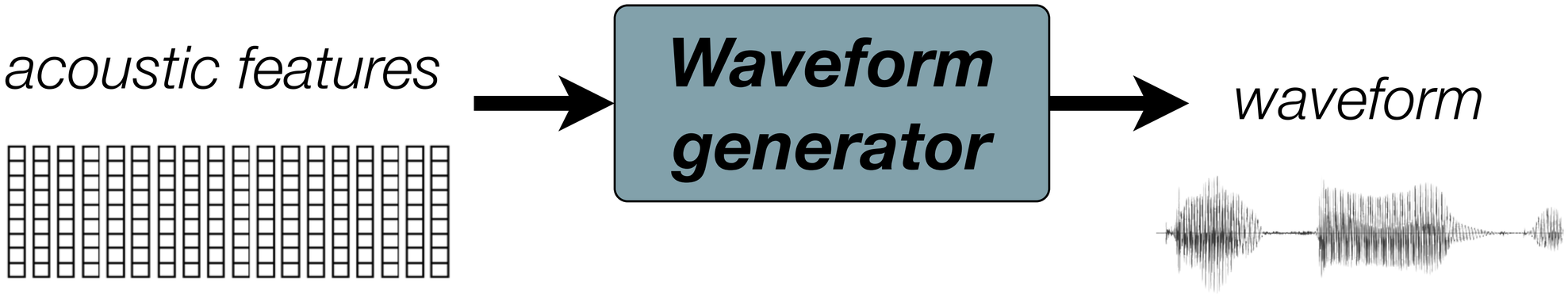
“Simon”


arctic_a0212_pau-s.wav
arctic_b0516_s-ay.wav
arctic_b0452_ay-m.wav
arctic_a0212_m-ax.wav
arctic_a0212_ax-n.wav
arctic_a0212_n-pau.wav
arctic_b0448_pau-s.wav
arctic_b0465_s-ay.wav
arctic_b0438_ay-m.wav
arctic_b0429_m-ax.wav
arctic_b0429_ax-n.wav
arctic_b0382_m-ax.wav
arctic_b0382_ax-n.wav
arctic_b0516_ax-n.wav
arctic_b0516_n-pau.wav

“Simon”

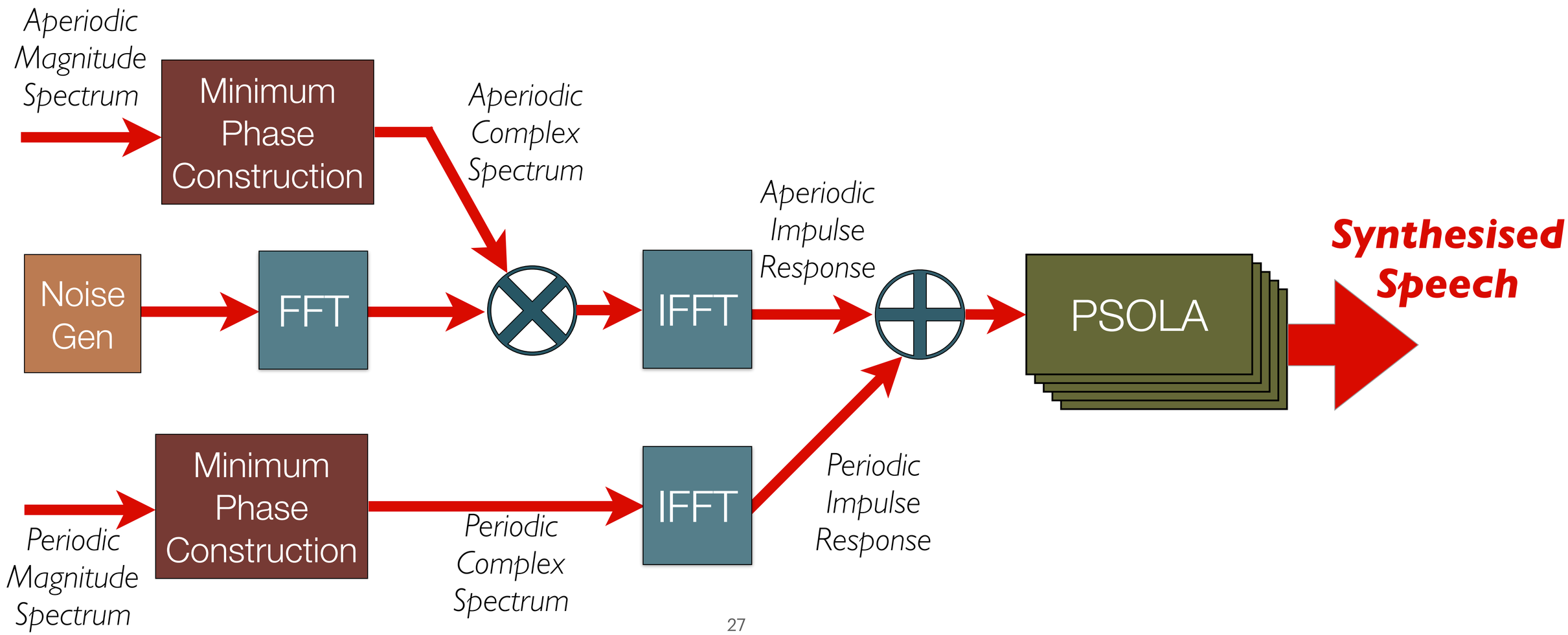

arctic_a0212_pau-s.wav
arctic_b0516_s-ay.wav
arctic_b0452_ay-m.wav
arctic_a0212_m-ax.wav
arctic_a0212_ax-n.wav
arctic_a0212_n-pau.wav
arctic_b0448_pau-s.wav
arctic_b0465_s-ay.wav
arctic_b0438_ay-m.wav
arctic_b0429_m-ax.wav
arctic_b0429_ax-n.wav
arctic_b0382_m-ax.wav
arctic_b0382_ax-n.wav
arctic_b0516_ax-n.wav
arctic_b0516_n-pau.wav

Waveform generator: Vocoder



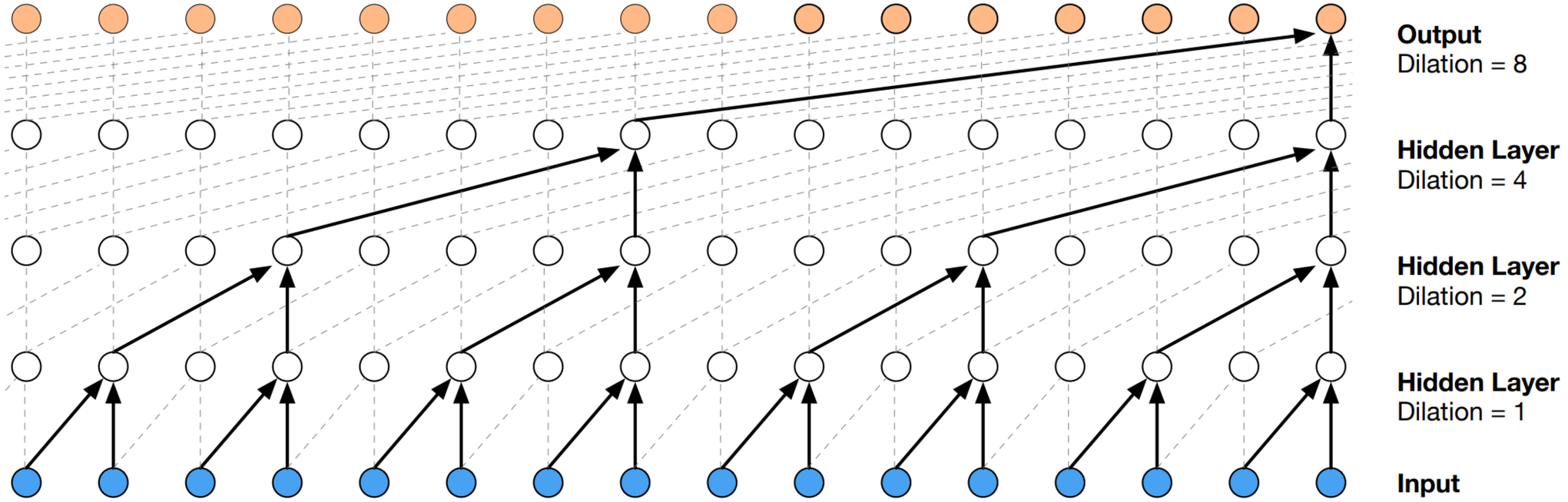
Vocoder - Signal processing based

- WORLD vocoder



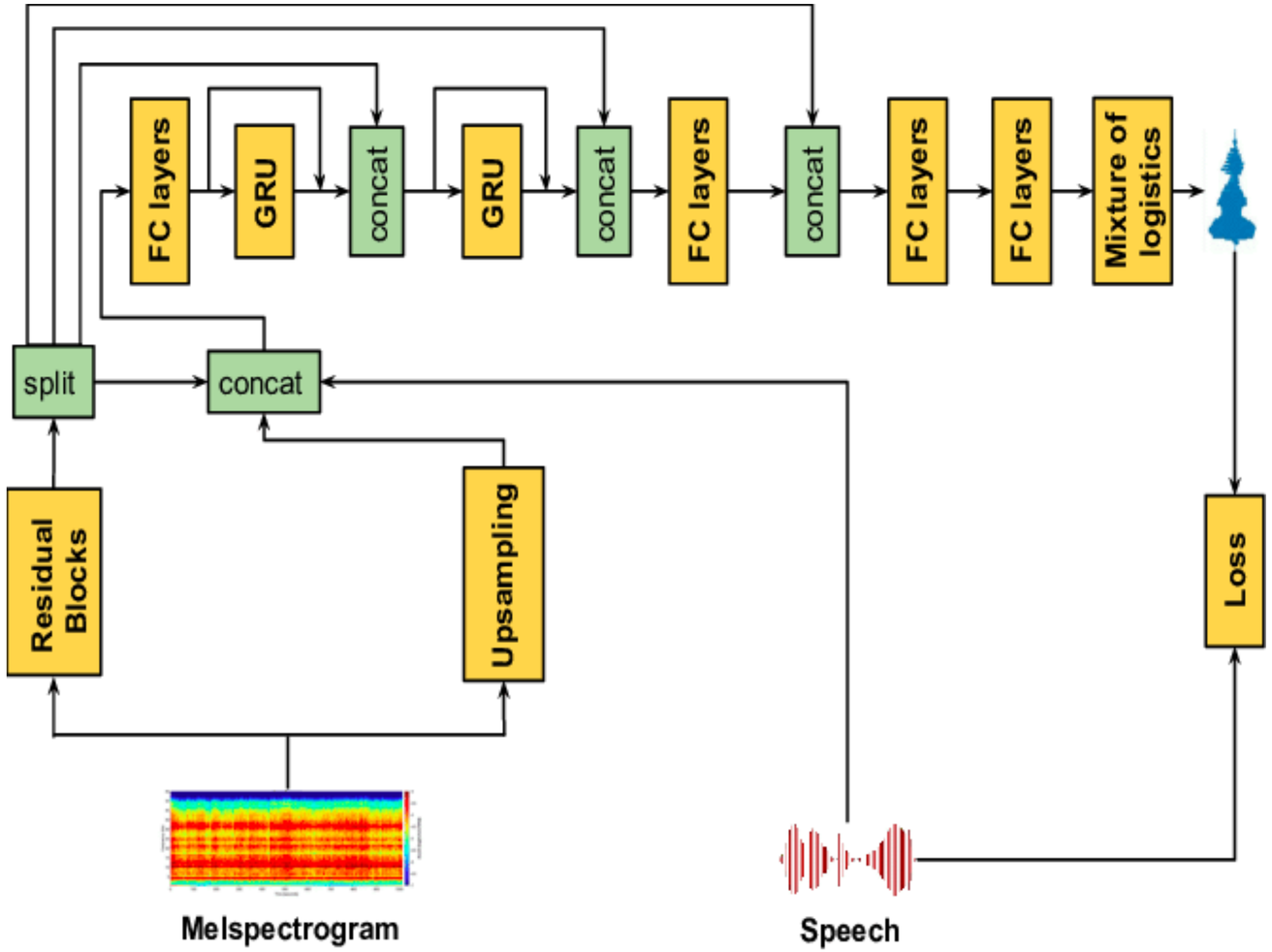
Vocoder: Autoregressive

- WaveNet: autoregressive model with dilated causal convolution



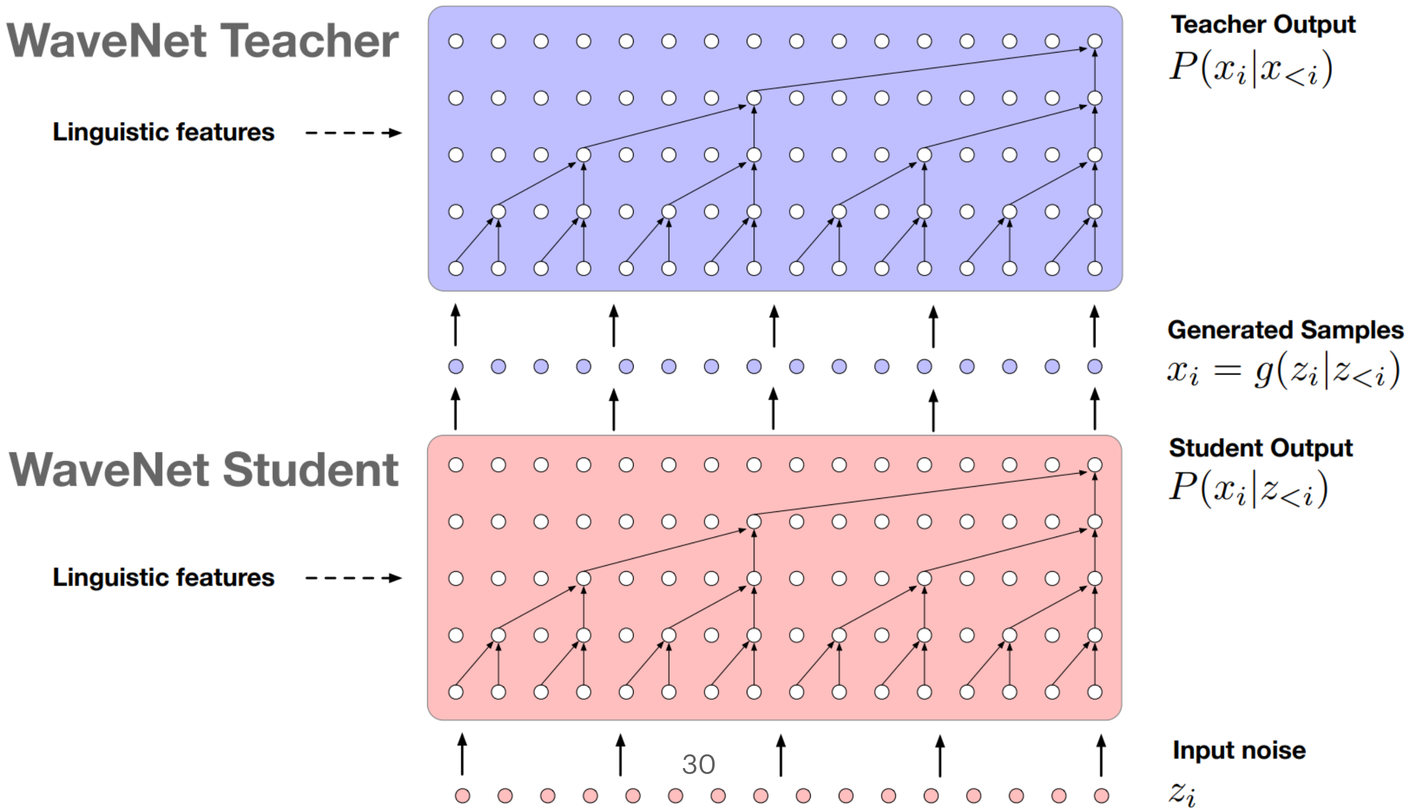
Vocoder: Autoregressive

- WaveRNN: autoregressive model with RNN



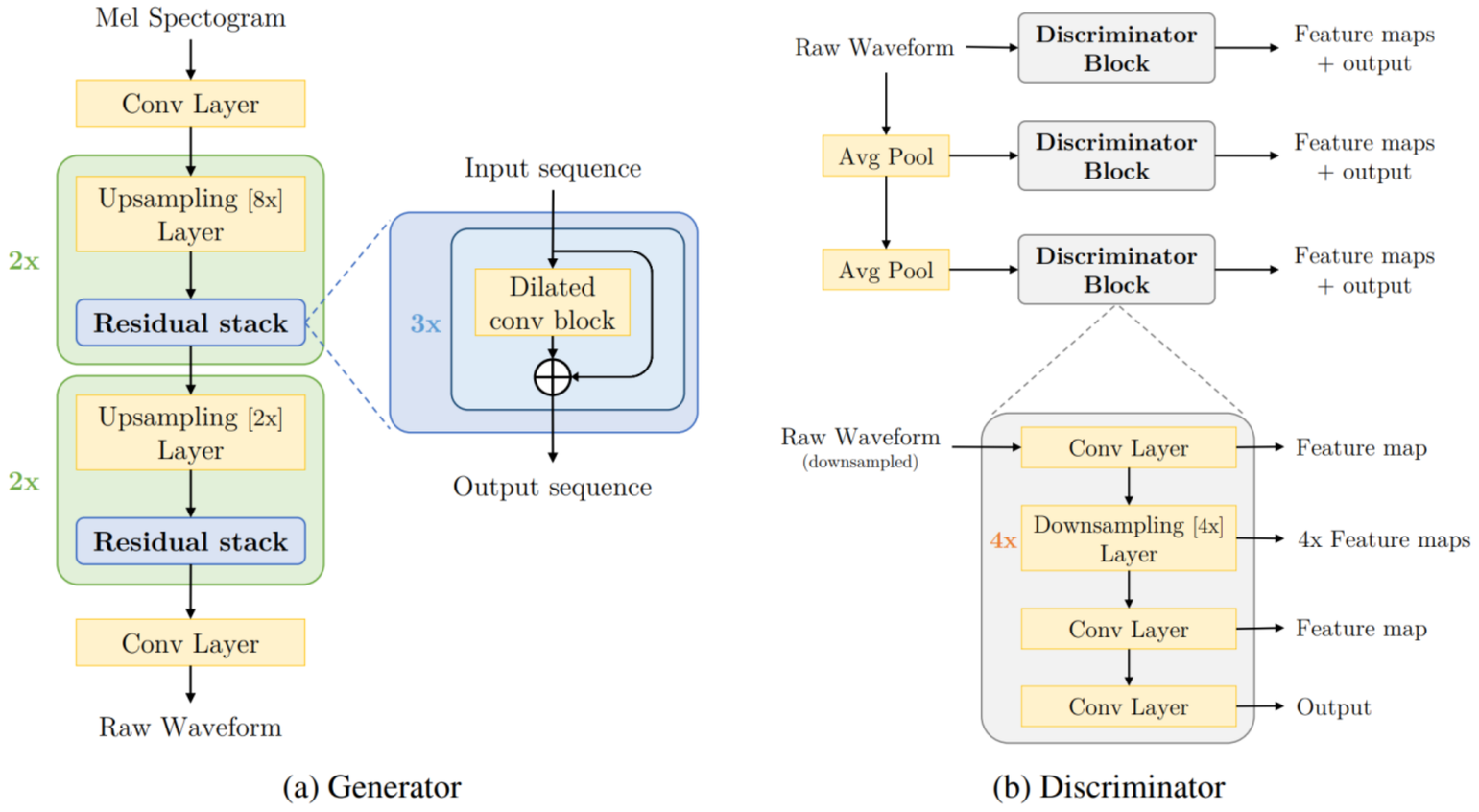
Vocoder: Flow based

- AF (autoregressive flow) and IAF (inverse autoregressive flow)
  - Parallel inference of IAF student
  - Parallel training of AF teacher



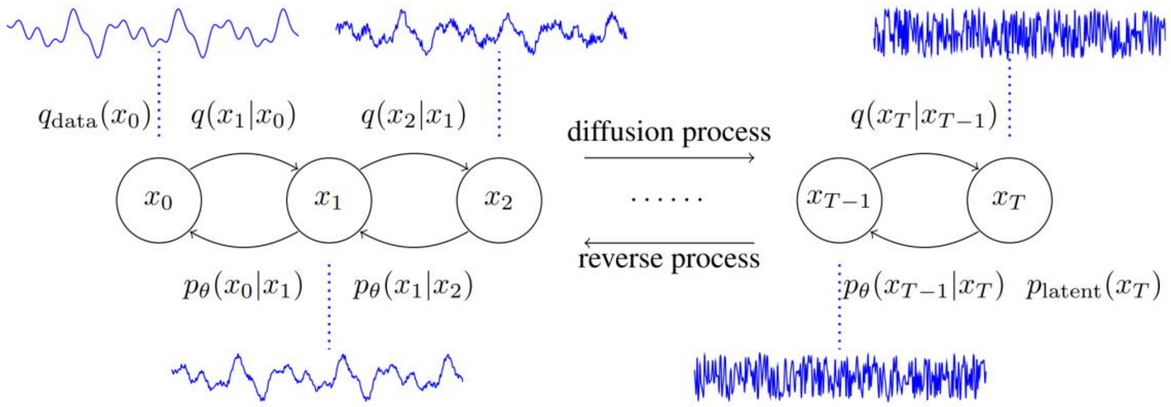
Vocoder: GAN based

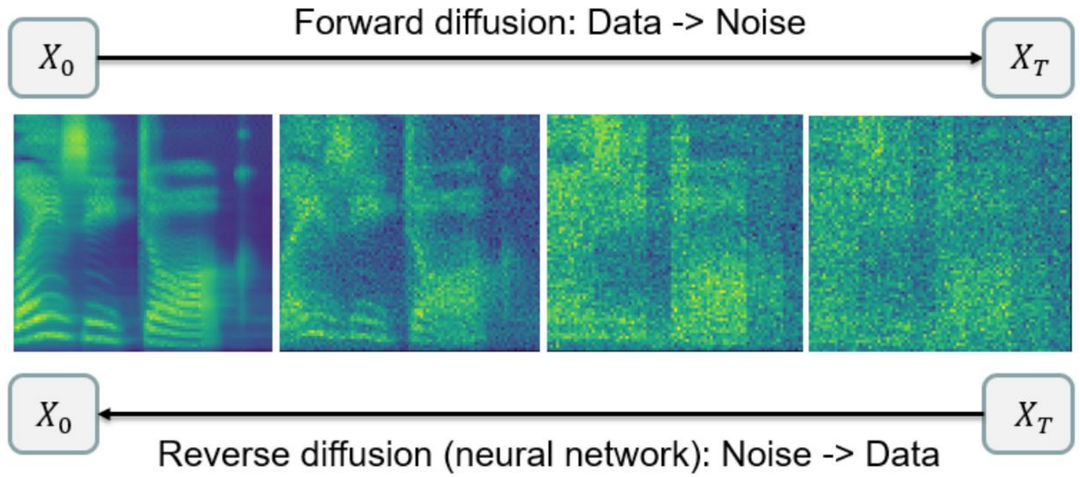
- MelGAN: Generator + Discriminator



Vocoder: Diffusion based

- Diffusion probabilistic model
  - Forward process: diffusion
  - Reverse process: denoising





Tools

- TTS open-source
  - https://github.com/coqui-ai/TTS
  - https://github.com/espnet/espnet

- Acoustic models
  - Tacotron2: https://github.com/NVIDIA/tacotron2
  - Fastspeech2: https://github.com/ming024/FastSpeech2

- Vocoder
  - https://github.com/coqui-ai/TTS/tree/dev/TTS/vocoder
  - https://github.com/NVIDIA/BigVGAN

Readings

- Interspeech 2022 TTS tutorial
  - https://github.com/tts-tutorial/interspeech2022/blob/main/INTERSPEECH_Tutorial_TTS.pdf
- Text-to-Speech Synthesis
  - https://www.cambridge.org/core/books/texttospeech-synthesis/D2C567CEF939C7D15B2F1232992C7836