Text-To-Speech Synthesis

Vasileios Tsiaras

3 December 2025

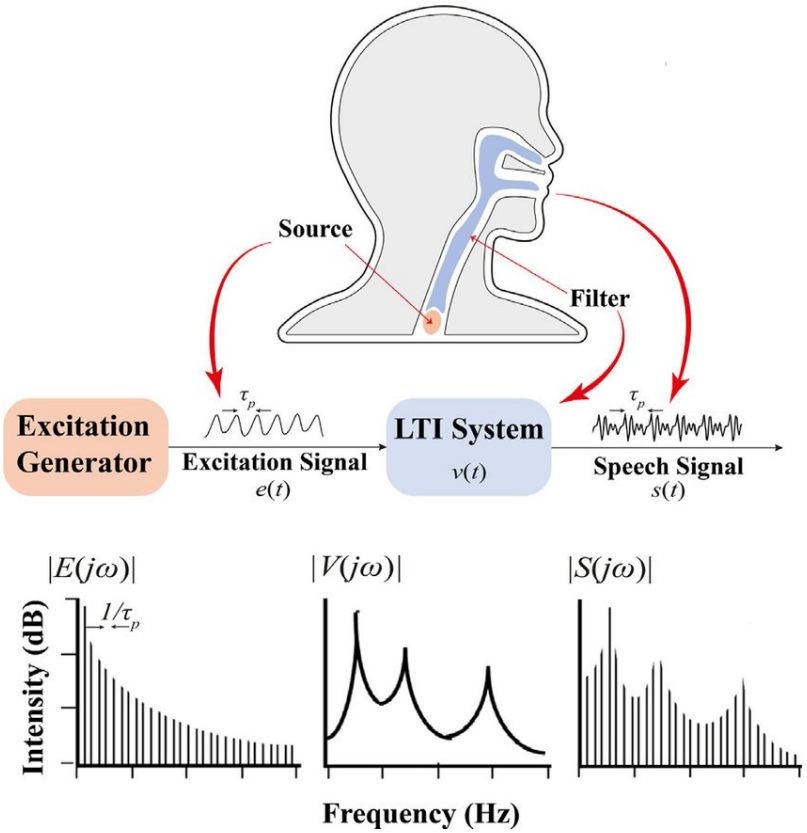
Text-To-Speech Synthesis

- Text-to-Speech (TTS) is a technology that converts text into human-like speech.
- 1961, Early speech synthesis:
  - Researchers John Kelly, Carol Lochbaum, and Max Mathews at Bell Labs created the first computer-generated recording of “Daisy Bell (Bicycle Built for Two)”.

- 1978, Speak & Spell:
  - A popular educational toy by Texas Instruments
  - It used a single-chip linear predictive coding (LPC) synthesizer, making speech synthesis accessible to the public.

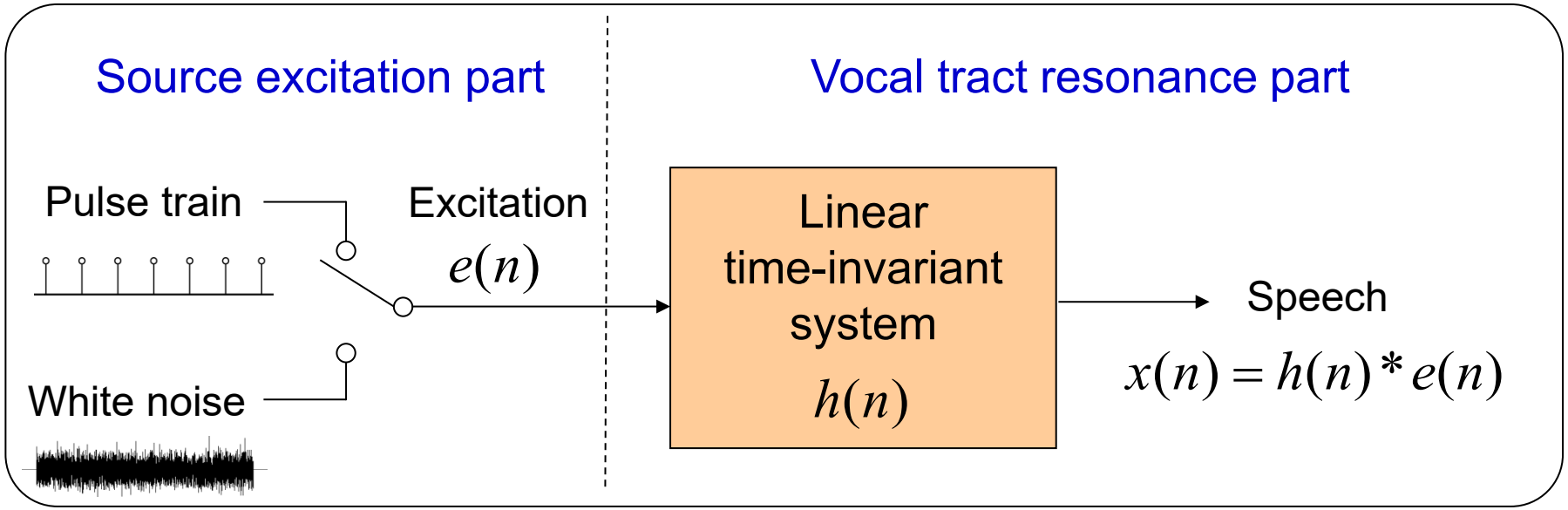
Text-To-Speech Synthesis

- 1984, DECtalk Voice:
- The process involves several stages:
    1. Text Analysis
        - Letter-to-sound conversion
        - Phonetic and Prosodic Processing
    2. Vocal Tract Modeling: "source-filter" model
        - Sound Source:
            a) A periodic pulse train for voiced sounds (vowels, 'm', 'z', etc.).
            b) White noise for unvoiced sounds (fricatives like 's', 'f', 'sh', etc.).
        - Acoustic Filtering (Formants): The sound source is passed through an all-pole filter that simulates the resonant frequencies (formants) of the human vocal tract (throat, tongue, lips, nasal cavity).



Text-To-Speech Synthesis

- Source-Filter model:



$$x(n) = h(n) * e(n)$$

Fourier transform

$$X(e^{j\omega}) = H(e^{j\omega})E(e^{j\omega})$$

Text-To-Speech Synthesis

- Source-Filter model:

Autoregressive (AR) model	Exponential (EX) model
$H(z) = \dfrac{K}{1 - \sum_{m=0}^{M} c(m) z^{-m}}$	$H(z) = \exp \sum_{m=0}^{M} c(m) z^{-m}$

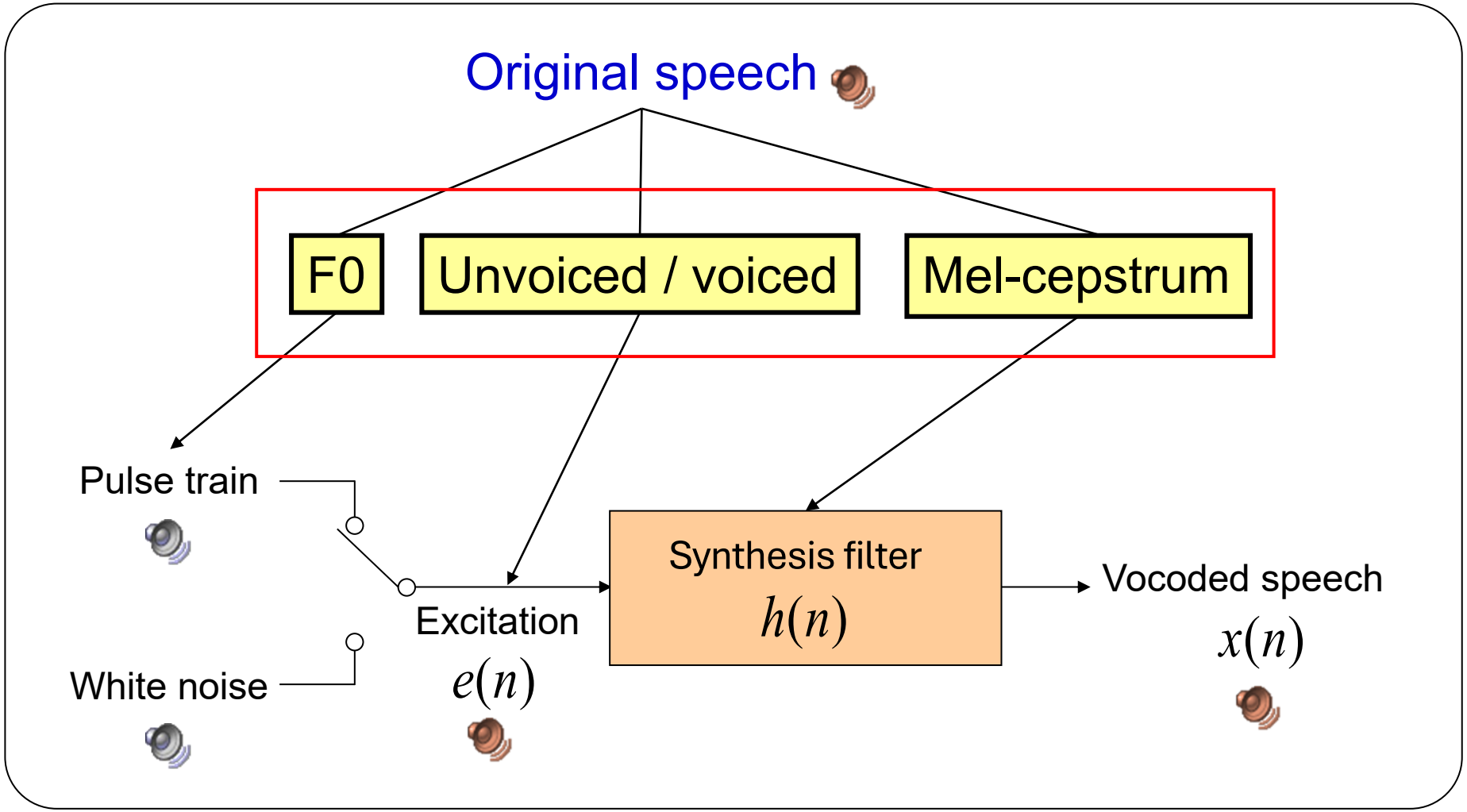
Estimate model parameters based on ML

$$\boldsymbol{c} = \arg\max_{\boldsymbol{c}} p(\boldsymbol{x} \mid \boldsymbol{c})$$

- $p(\boldsymbol{x} \mid \boldsymbol{c})$: AR model $\rightarrow$ Linear predictive analysis
- $p(\boldsymbol{x} \mid \boldsymbol{c})$: EX model $\rightarrow$ (ML-based) cepstral analysis

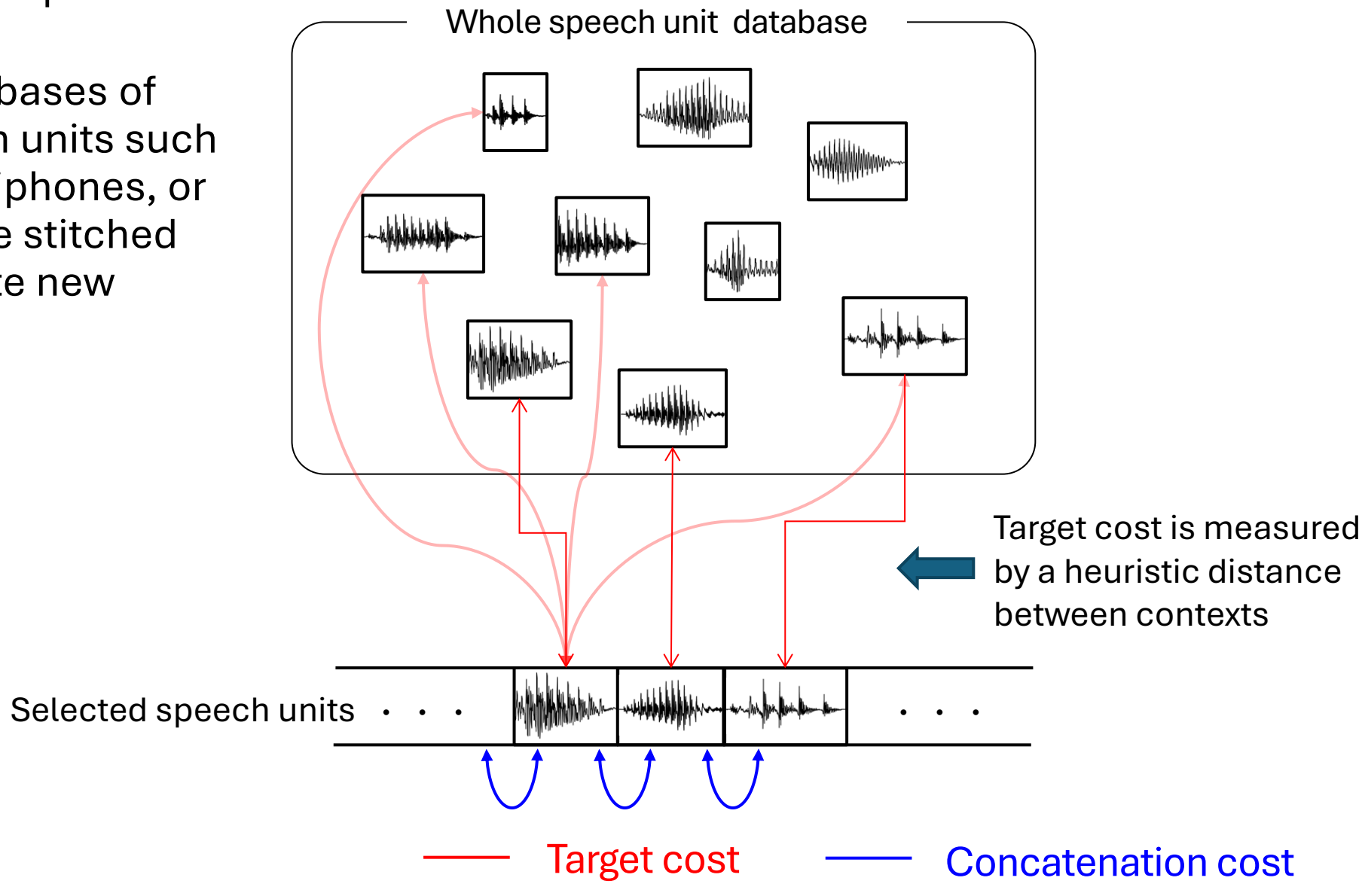
Text-To-Speech Synthesis

- Speech Vocoding



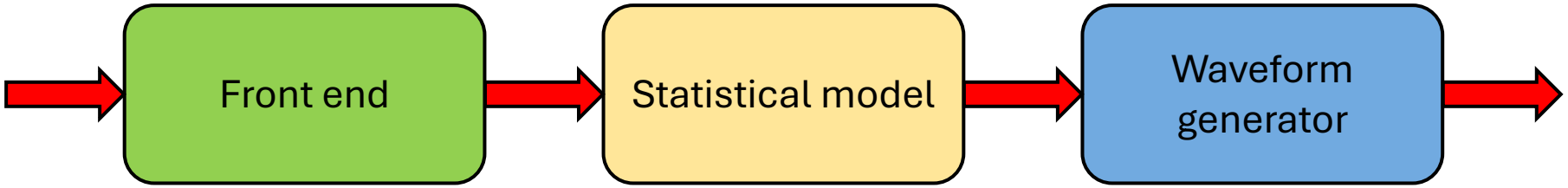
Text-To-Speech Synthesis

- 1990, Concatenative speech synthesis:
  - Using large databases of recorded speech units such as phonemes, diphones, or syllables that are stitched together to create new speech.
  - Natural
  - Inflexible



Text-To-Speech Synthesis

- 1994, Statistical Parametric Speech Synthesis:



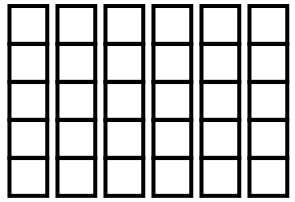
text

There was a change now

Linguistic specification

pau dh eh r w aa z ax ch ey n jh n aw pau

Speech features

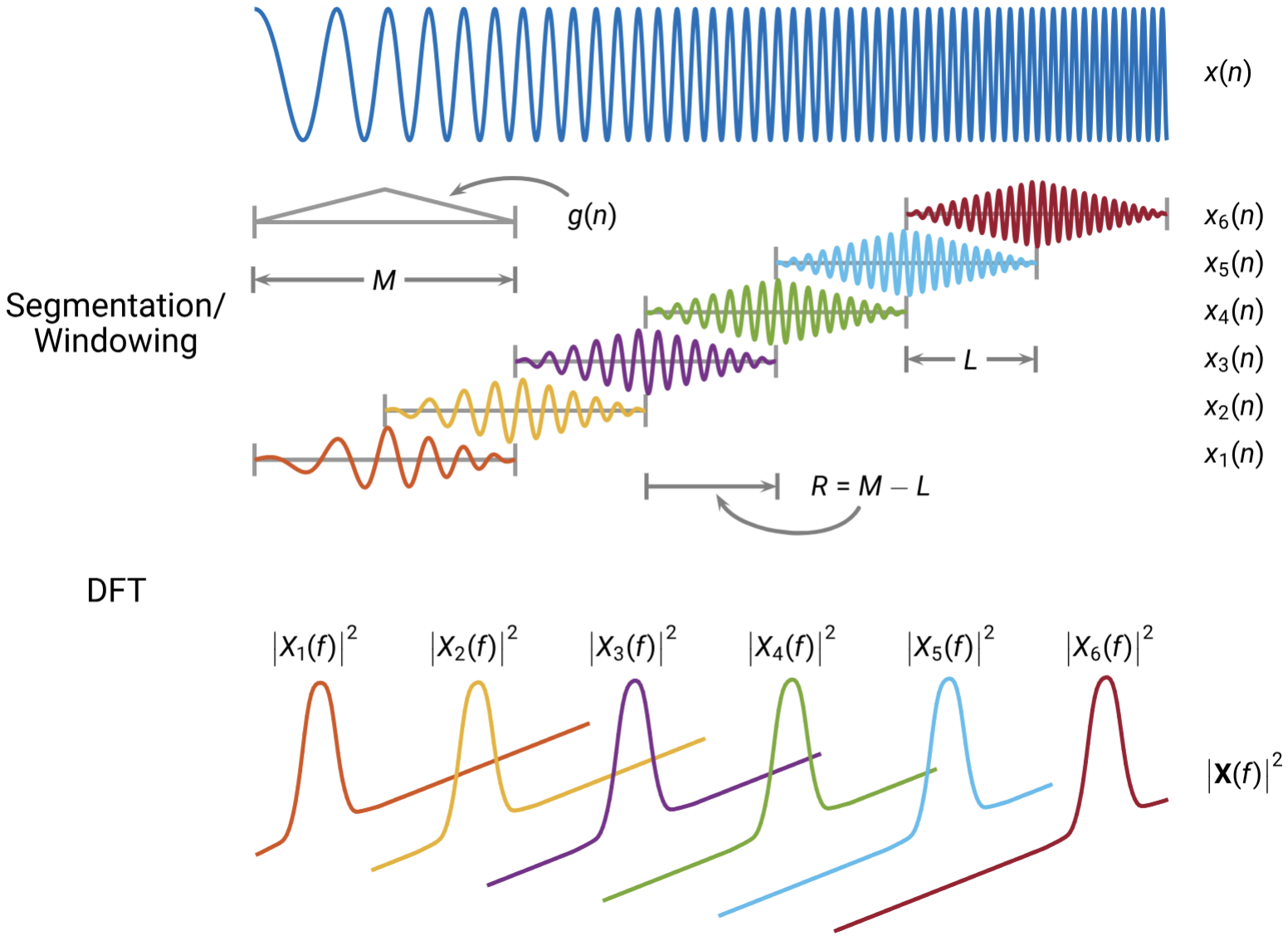


mcep, F0, bap

waveform

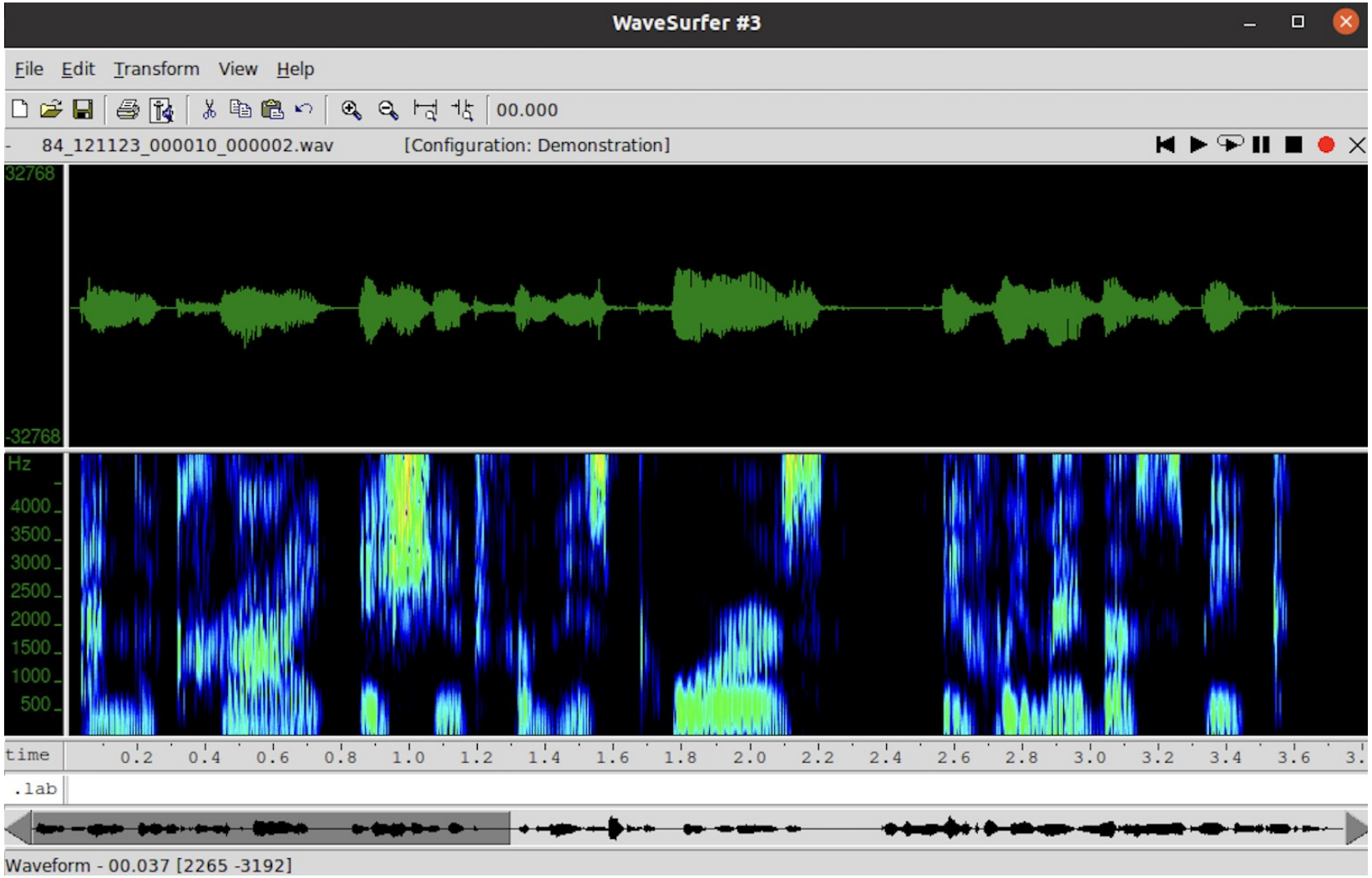
Text-To-Speech Synthesis

- Spectrogram



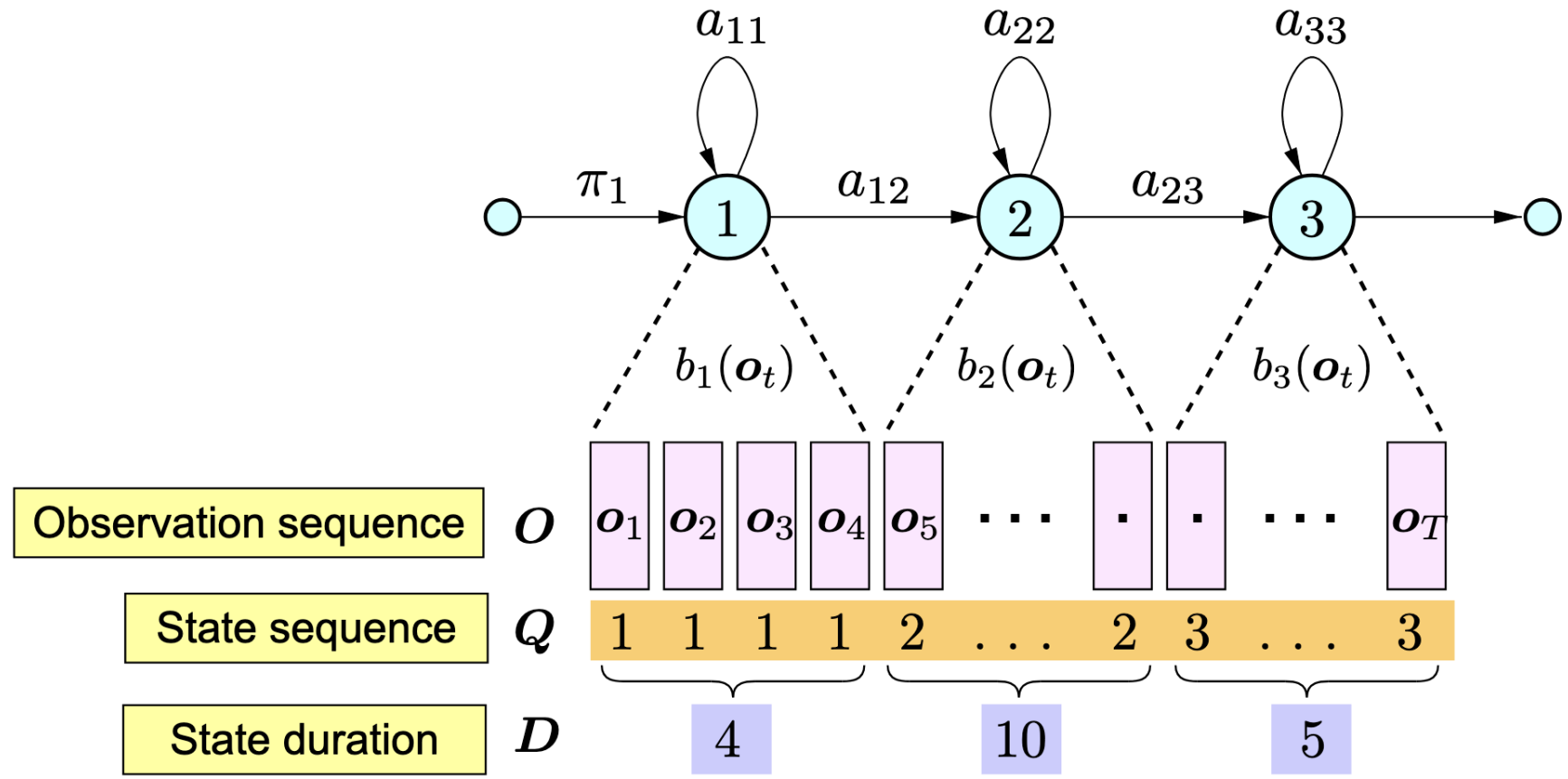
Text-To-Speech Synthesis

- Spectrogram



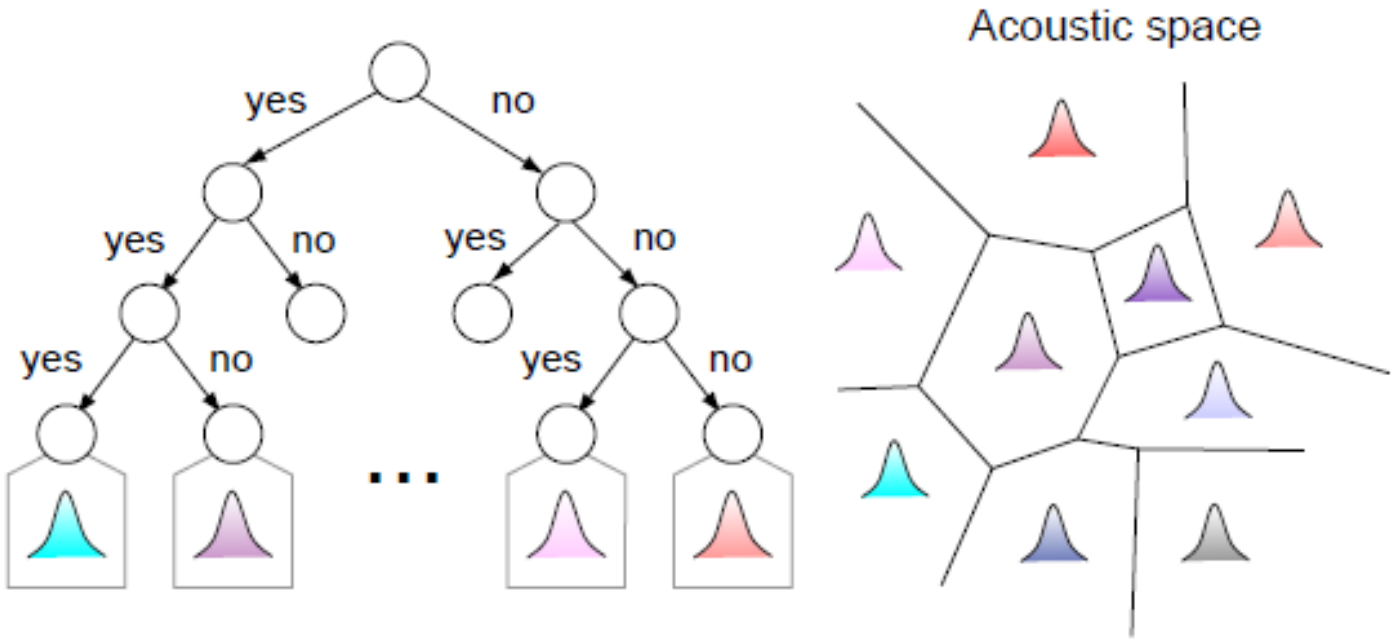
Text-To-Speech Synthesis

- 1994, Statistical Parametric Speech Synthesis:
  - Hidden Markov Models (HMM)
  - Over-smoothing
  - Muffled



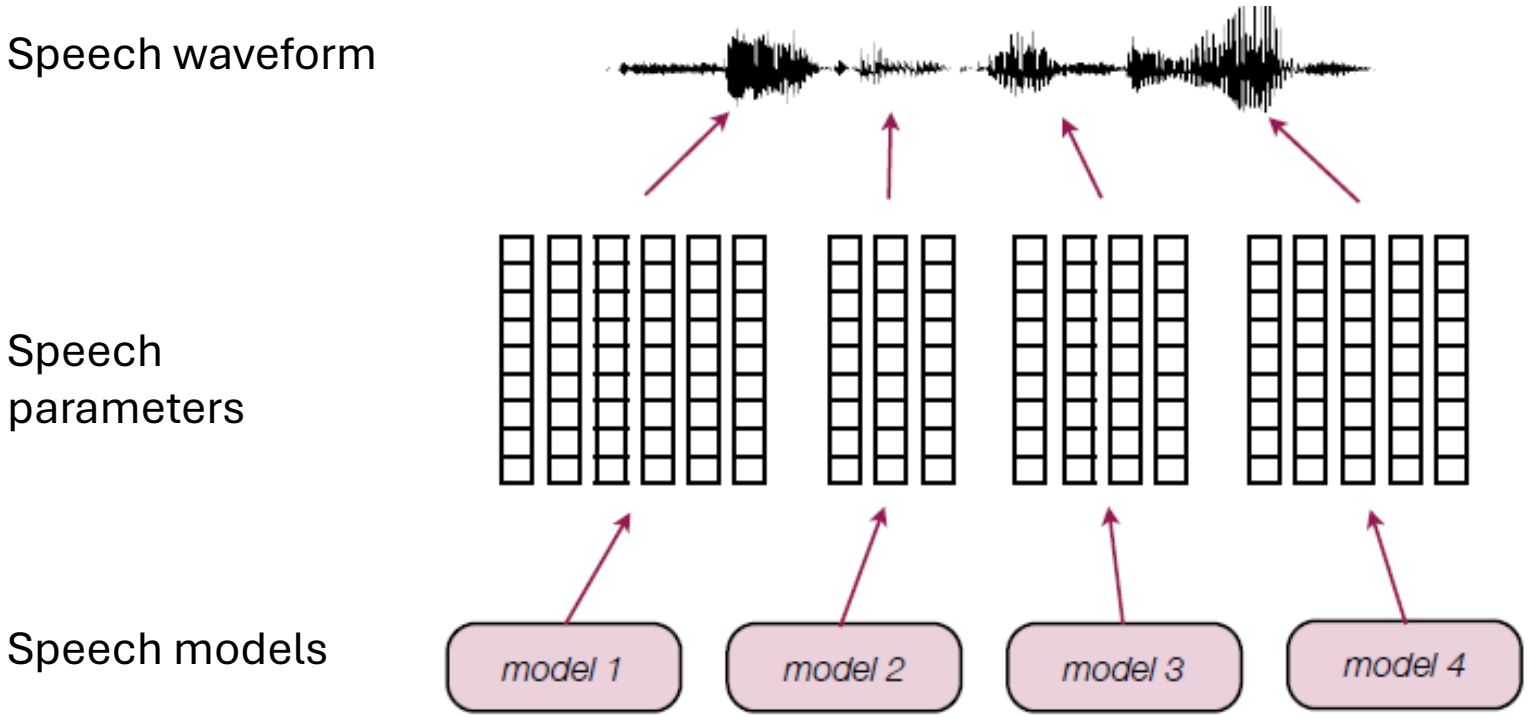
Text-To-Speech Synthesis

- 1994, Statistical Parametric Speech Synthesis:
  - Linguistic-to-acoustic mapping by decision trees



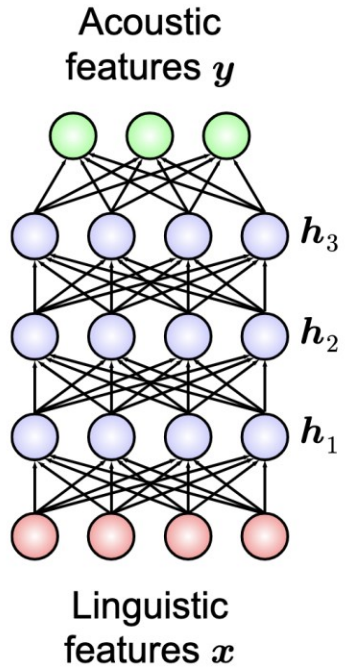
Text-To-Speech Synthesis

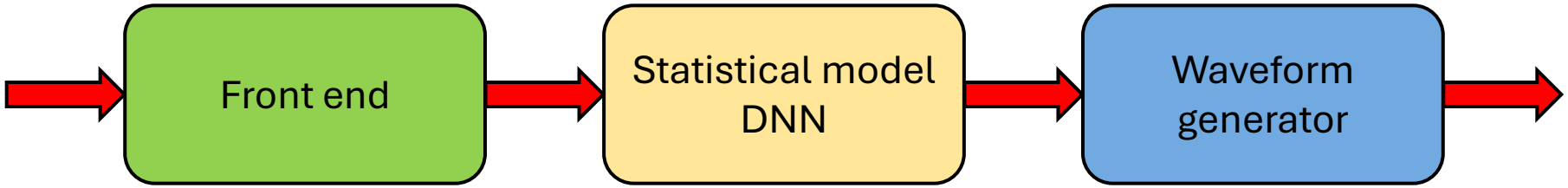
- Given sequence of linguistic labels, each label is mapped, via the decision tree(s) to an HMM model and a corresponding duration.
- Each model in the sequence of models generates a number of speech parameters equal to its duration.
- The speech parameters of the models are concatenated and then are used as input to a vocoder to produce speech.



Text-To-Speech Synthesis

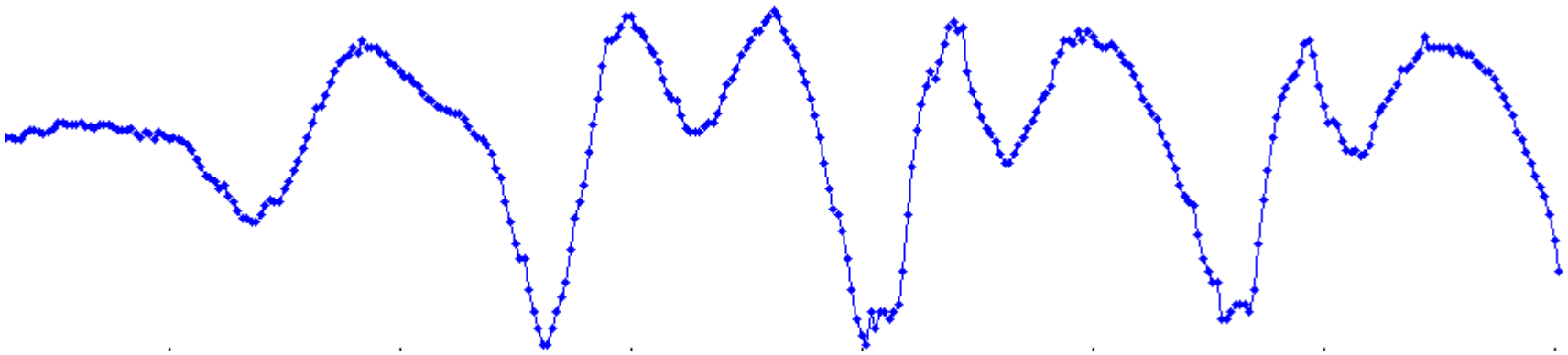
- 2010, DNN-based:
  - The sequencing or alignment task is solved using Hidden Semi-Markov Models.
  - The prediction task is solved by the DNN.
  - No decision tree is needed.
  - All examples are used to train a single model.





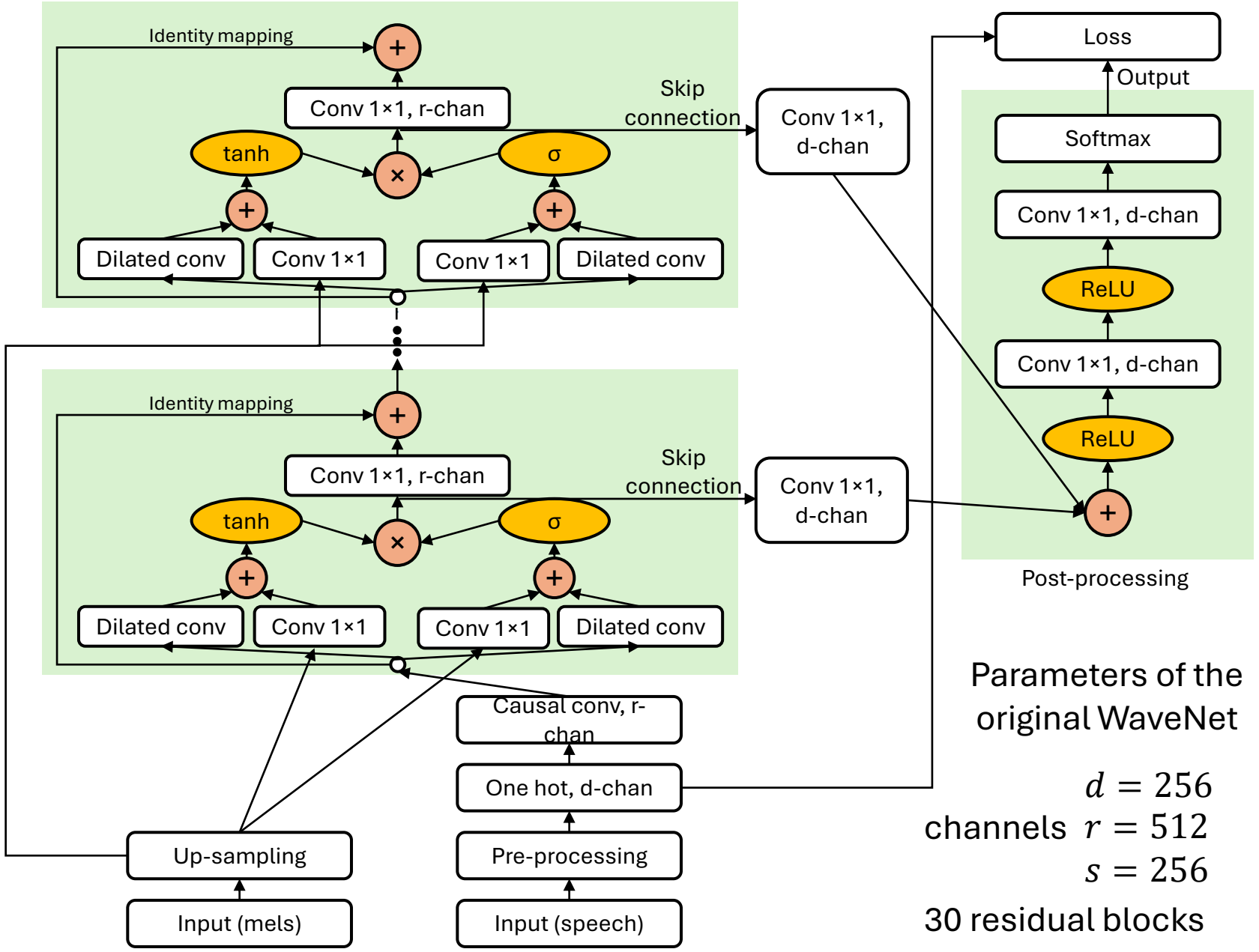
Text-To-Speech Synthesis

- 2016, WaveNet:
  - WaveNet revolutionized SPSS by directly modelling the raw waveform of the audio signal.
  - WaveNet, uses one-dimensional causal convolutional neural networks to represent $P(x_t|x_1, \dots, x_{t-1})$.



Text-To-Speech Synthesis

- 2016, WaveNet:



Parameters of the original WaveNet

$d = 256$

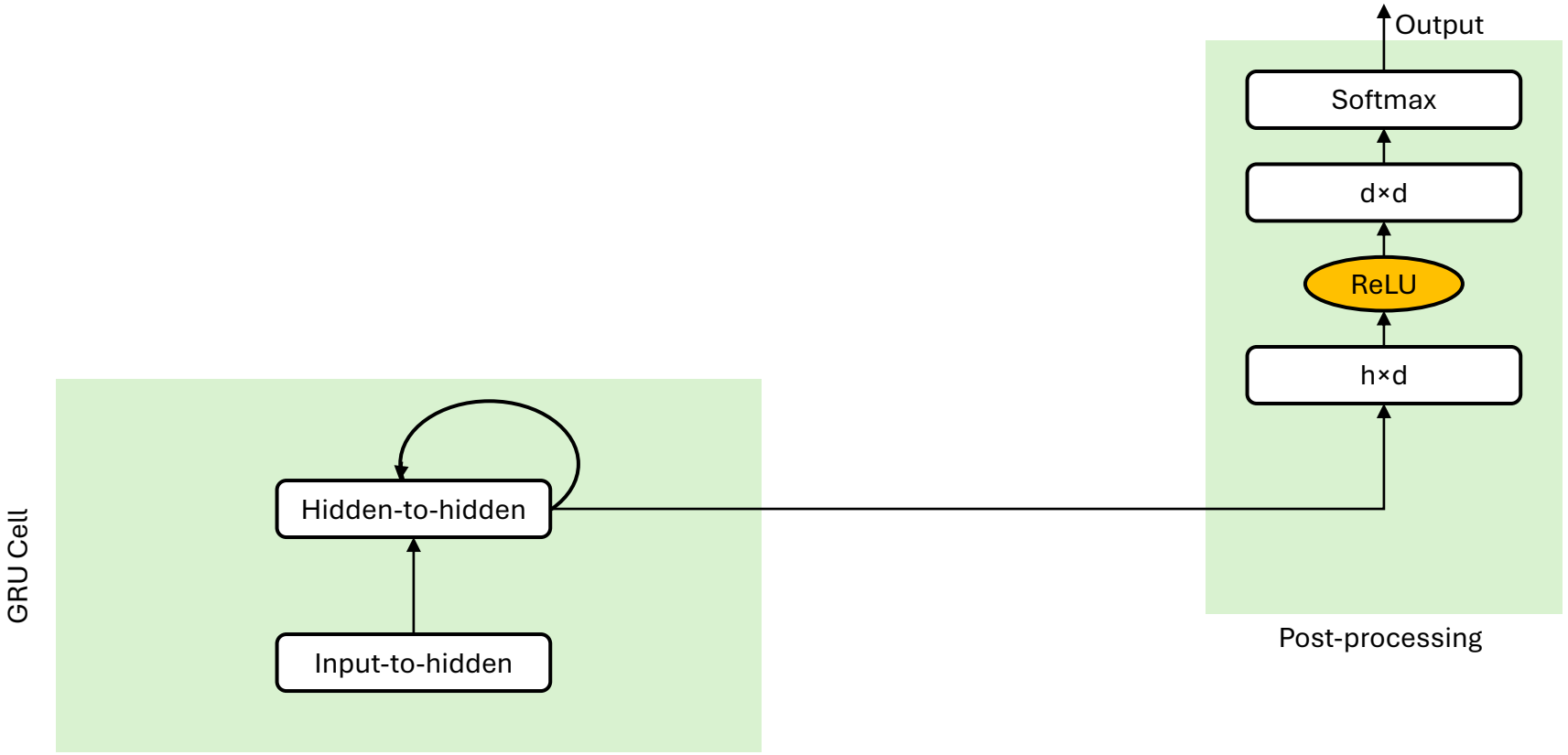
channels $r = 512$

$s = 256$

30 residual blocks

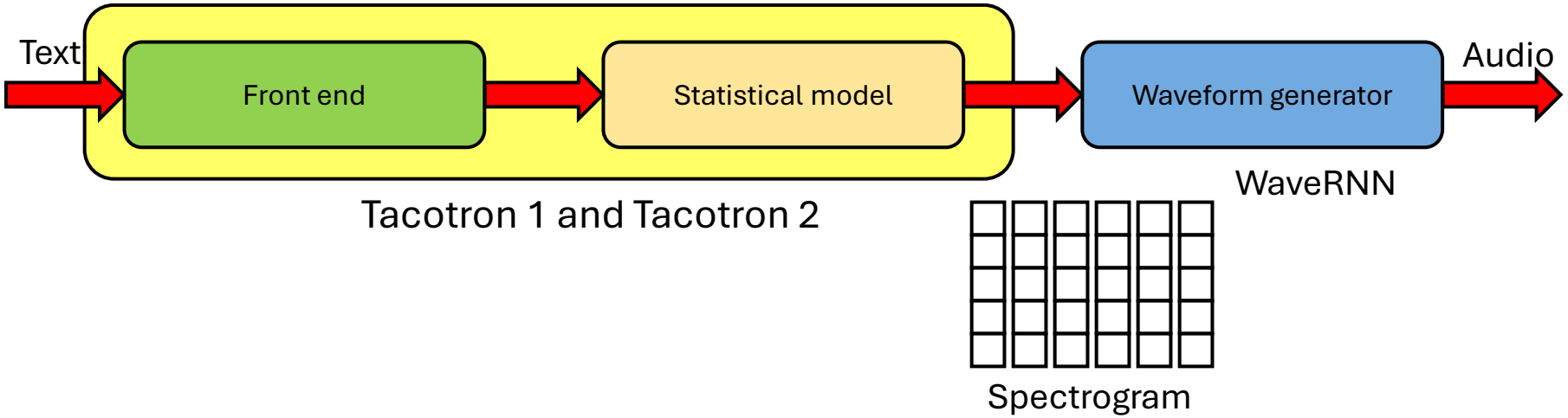
Text-To-Speech Synthesis

- 2018, WaveRNN:
  - To increase the efficiency of sampling without compromising their quality, Kalchbrenner et al., proposed to substitute deep layers of dilated convolutions of WaveNet with a single layer RNNs



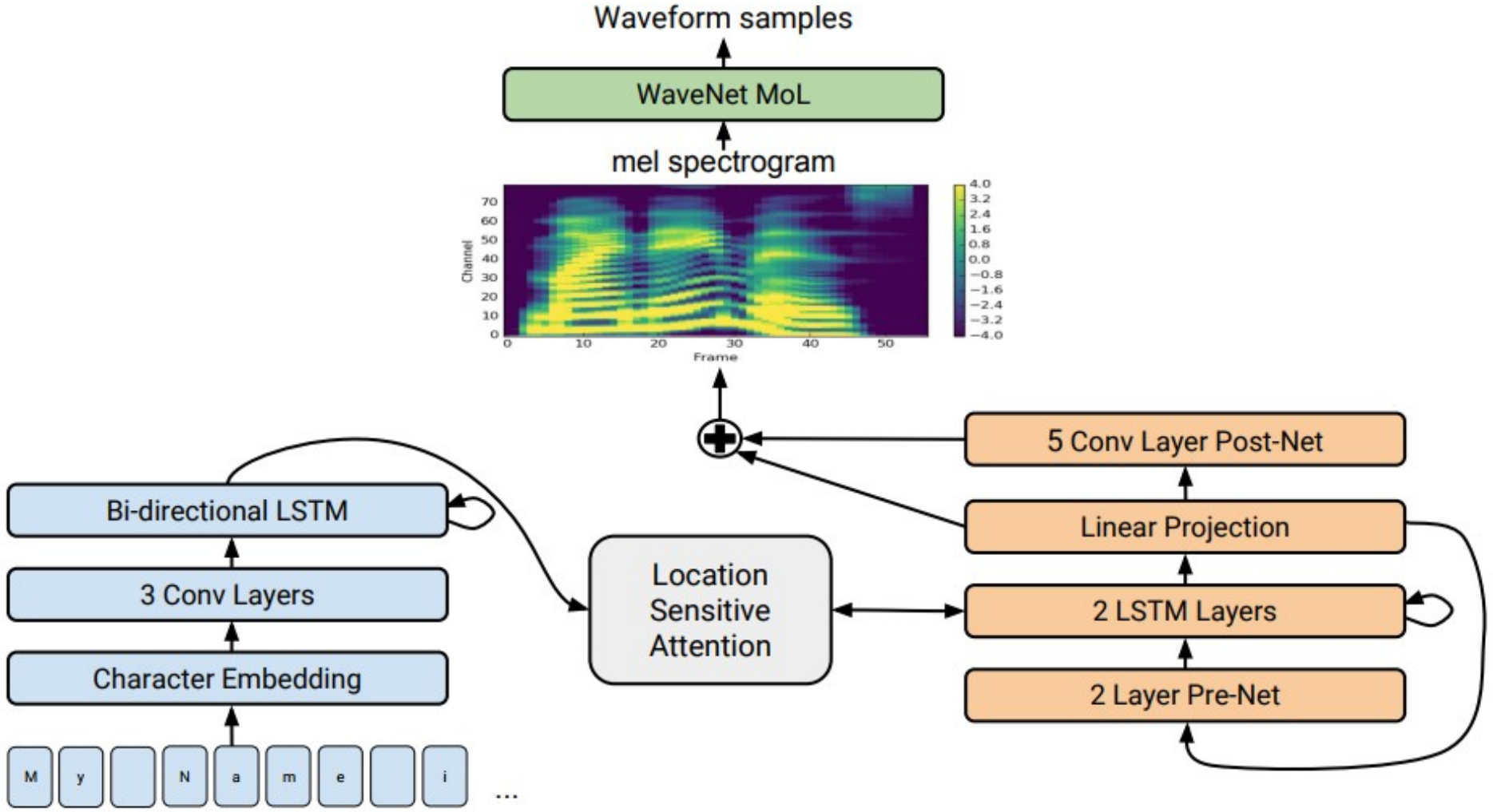
Text-To-Speech Synthesis

- 2017, Sequence-to-sequence modelling – Tacotron & Tacotron 2:



Text-To-Speech Synthesis

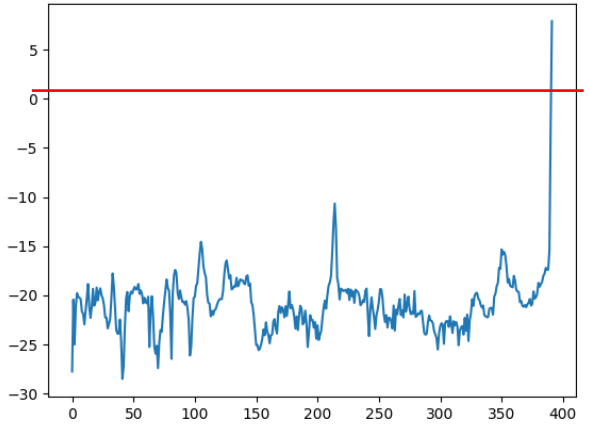
- 2017, 2018, Sequence-to-sequence modelling – Tacotron & Tacotron 2:



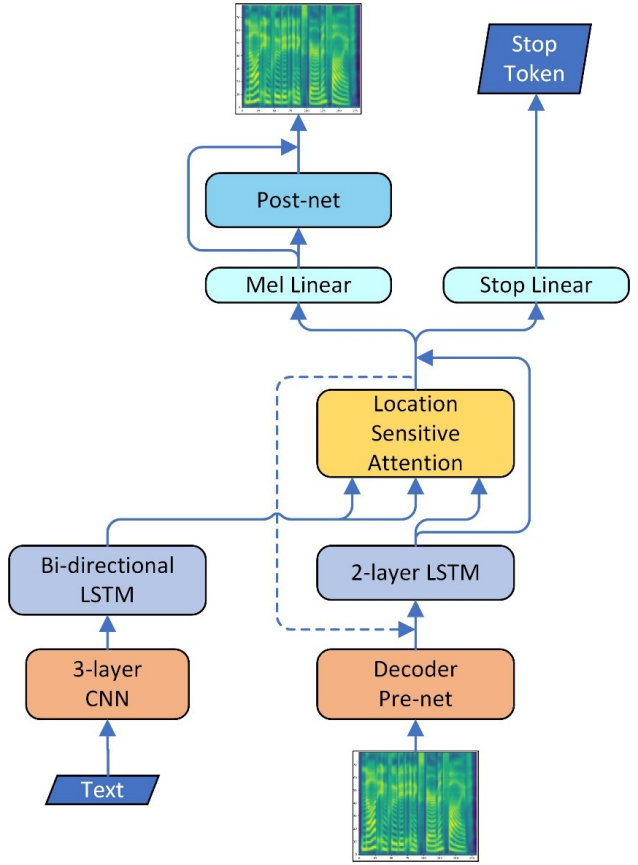
Text-To-Speech Synthesis

- 2017, 2018, Sequence-to-sequence modelling – Tacotron & Tacotron 2:
  - Issues with autoregressive acoustic models
    - Slow inference speed for autoregressive generation.
    - Synthesized speech is usually not robust, with words being skipped and repeated.

Text: Bad speech synthesis

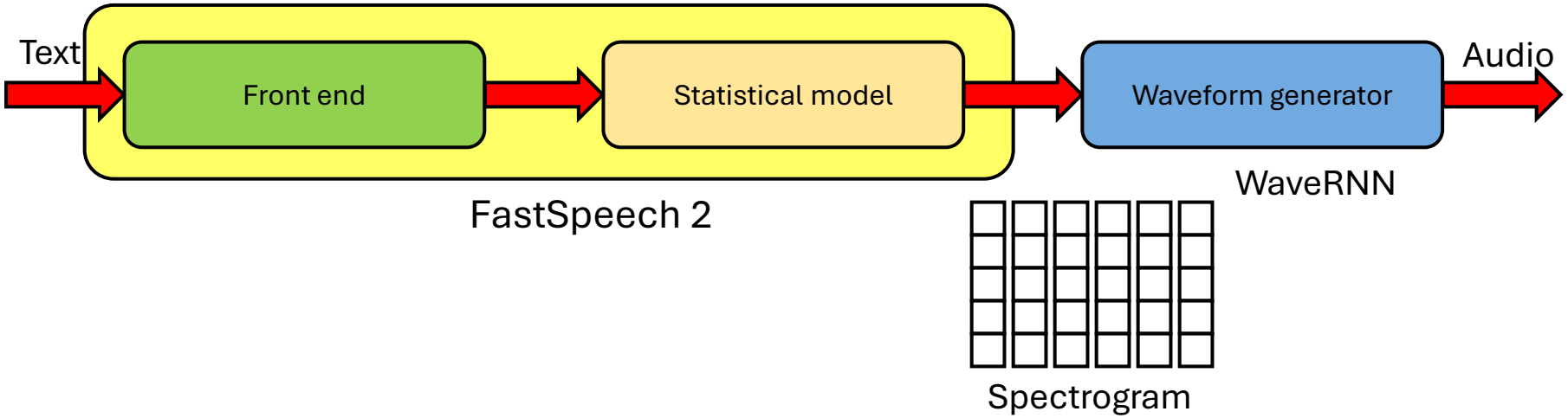


Stop token plot



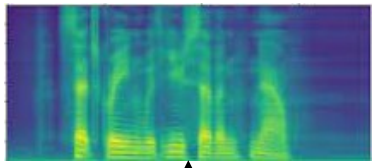
Text-To-Speech Synthesis

- 2020, FastSpeech 2:



Text-To-Speech Synthesis

- FastSpeech 2



Mel-spectrogram Decoder

Positional Encoding

Variance Adaptor

Encoder

Positional Encoding

Phoneme Embedding

Phoneme

LR

Duration

Duration Predictor

Energy

Energy Predictor

Pitch

Pitch Predictor

Linear Layer

LN + Dropout

Conv1D + ReLU

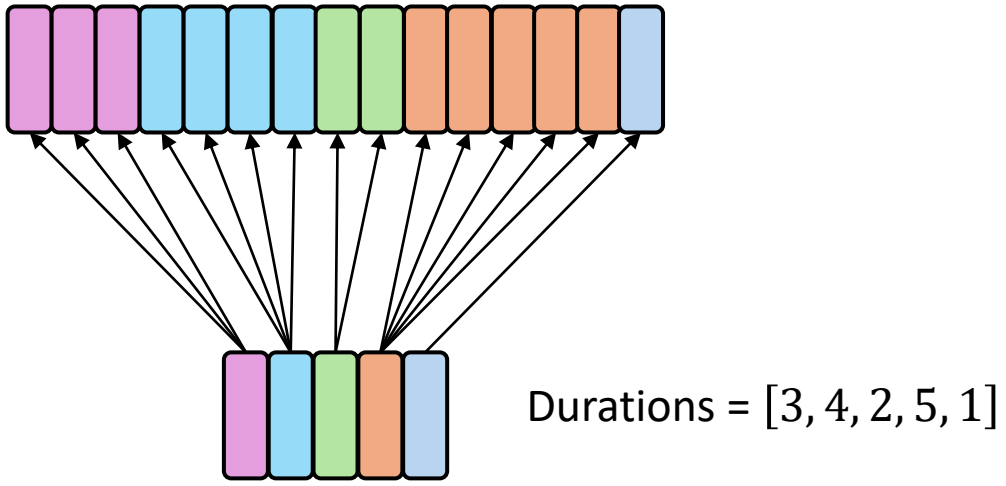
LN + Dropout

Conv1D + ReLU

Duration/pitch/energy predictor

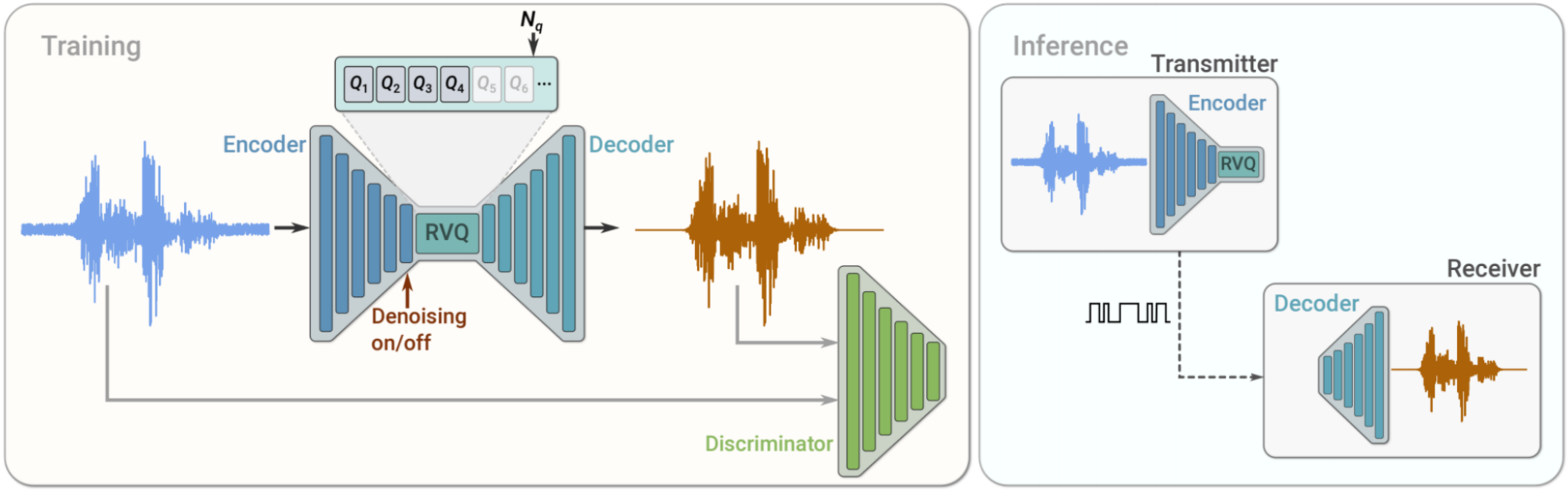
Text-To-Speech Synthesis

- Length Regulator
  - The number of mel-spectrograms that aligns to a phoneme is called phoneme duration.
  - The length regulator expands the hidden sequence of phonemes according to the duration to match the length of a mel-spectrogram sequence.



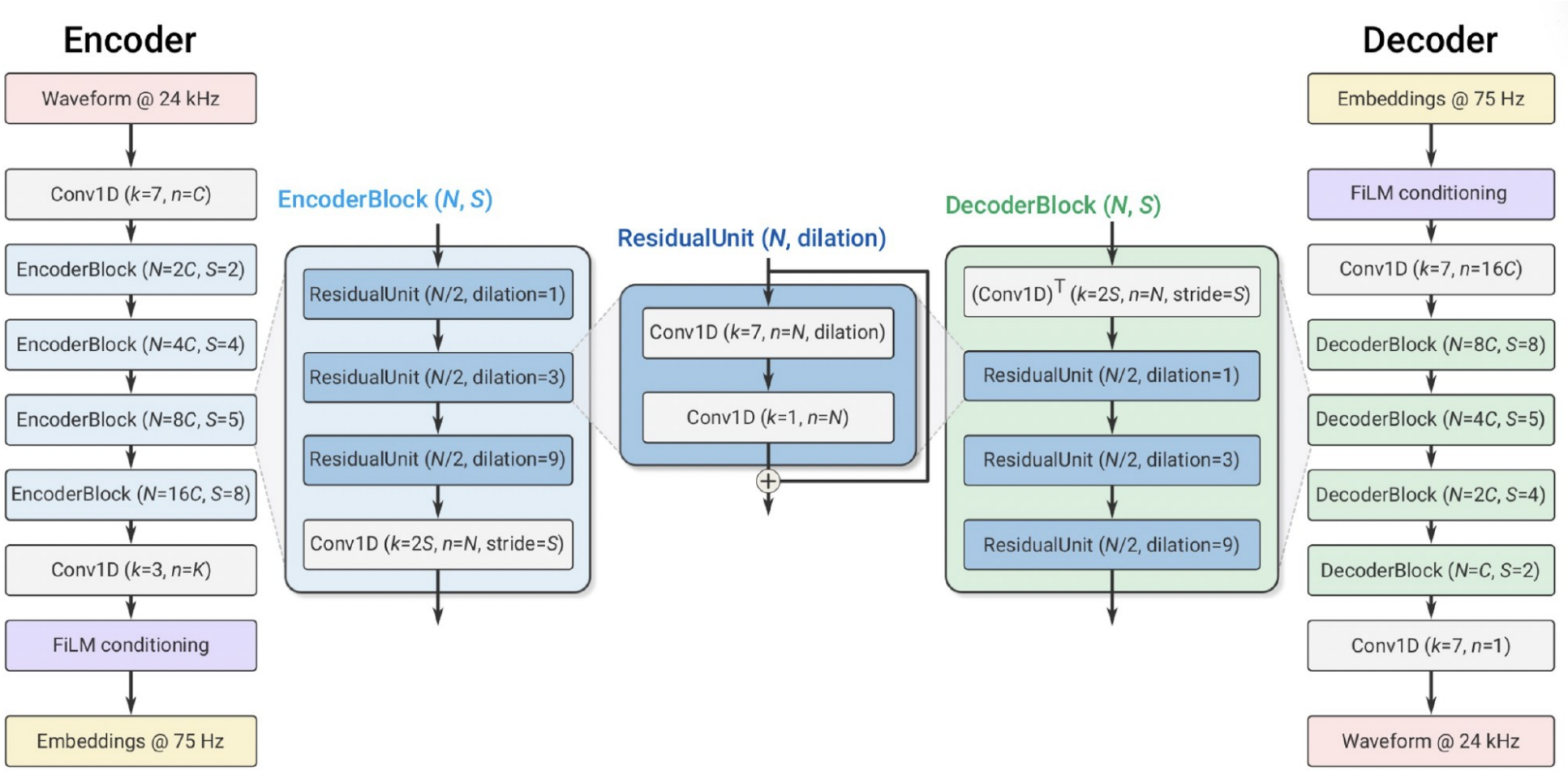
Text-To-Speech Synthesis

- Discrete audio representation:
  - 2021, SoundStream



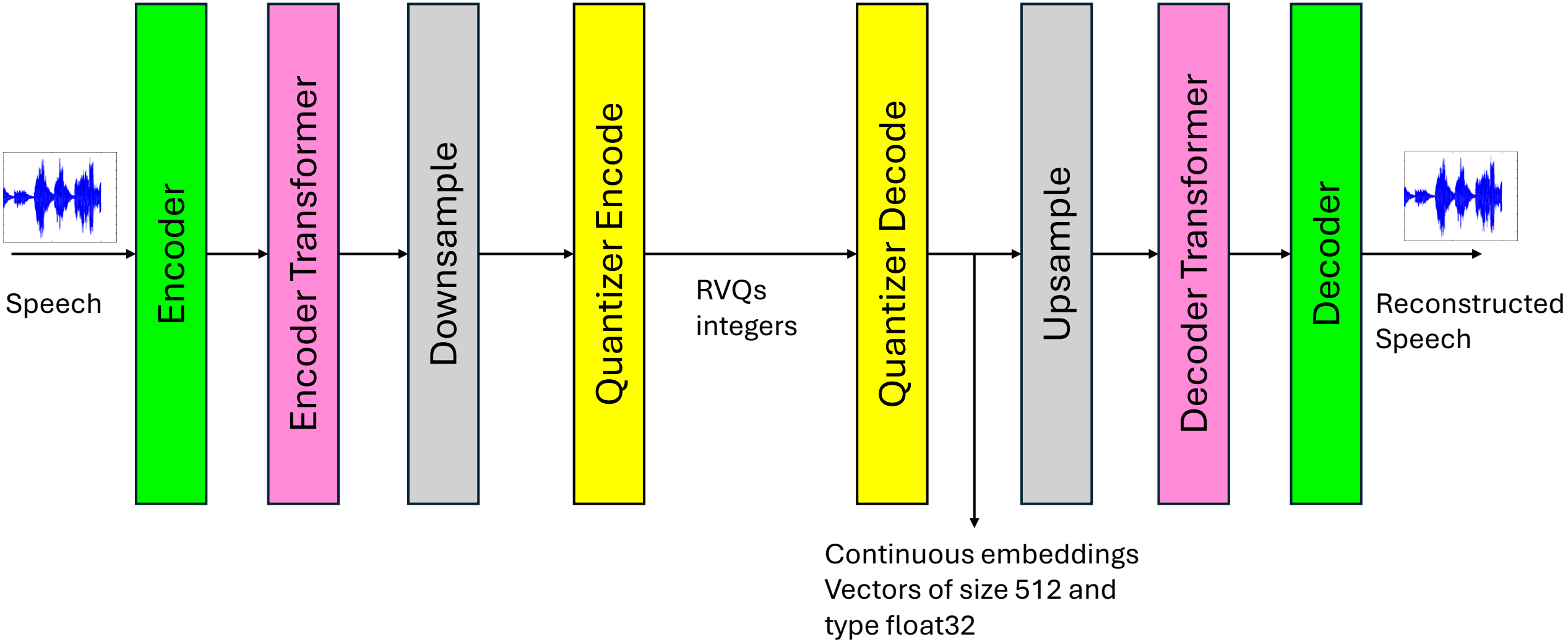
SoundStream Architecture

- Causal convolutions
- Double the number of feature maps every time we stride
- Symmetric encoder and decoder



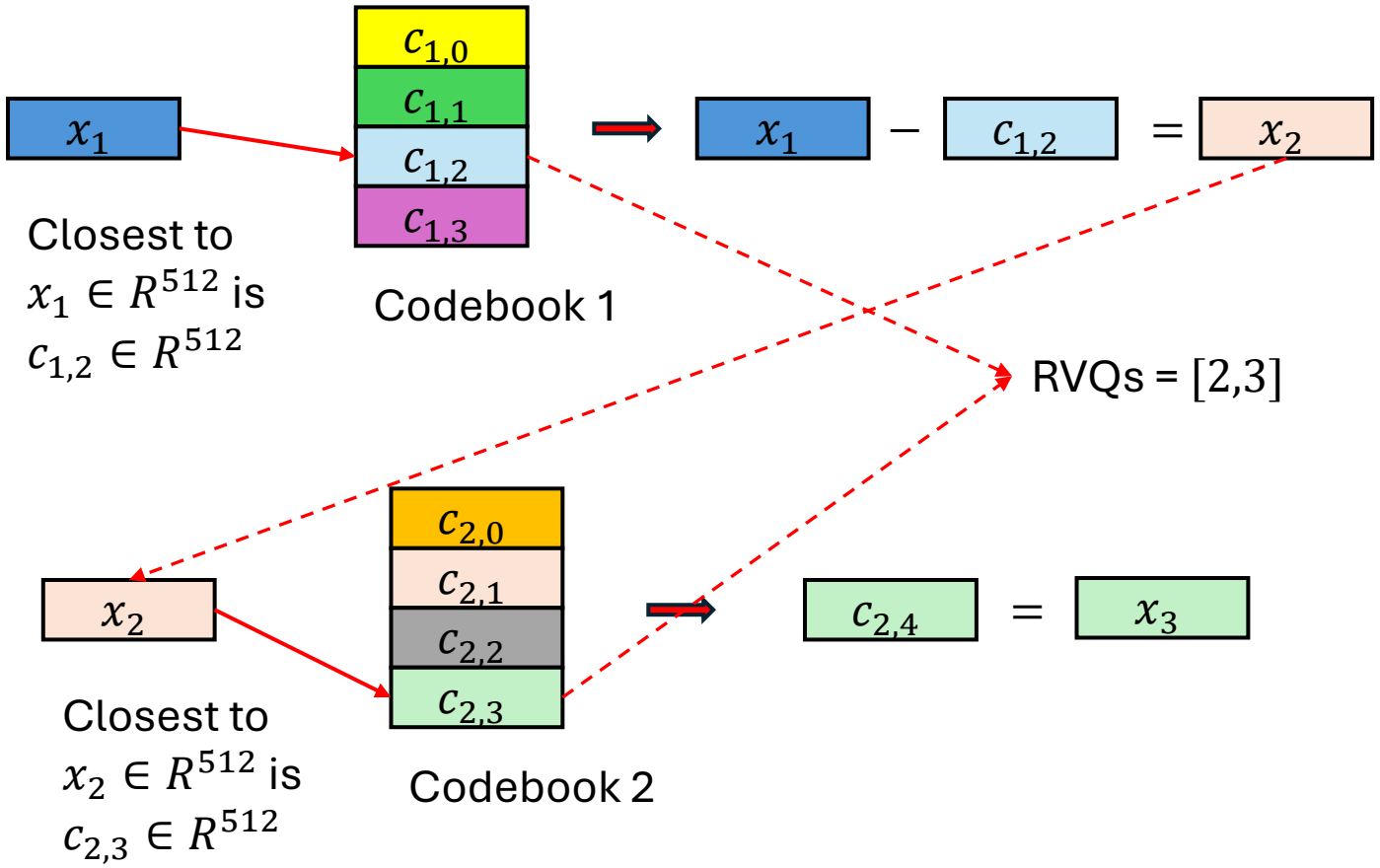
Text-To-Speech Synthesis

- Discrete audio representation:
  - 2024, Mimi



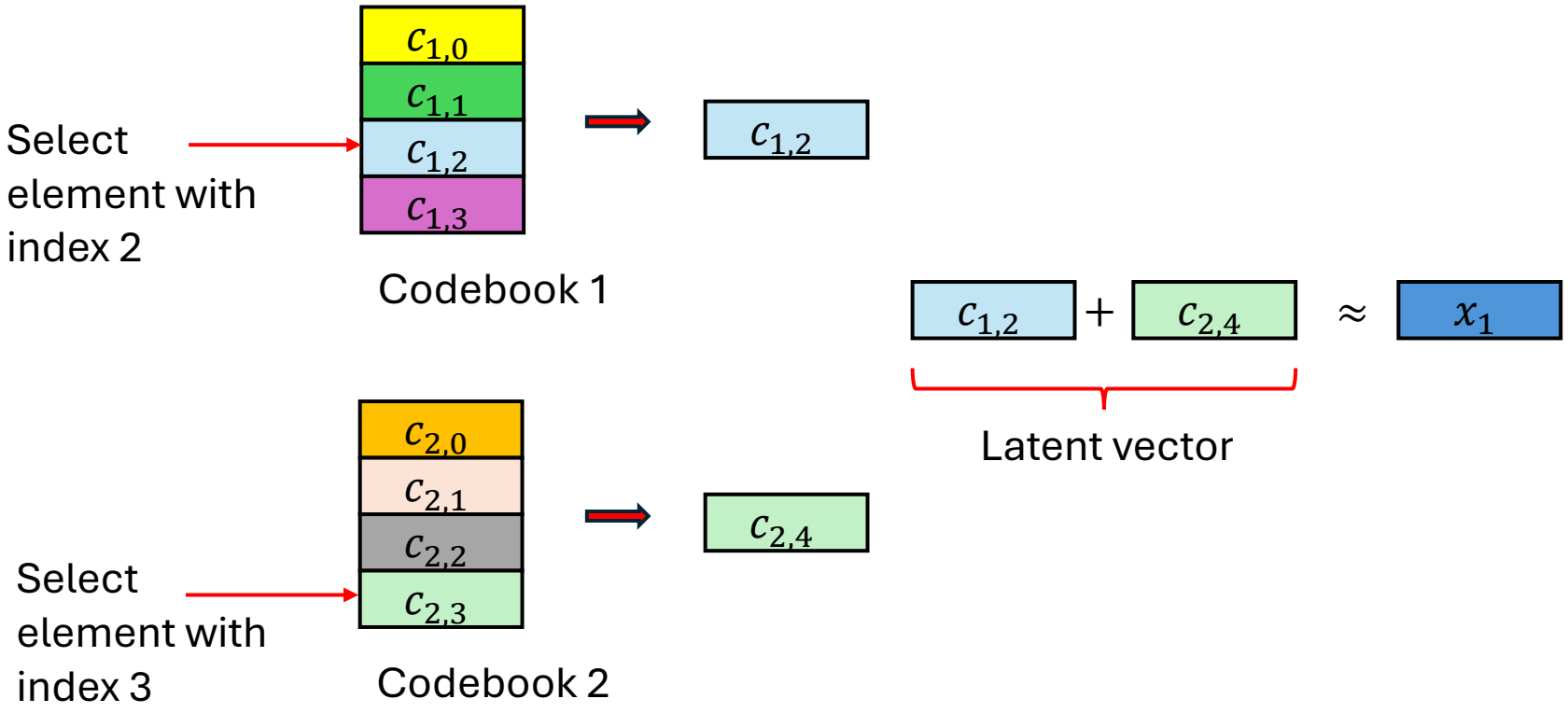
Residual Vector Quantization - Encode

- Example with two codebooks, each of 4 elements, and dimension 512



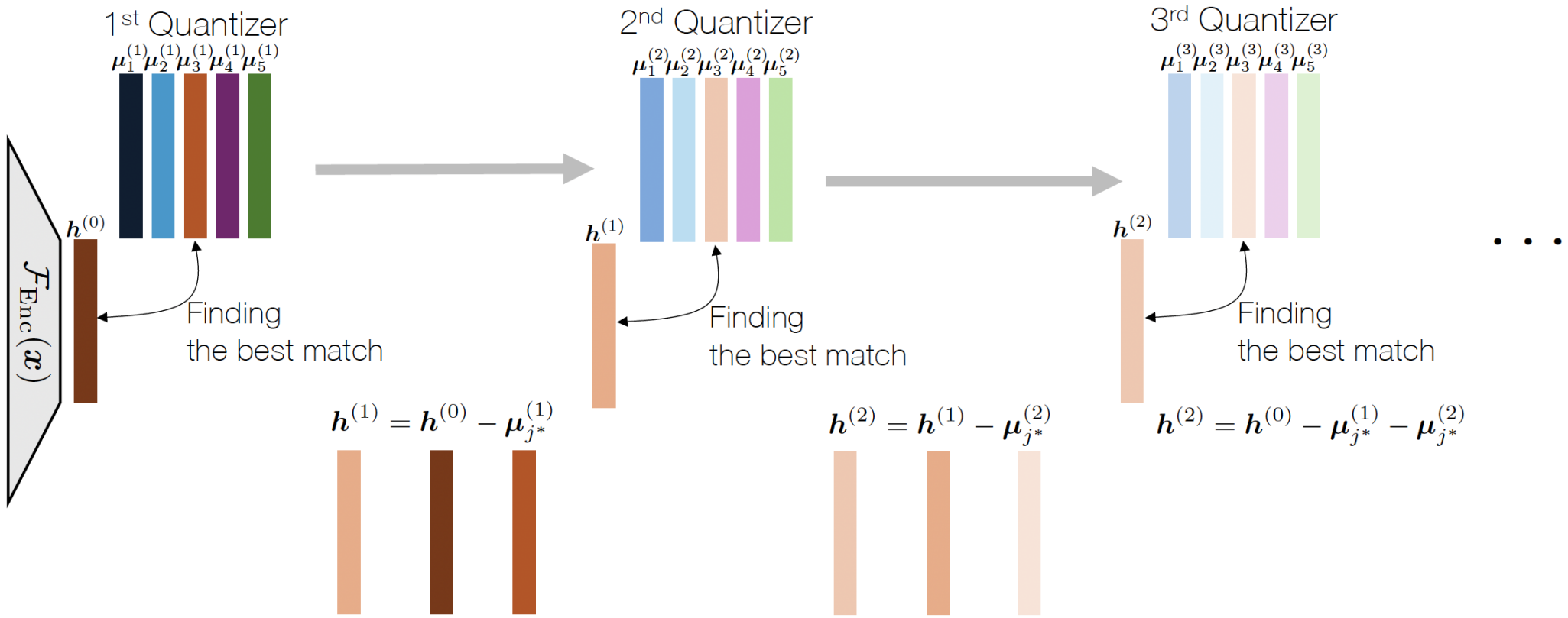
Residual Vector Quantization - Decode

- Example with two codebooks, each of 4 elements, and dimension 512
- Decoding RVQs = [2,3]



Residual Vector Quantization

- More general example
- $h \approx \sum_{i-1}^{N_q} \mu_{j^*}^{(i)}$



Residual Vector Quantization (RVQ)

Algorithm 1: Residual Vector Quantization

Input: $y = \text{enc}(x)$ the output of the encoder, vector quantizers Q_i for $i = 1..N_q$

Output: the quantized $\hat{y}$

$\hat{y} \leftarrow 0.0$

$\text{residual} \leftarrow y$

for $i = 1$ *to* N_q **do**

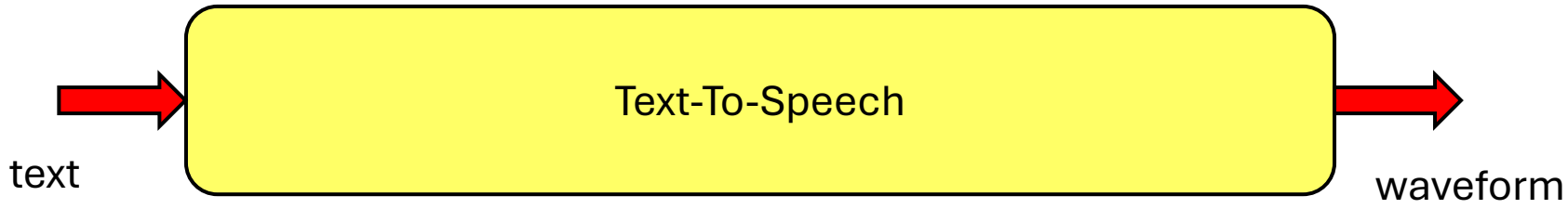
  $\hat{y} \mathrel{+}= Q_i(\text{residual})$

  $\text{residual} \mathrel{-}= Q_i(\text{residual})$

return $\hat{y}$

Text-To-Speech Synthesis

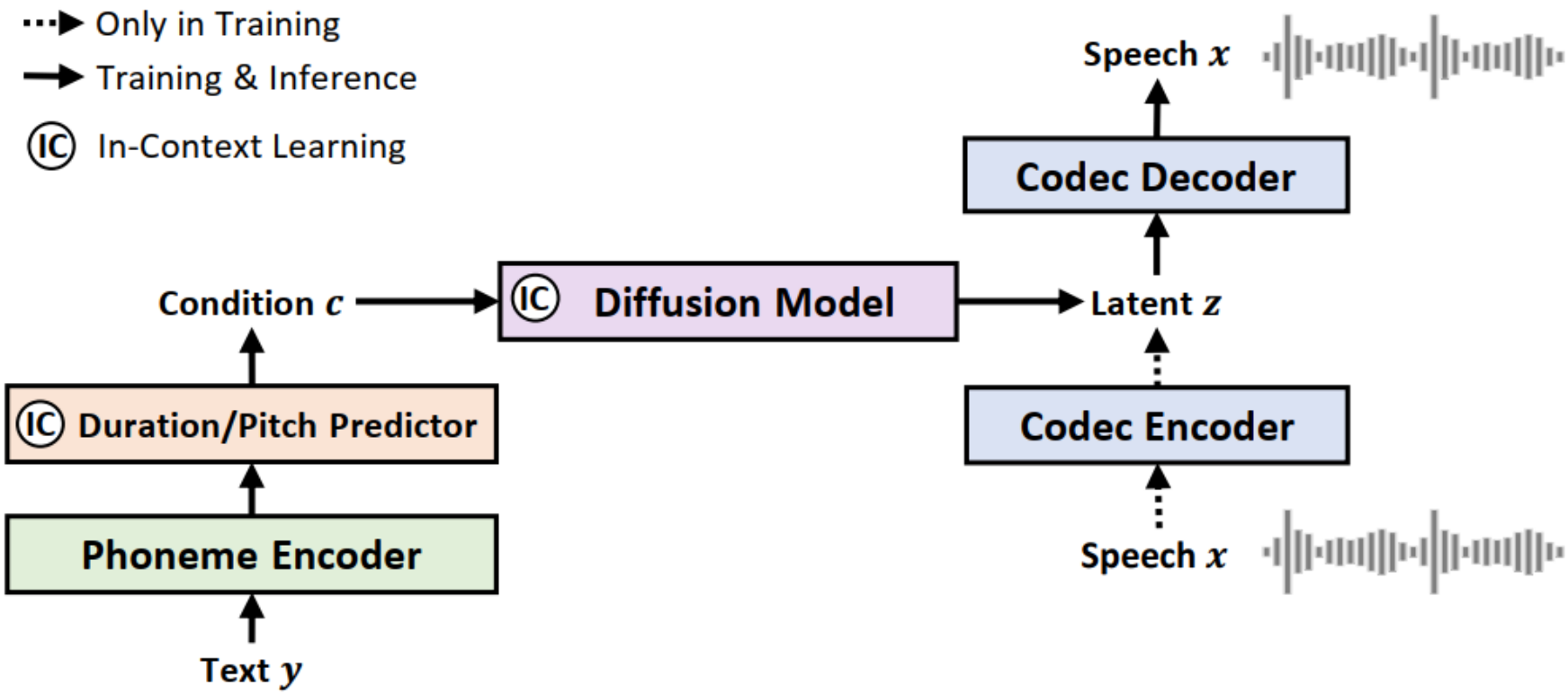
- End-to-End TTS



There was a
change now

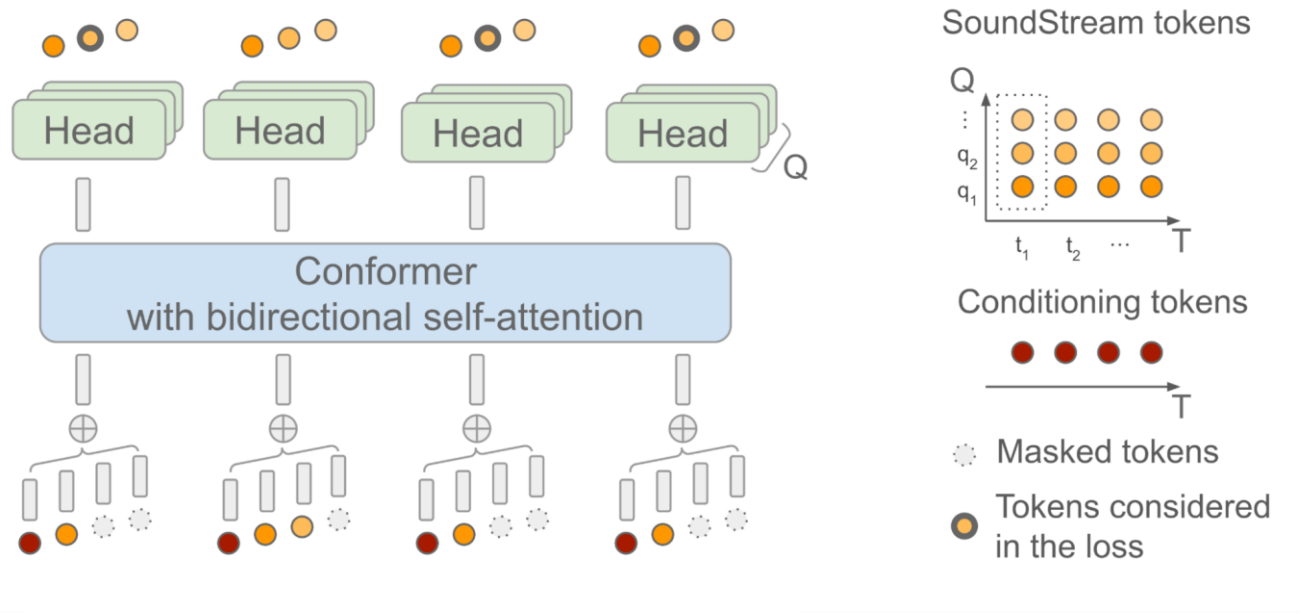
Text-To-Speech Synthesis

- 2023, audio tokens in speech synthesis:
  - NaturalSpeech 2



Text-To-Speech Synthesis

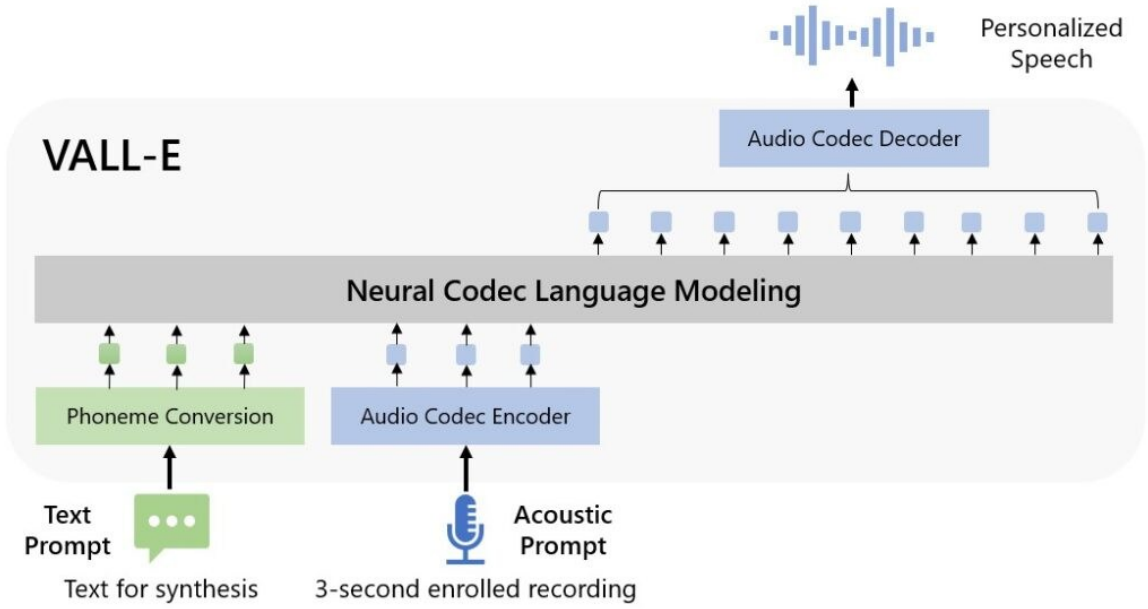
- 2023, audio tokens in speech synthesis:
  - SoundStorm



SoundStorm:
Efficient Parallel Audio Generation

Text-To-Speech Synthesis

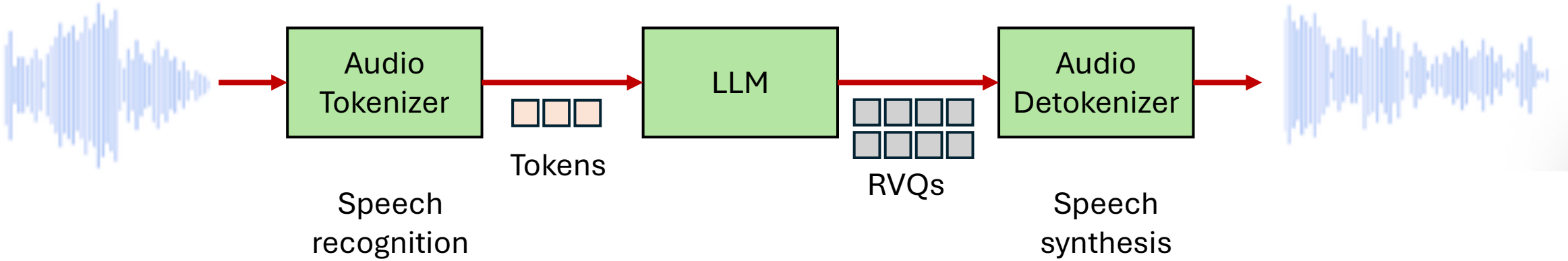
- 2023, audio tokens in speech synthesis:
  - VALL-E & VALL-E 2

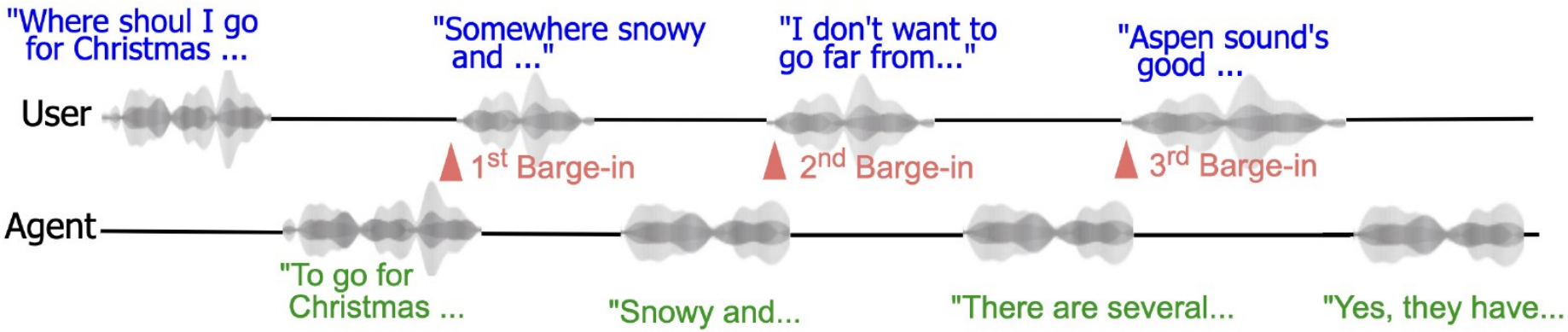


VALL-E

VALL-E 2

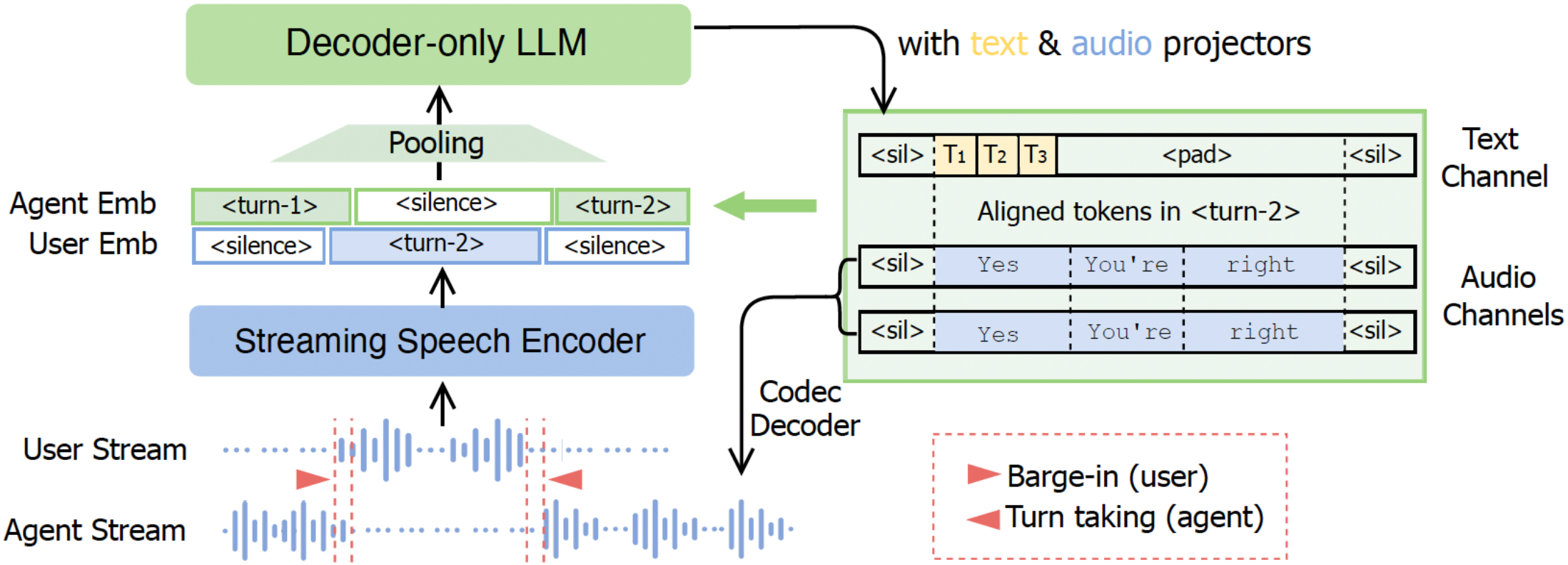
Dialog Systems





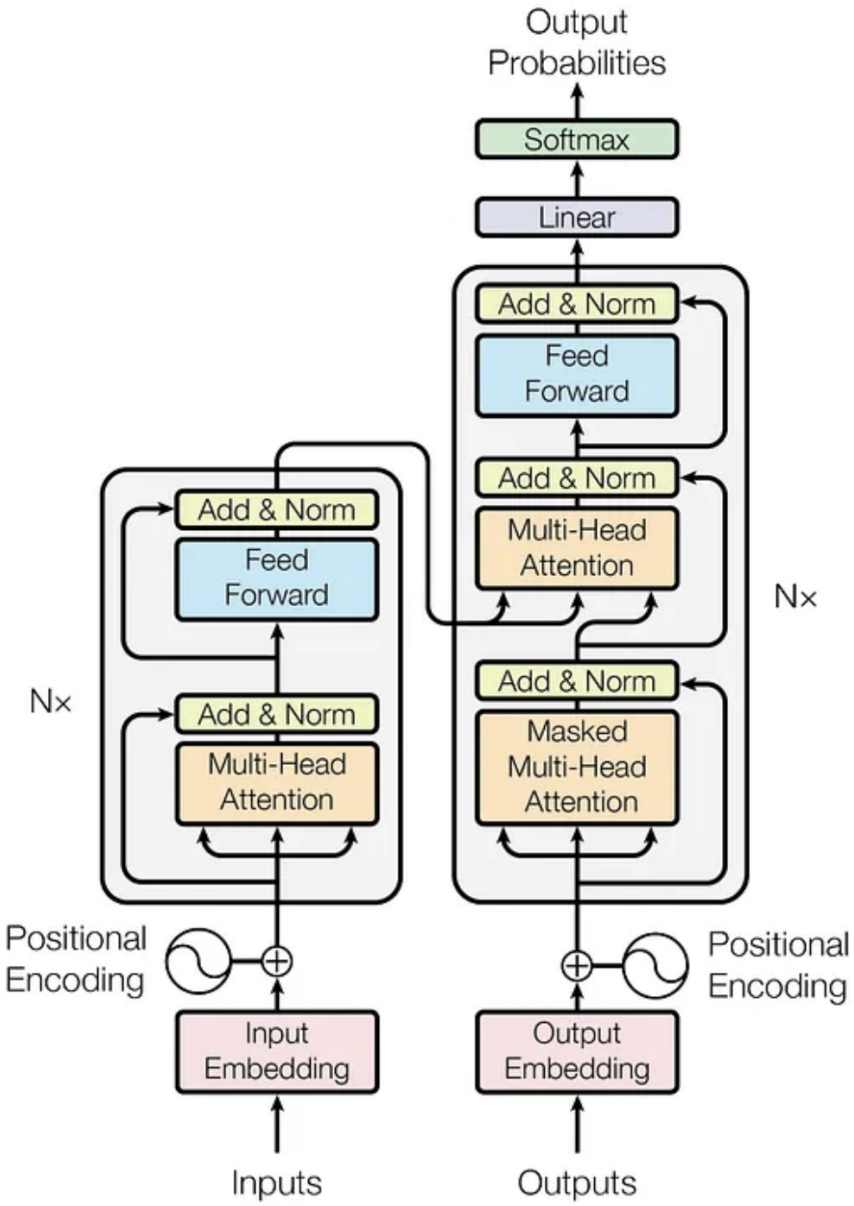
Full Duplex Dialog Systems

- 2024, Moshi
- 2025, SALM-Duplex



LLM

- Encoder only
  - Classification
- Decoder only
  - Text generation,
- Encoder-Decoder
  - Machine translation



References

1. Alan W Black, Heiga Ze, Keiichi Tokuda, Statistical parametric speech synthesis, Speech Communication 2009
2. van den Oord, Aaron; Dieleman, Sander; Zen, Heiga; Simonyan, Karen; Vinyals, Oriol; Graves, Alex; Kalchbrenner, Nal; Senior, Andrew; Kavukcuoglu, Koray, WaveNet: A Generative Model for Raw Audio, arXiv:1609.03499
3. Le Paine, Tom; Khorrami, Pooya; Chang, Shiyu; Zhang, Yang; Ramachandran, Prajit; Hasegawa-Johnson, Mark A.; Huang, Thomas S., Fast Wavenet Generation Algorithm, eprint arXiv:1611.09482
4. Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, Koray Kavukcuoglu , Efficient Neural Audio Synthesis, PMLR 80:2410-2419, 2018.
5. Yuxuan Wang and RJ Skerry-Ryan and Daisy Stanton and Yonghui Wu and Ron J. Weiss and Navdeep Jaitly and Zongheng Yang and Ying Xiao and Zhifeng Chen and Samy Bengio and Quoc Le and Yannis Agiomyrgiannakis and Rob Clark and Rif A. Saurous, Tacotron: Towards End-to-End Speech Synthesis,, Interspeech 2017.
6. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu, Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, ICASSP 2017.
7. Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, FastSpeech: Fast, Robust and Controllable Text to Speech, arXiv:1905.09263, 2019

References

8. Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, arXiv:2006.04558, 2020
9. Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, Marco Tagliasacchi, SoundStream: An End-to-End Neural Audio Codec, 2021
10. Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, Marco Tagliasacchi, SoundStorm: Efficient Parallel Audio Generation, 2023
11. Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, sheng zhao, Jiang Bian , NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers, ICLR 2024
12. Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, Furu Wei, Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, IEEE Transactions on Audio, Speech, and Language Processing, 2023
13. Sanyuan Chen, Shujie LIU, Long Zhou, Eric Liu, Xu Tan, Jinyu Li, sheng zhao, Yao Qian, Furu Wei, VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers, ICLR 2025

References

14. Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, Neil Zeghidour, Moshi: a speech-text foundation model for real-time dialogue, ArXiv, 2024
15. Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, Boris Ginsburg, Efficient and Direct Duplex Modeling for Speech-to-Speech Language Model, Interspeech 2025
16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, NIPS 2017
17. Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, Ohad Shamir, Proving the Lottery Ticket Hypothesis: Pruning is All You Need, ICML 2020
18. Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, PNAS 2019