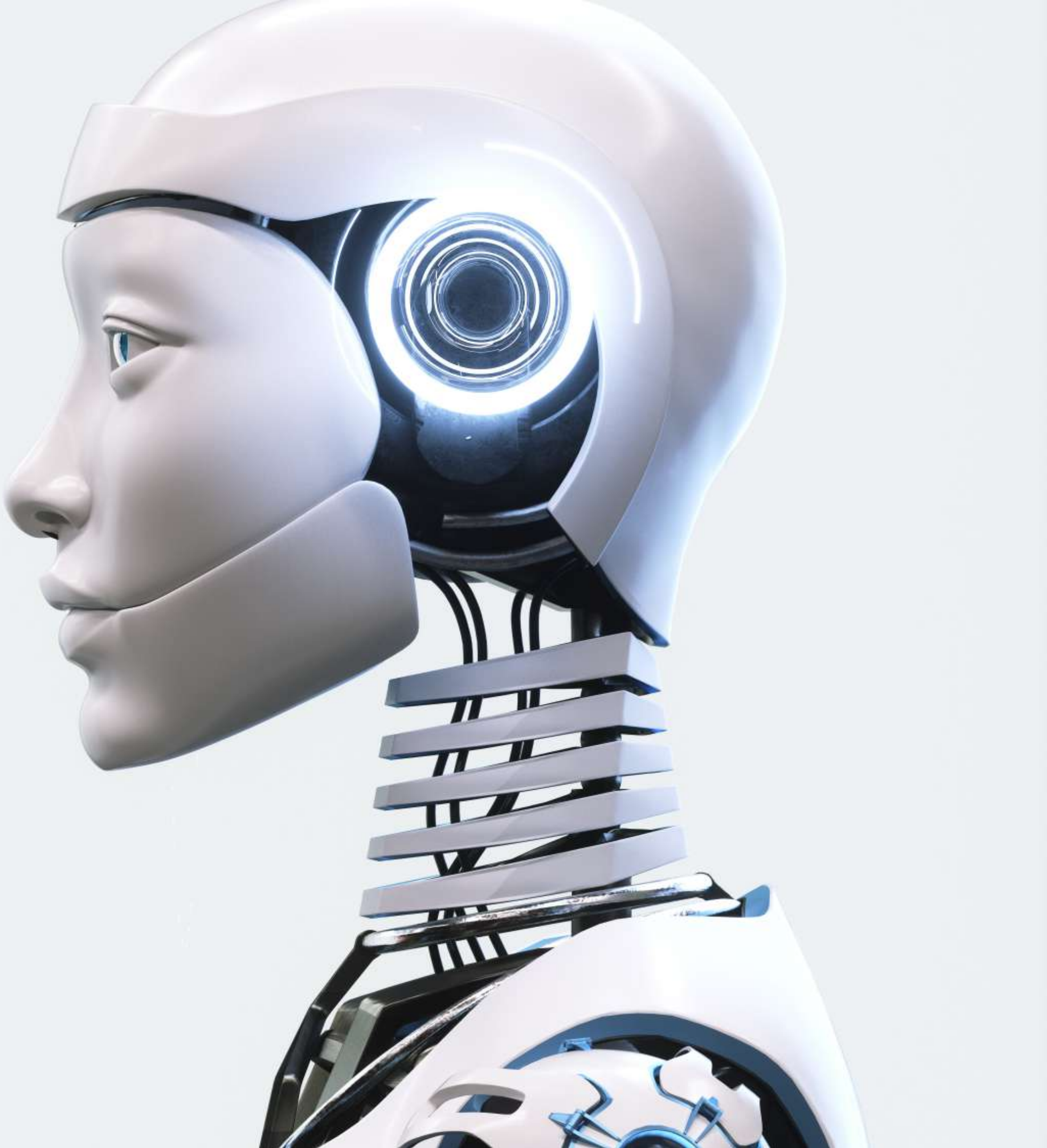
Speech Synthesis

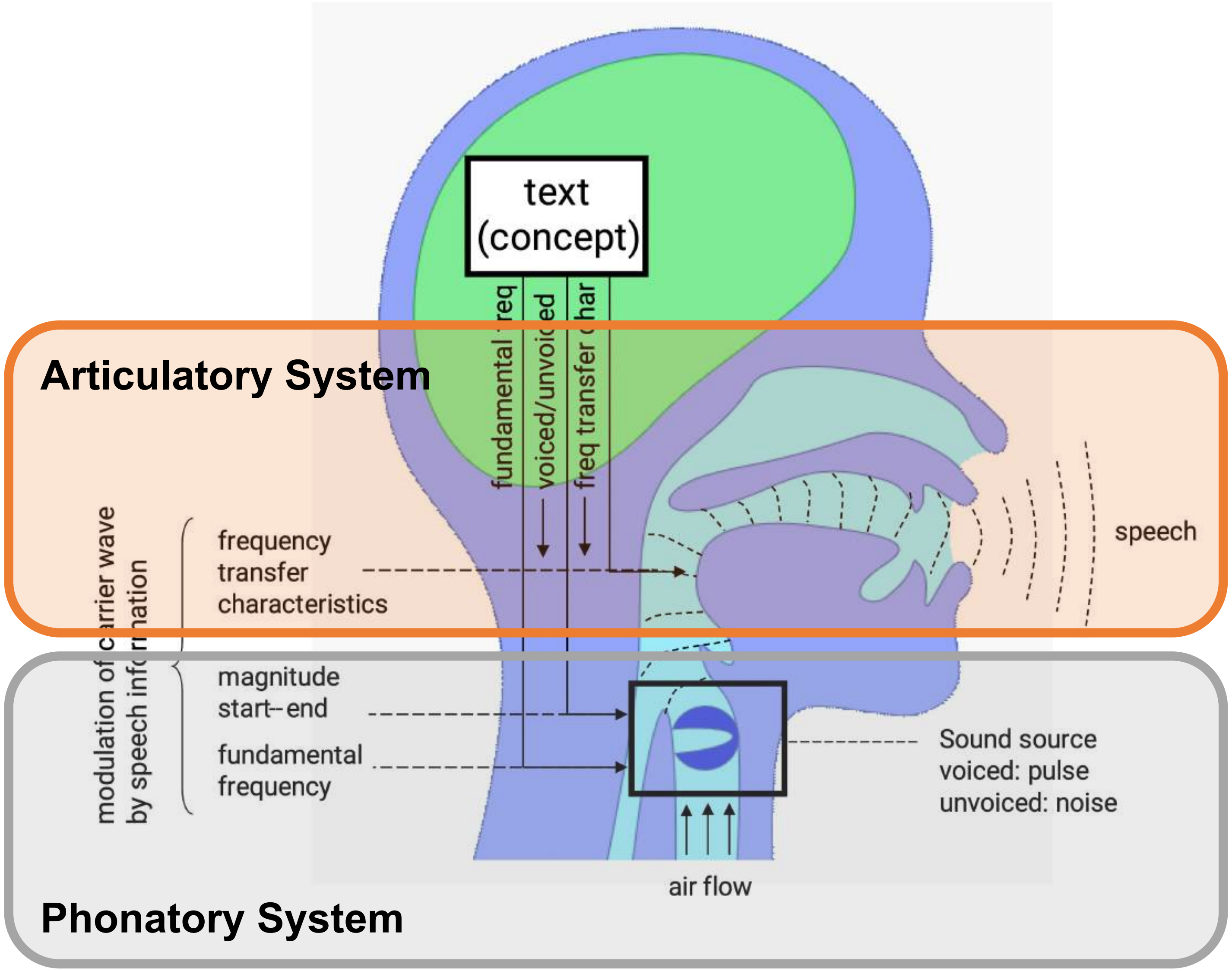
Luis Caldas de Oliveira

Part I: Fundamental Concepts



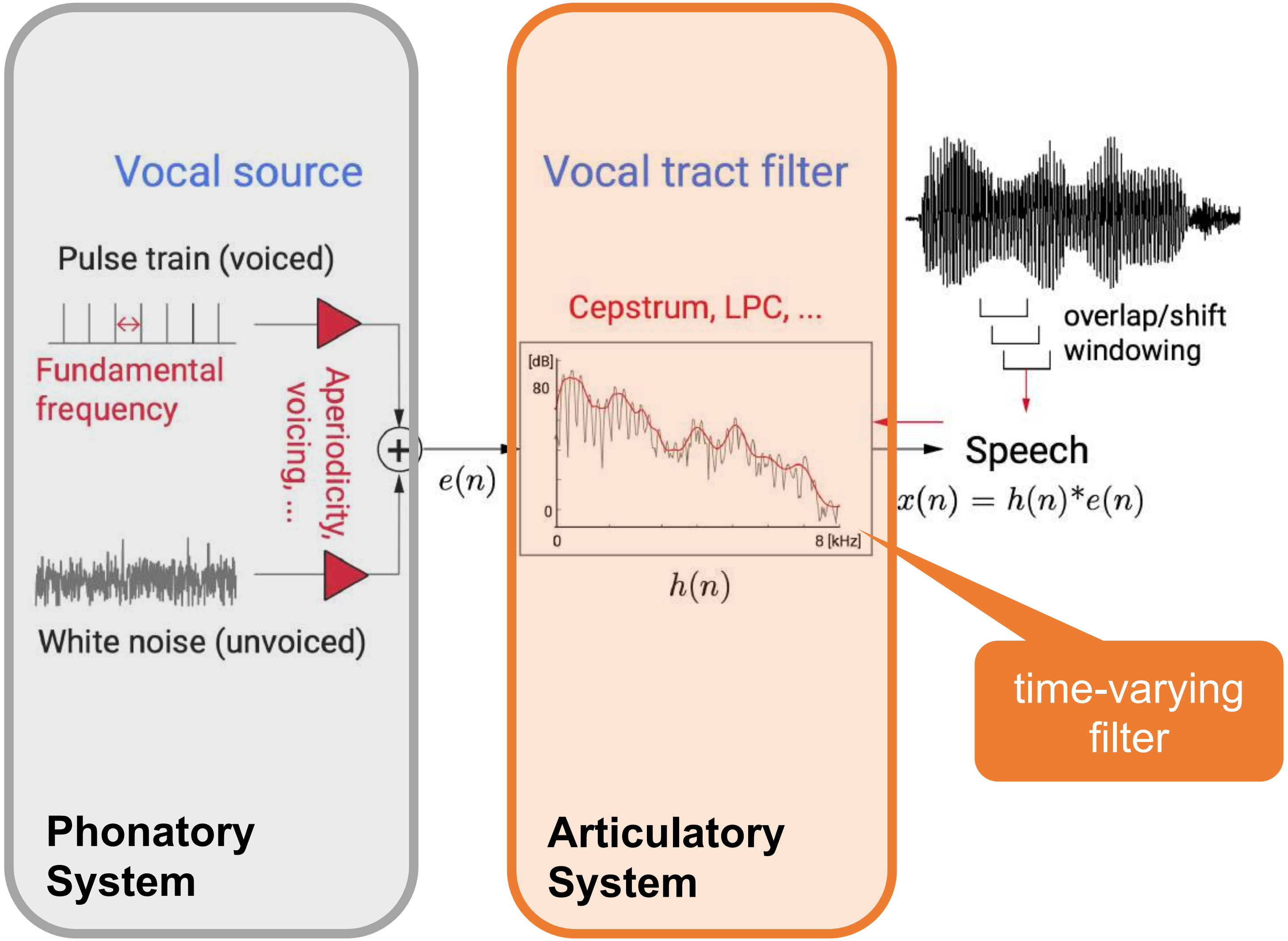
Introduction to Speech Synthesis

Human Speech Production



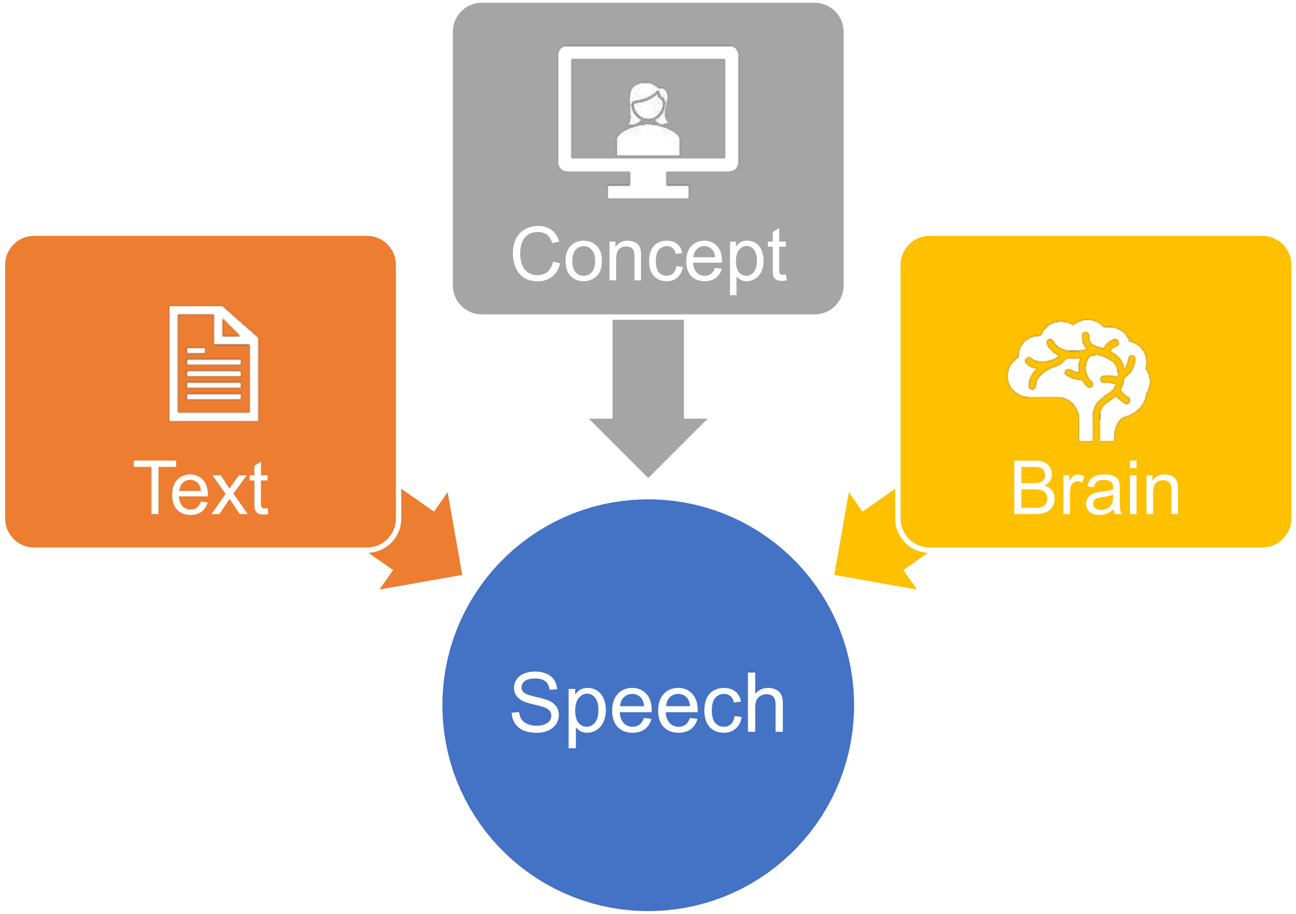
Heiga Zen (2017)

Human Speech Production Model



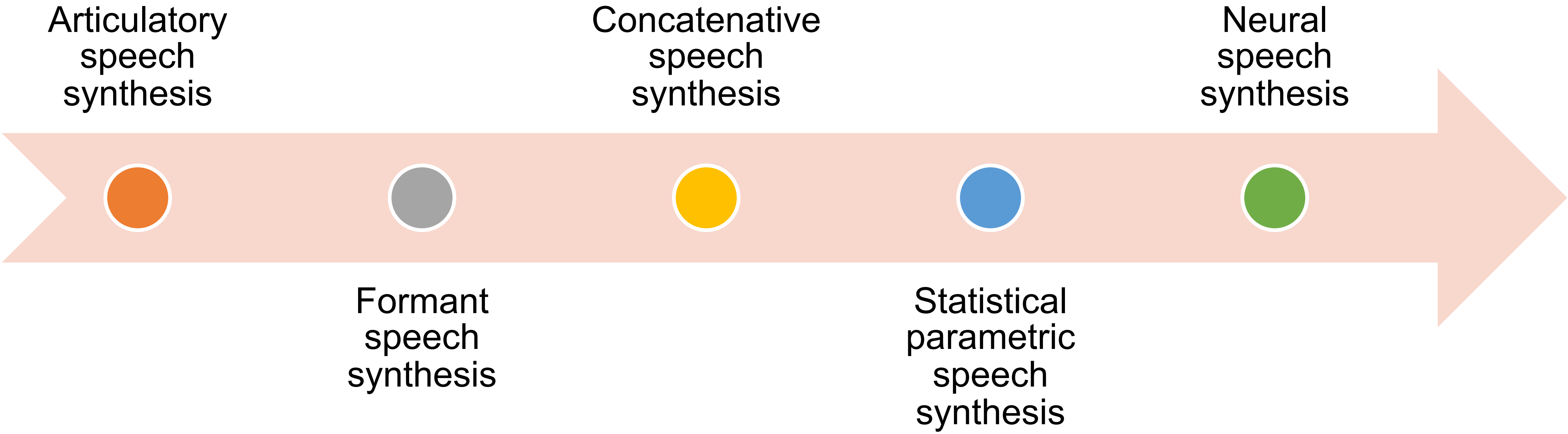
Heiga Zen (2017)

Speech Synthesis



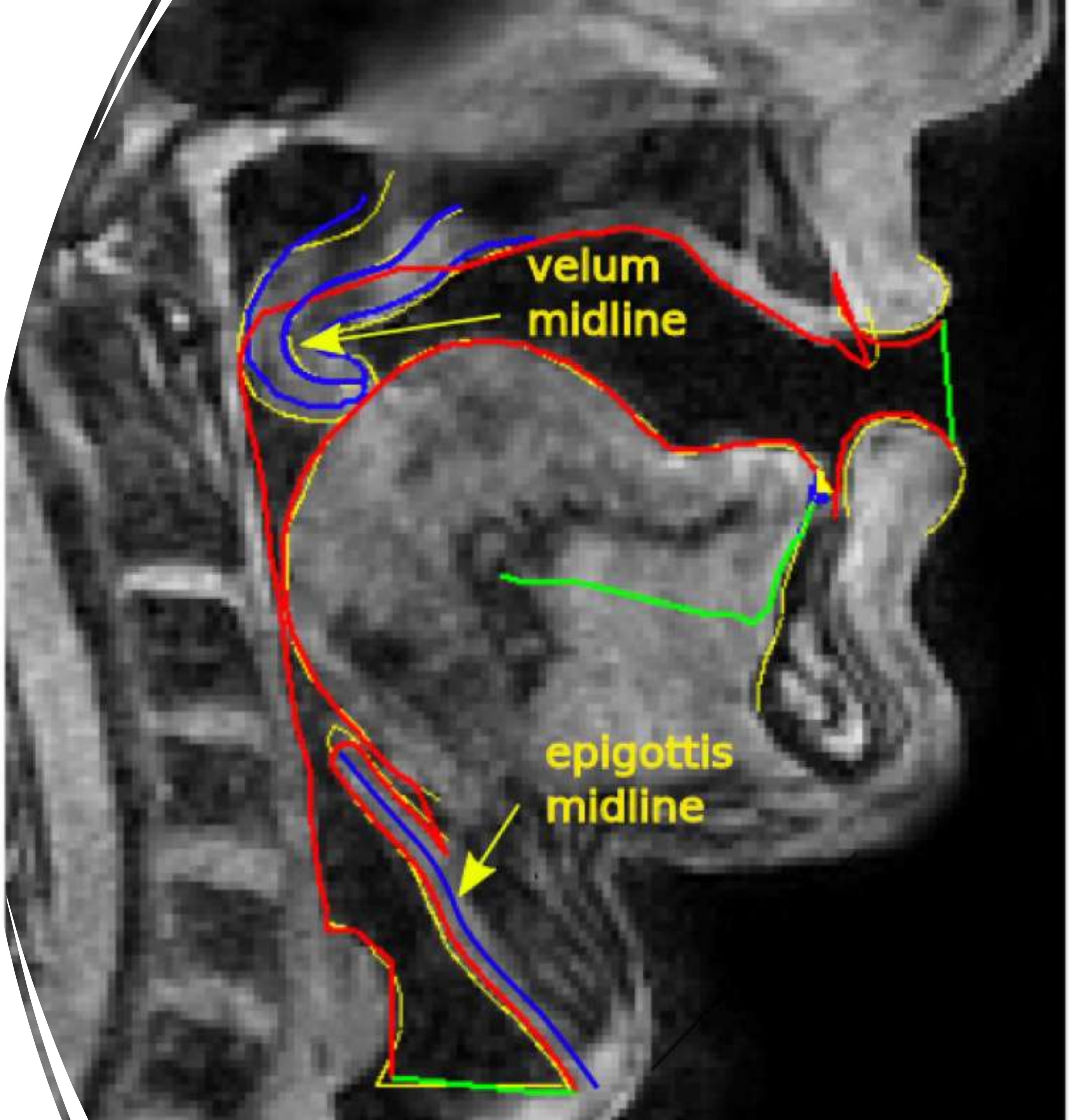
Speech Synthesis Technologies

Evolution of Speech Synthesis

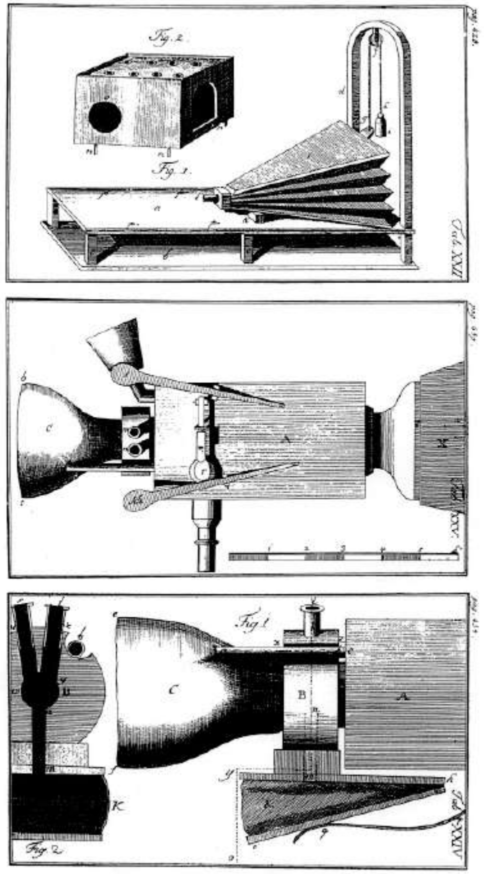


Articulatory Speech Synthesis

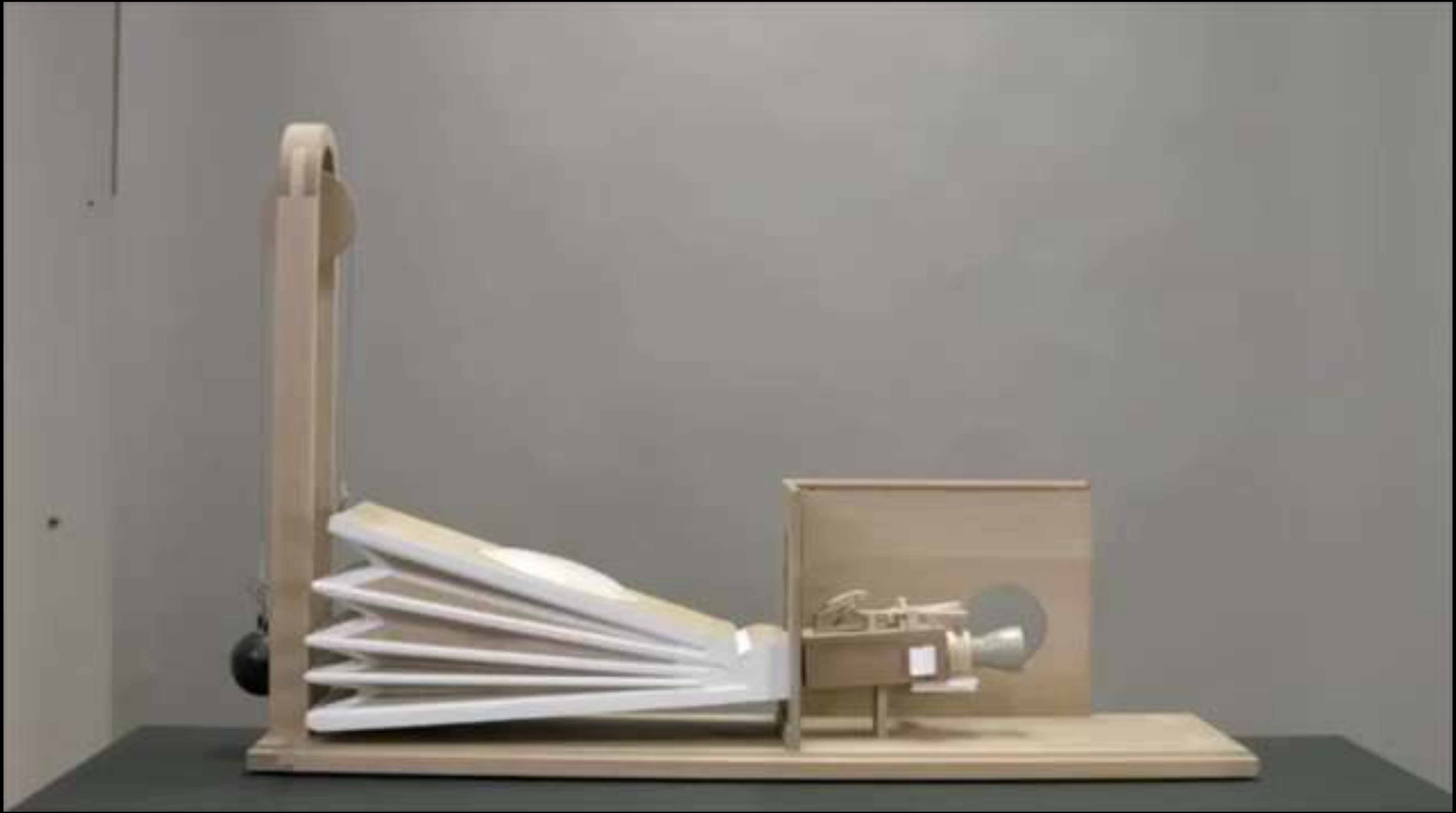
- Replication of the movements of human articulators
- Difficult to gather data
- Useful for phonological studies
- Inferior synthetic speech quality



von Kempelen Speaking Machine (1800)



Google Arts & Culture



https://youtu.be/oljkzZGe2I8

Voder 1939

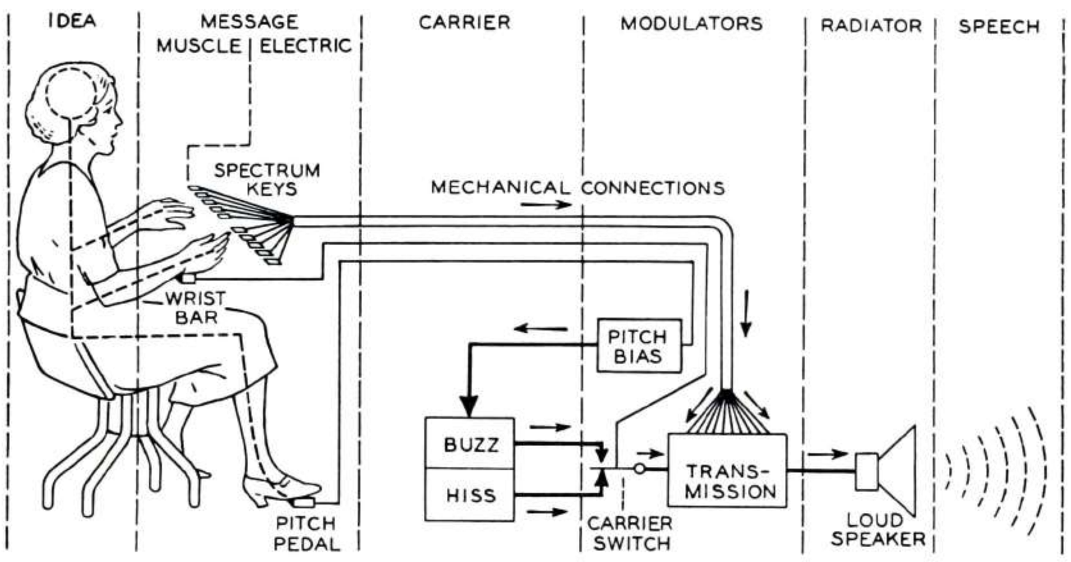
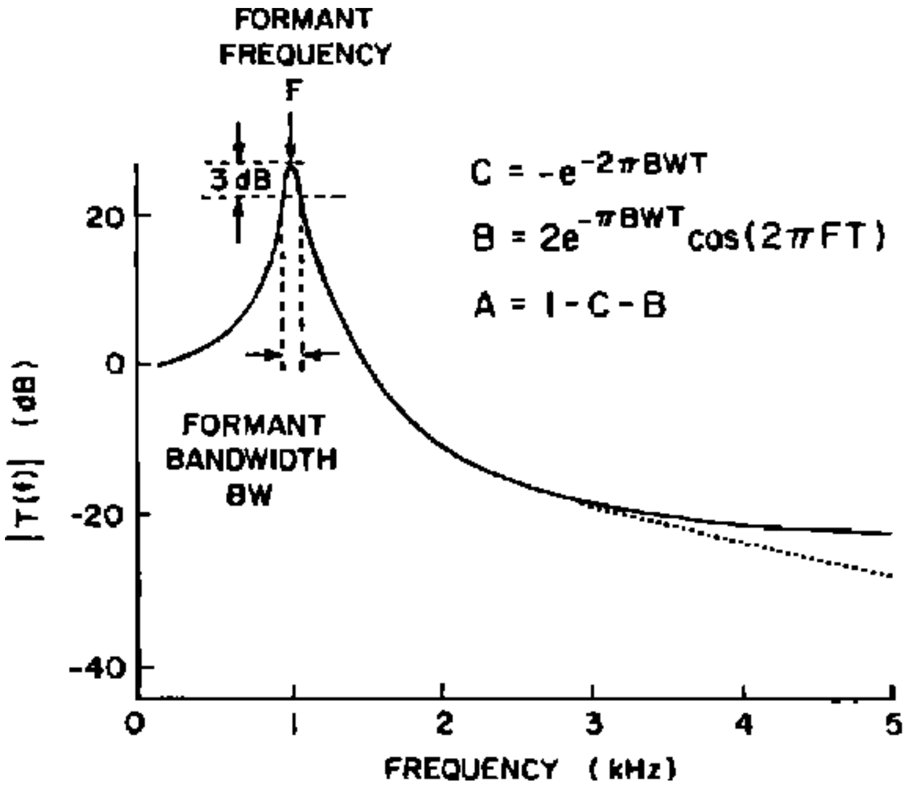


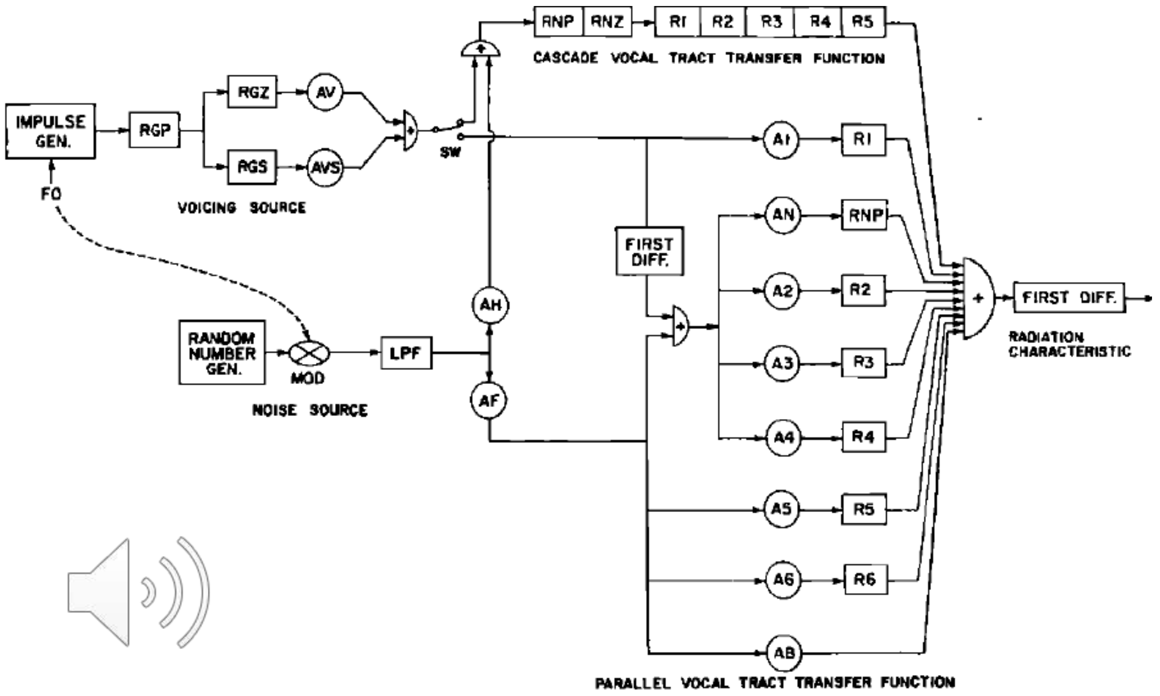
Fig. 8—Schematic circuit of the voder.

https://youtu.be/0rAyrmm7vv0

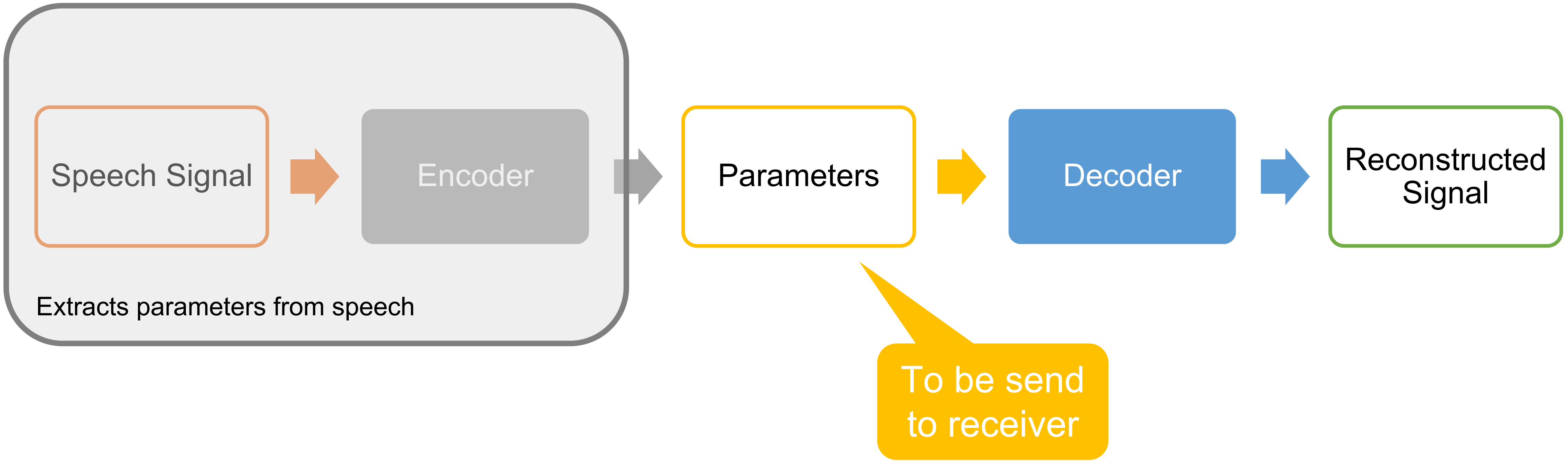
Formant Speech Synthesis

- Simplified source-filter model
- Each vocal tract resonance is model by a 2nd order filter
- Cascade/parallel association
- Rule-based parameter control

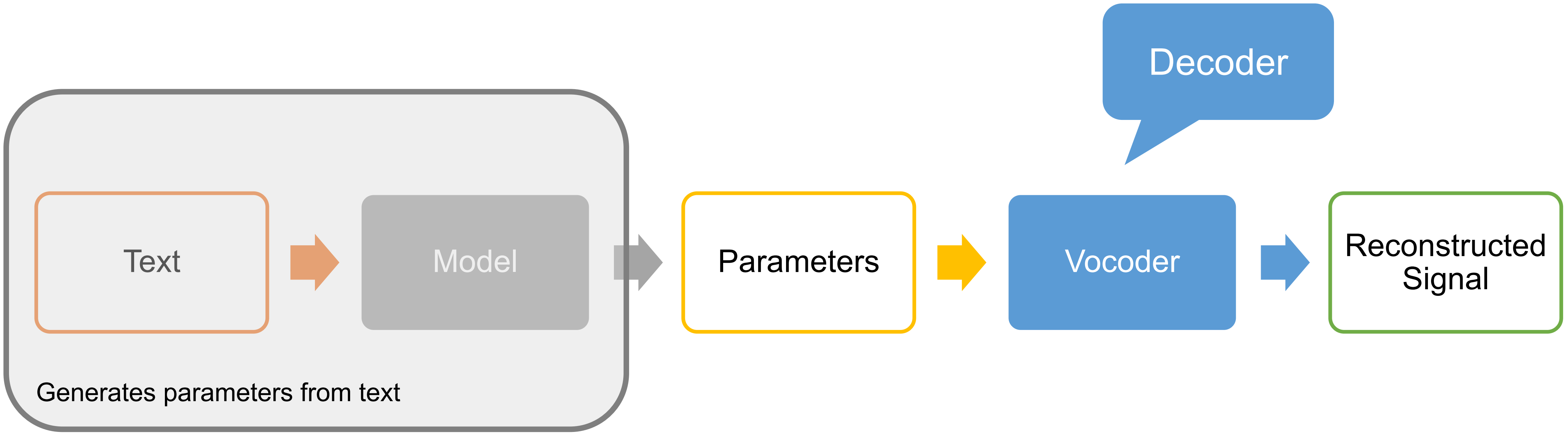




Speech Coding: Removing Redundancy



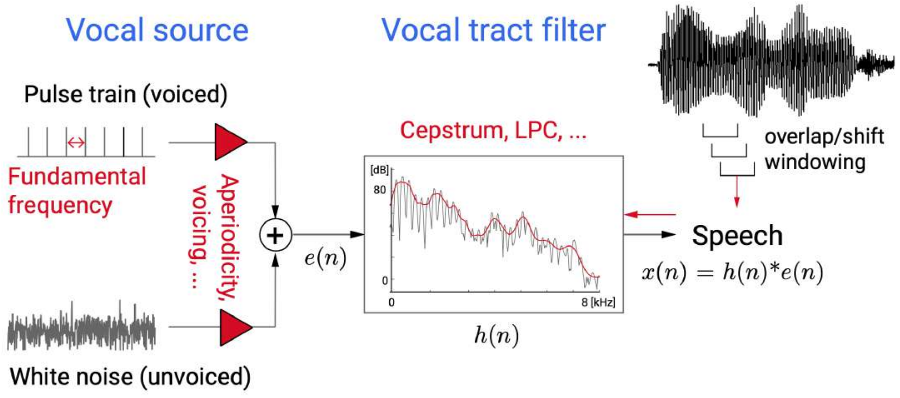
Text-to-Speech Synthesis



Text-to-Speech LPC Synthesizer

Very simplified parametrization

Text → Model → f0, RMS LPC Coeffs → Vocoder → Reconstructed Signal

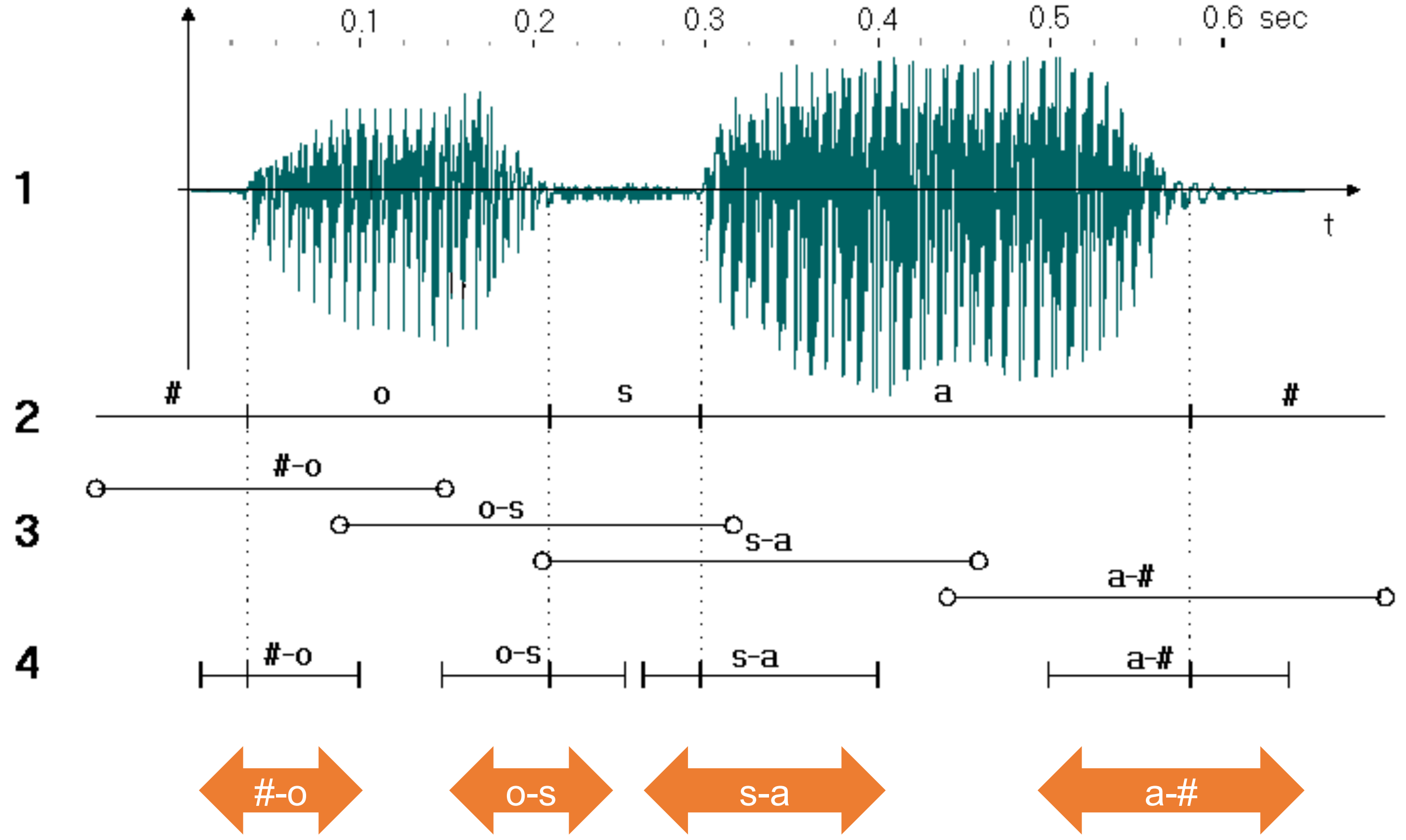


Advantages of the LPC Model

- Parameters easily extracted from speech
- Parameters can be interpolated
- Modifiable fundamental frequency
- Modifiable duration

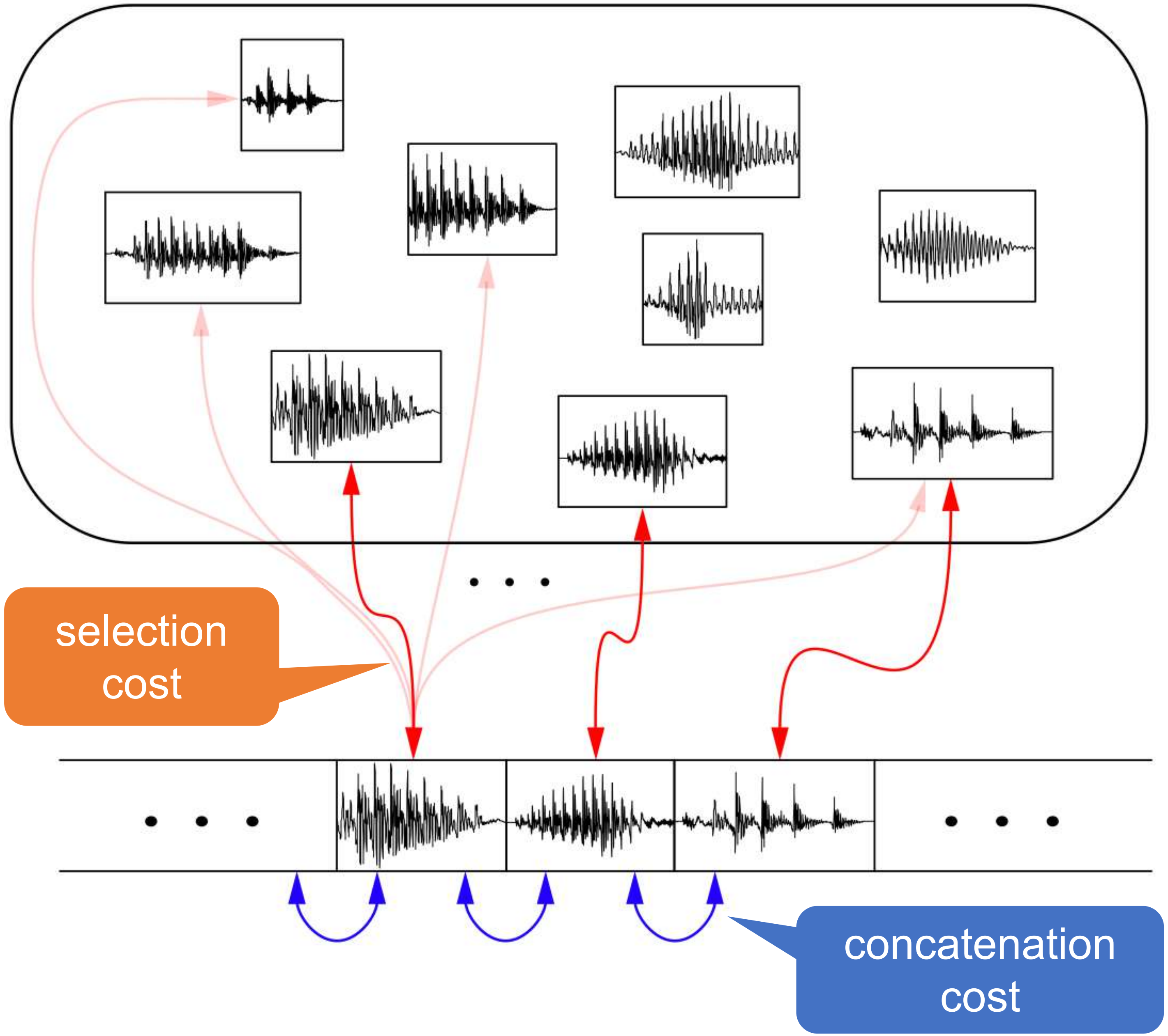
Diphone

- Speech sound segment between the stable regions of two phones
- Captures the transition dynamics
- Can be extended to larger segments



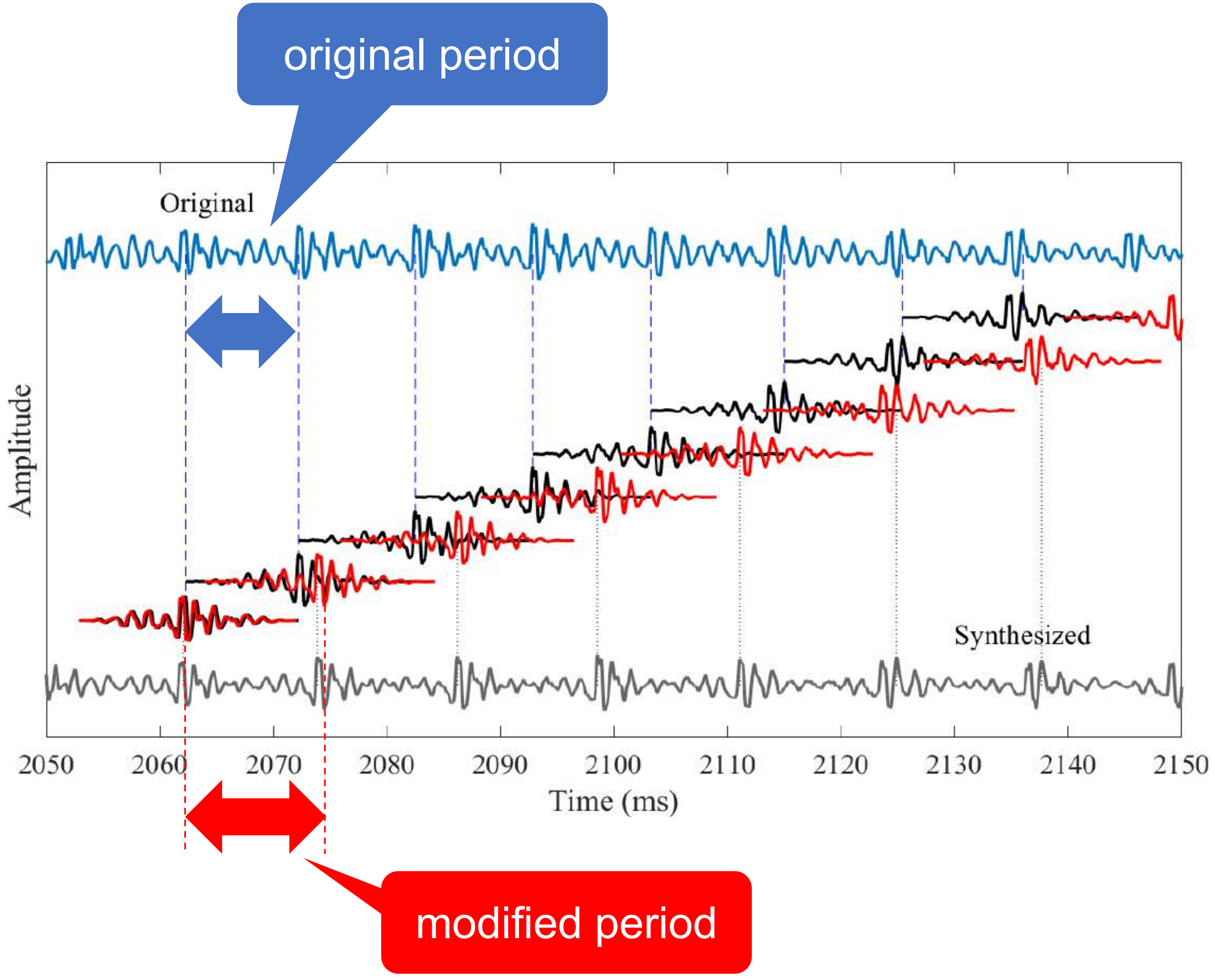
Concatenative Speech Synthesis

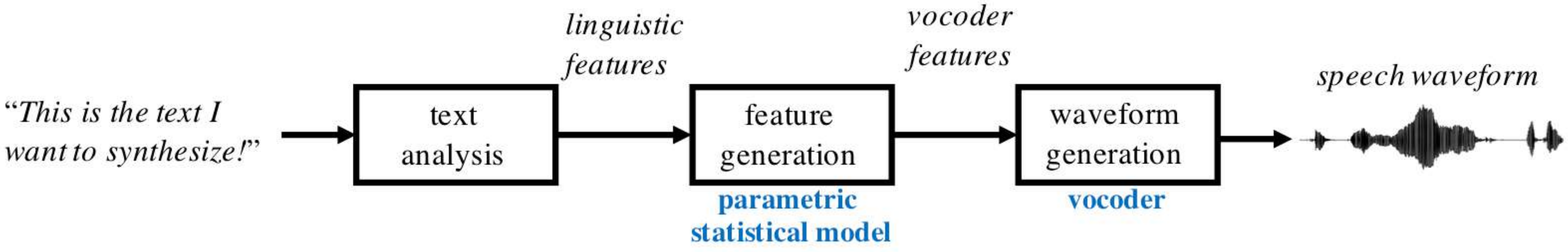
- Piecing together recorded speech units from a database
- Minimization of selection and concatenation costs
- Requires extensive recordings from a single speaker
- Highly intelligible
- Lack of naturalness and emotional expressiveness



Pitch Synchronous Overlap Add (TD-PSOLA)

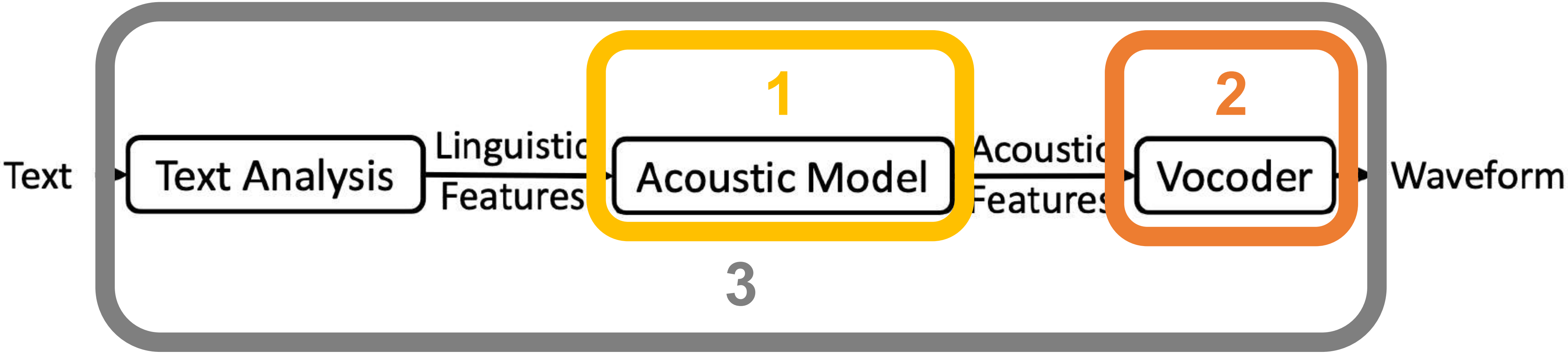
- Find the largest peak in each period
- Window covering two pitch periods
- Shift windows to match the desired period





Statistical Parametric Speech Synthesis (SPSS)

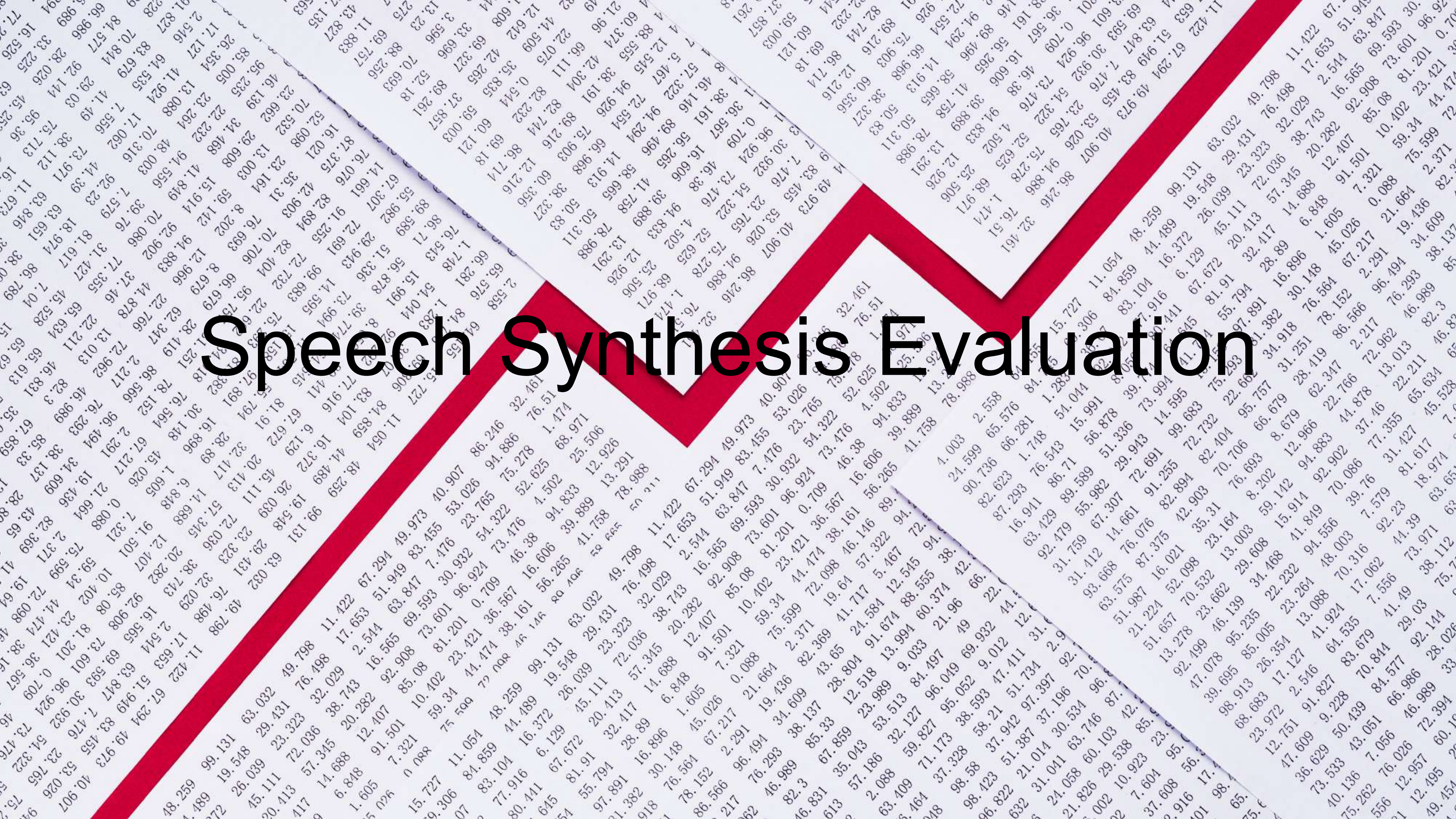
- HMM acoustic model to generate acoustic parameters
- The acoustic model is trained with paired linguistic features and acoustic features
- Requires fewer data than concatenative synthesis
- More flexibility but lower intelligibility and robotic voice quality



Neural Speech Synthesis

- Replace the HMM acoustic model with DNN
- Replace the vocoder with DNN
- Allows an end-to-end system
- High voice quality: intelligibility and naturalness

Speech Synthesis Evaluation

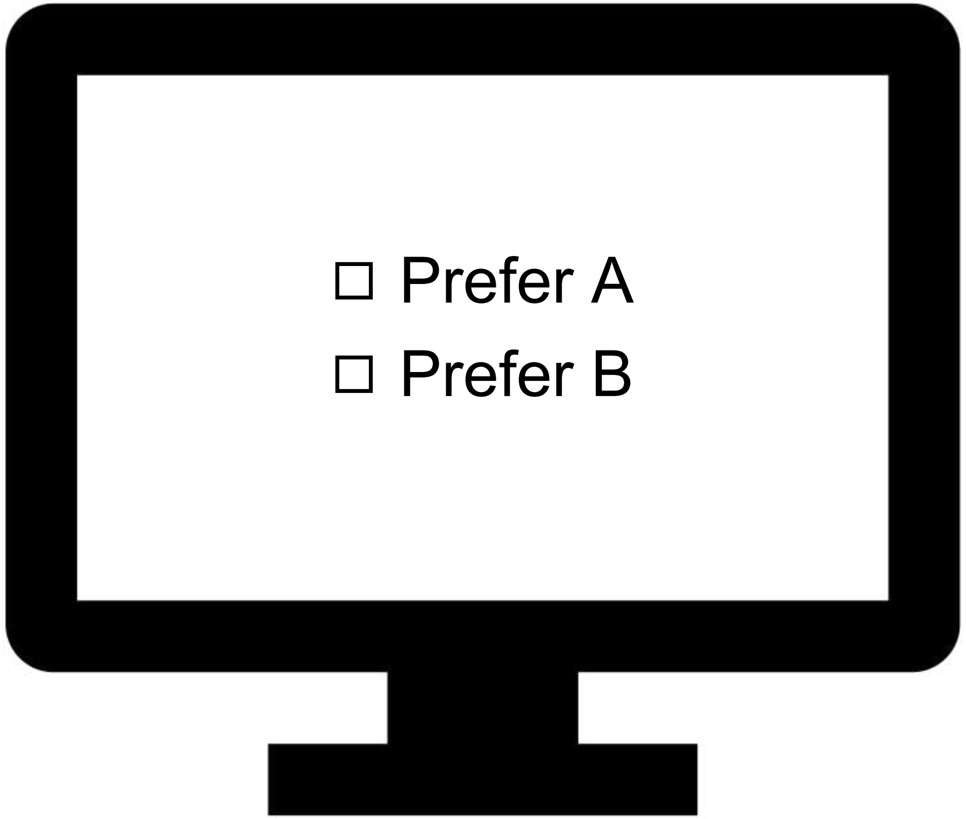


Mean Opinion Score

- Subjective test
- Rating on a 1-5 scale
- Careful design
- Native language listeners
- Baseline examples

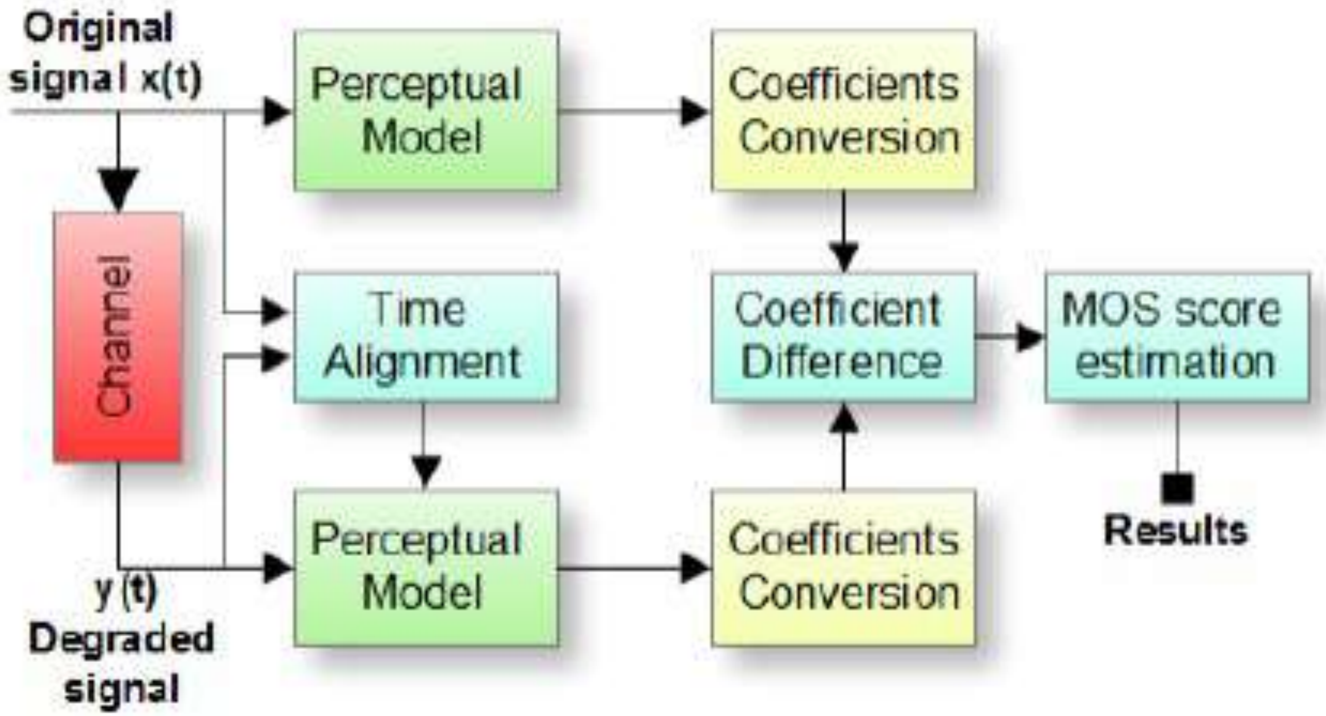
AB Test

- Subjective test
- Preference test
- Listeners are presented with two versions
- Careful design
- Native language listeners
- Baseline examples



Perceptual Evaluation Speech Quality (PESQ)

- Designed to assess voice communication systems
- Needs a reference speech signal
- Spectral distortion
- Temporal alignment
- Noise
- Score range between -0.5 and 4.5



Mel Cepstral Distortion

- Needs a reference speech signal
- Extracts MFCCs from both the synthesized and reference speech signals
- Euclidean distance between each corresponding pair of MFCCs is computed and averaged over all frames.
- Does not capture prosody, intonation, or pronunciation accuracy.

$$\mathrm{MCD}_{\mathrm{dB}} = \frac{\alpha}{N} \sum_{t=0}^{N-1} \sqrt{\sum_{k=1}^{P} (\mathrm{MC}_{\mathrm{syn}}(t,k) - \mathrm{MC}_{\mathrm{ref}}(t,k))^2}$$

Word Error Rate (WER)

- Measures the intelligibility of the synthesized speech
- Does not require a reference speech signal
- Requires a speech recognizer
- Percentage of incorrectly recognized words in the synthesized output compared to the reference text.

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\%$$

Probabilistic Formulation of TTS

Probabilistic Formulation of TTS

$\mathbf{w}$

text to be synthesize

$\mathbf{x}$

synthesized waveform

$$p(\mathbf{x}|\mathbf{w}, \mathcal{X}, \mathcal{W})$$

Separate inference from training

$(\mathcal{X}, \mathcal{W})$

dataset of waveform, transcriptions

$$p(\mathbf{x}|\mathbf{w}, \mathcal{X}, \mathcal{W}) = \int \underbrace{p(\mathbf{x}|\mathbf{w}, \lambda)}_{\text{inference}} \underbrace{p(\lambda|\mathcal{X}, \mathcal{W})}_{\text{training}} \, d\lambda$$

λ is the model

Linguistic and Acoustic Features

acoustic features

linguistic features (labels)

$$p(\mathbf{x}|\mathbf{w},\mathcal{X},\mathcal{W}) = \int\int \underbrace{p(\mathbf{x}|\mathbf{o})}_{\text{vocoder}}\, \underbrace{p(\mathbf{o}|\mathbf{l},\lambda)}_{\text{acoustic}} \underbrace{P(\mathbf{l}|\mathbf{w})}_{\text{linguistic}} \underbrace{p(\lambda|\mathcal{X},\mathcal{W})}_{\text{training}}\, d\lambda\, d\mathbf{o}$$

joint optimization

$$\left\{\hat{\mathbf{o}},\hat{\mathbf{l}},\hat{\lambda}\right\} = \arg\max_{\mathbf{o},\mathbf{l},\lambda} \{p(\mathbf{x}|\mathbf{o})p(\mathbf{o}|\mathbf{l},\lambda)P(\mathbf{l}|\mathbf{w})p(\lambda|\mathcal{X},\mathcal{W})\}$$

$$p(\mathbf{x}|\mathbf{w},\mathcal{X},\mathcal{W}) \approx \underbrace{p(\mathbf{x}|\hat{\mathbf{o}})}_{\text{vocoder}}\, \underbrace{p(\hat{\mathbf{o}}|\hat{\mathbf{l}},\hat{\lambda})}_{\text{acoustic}} \underbrace{P(\hat{\mathbf{l}}|\mathbf{w})}_{\text{linguistic}} \underbrace{p(\hat{\lambda}|\mathcal{X},\mathcal{W})}_{\text{training}}$$

TTS Pipeline

Training

$$\hat{\mathcal{O}} = \arg\max_{\mathcal{O}} p(\mathcal{X} \mid \mathcal{O})$$

Extract *acoustic features*

$$\hat{\mathcal{L}} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid \mathcal{W})$$

Extract *linguistic features*

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\mathcal{O}} \mid \hat{\mathcal{L}}, \lambda) p(\lambda)$$

Learn *mapping*

Inference

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} p(\boldsymbol{l} \mid \boldsymbol{w})$$

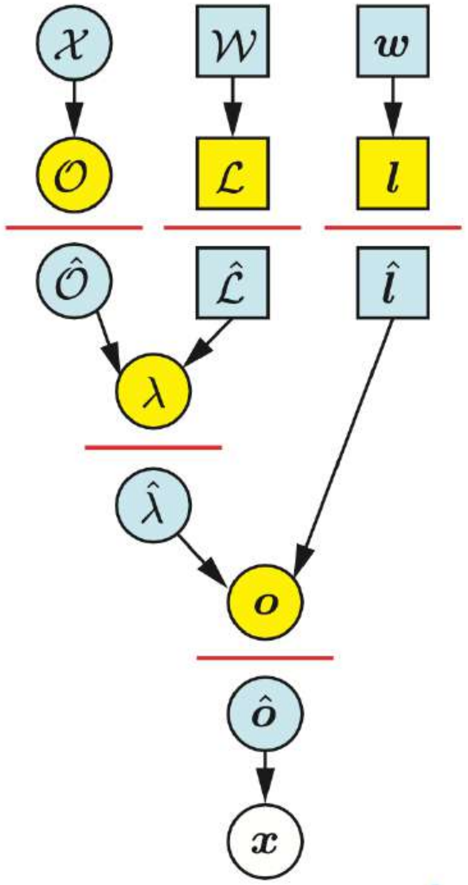
Predict *linguistic features*

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$

Predict *acoustic features*

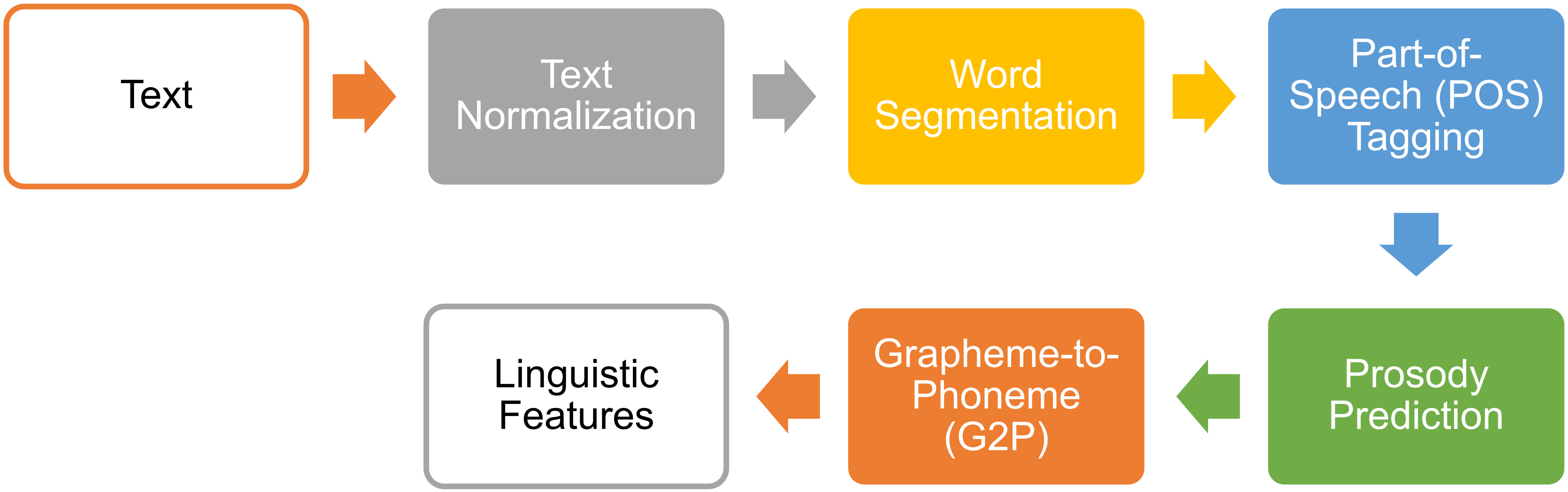
$$\bar{\boldsymbol{x}} \sim f_{\boldsymbol{x}}(\hat{\boldsymbol{o}}) = p(\boldsymbol{x} \mid \hat{\boldsymbol{o}})$$

Synthesize waveform



TTS Front End

TTS Front End



Text Normalization

Non-standard words	Semiotic class
Earthquake of 1755	Year
My phone is 123451755	Phone number
I paid € 1755	Money amount
I traveled 1775 km	Distance

Examples of Non-Standard Words

ordinal numbers	• 13th (thirteen)
roman numbers	• Charles ||| (Charles third)
percentage	• 3.5% (three point five per cent)
time	• 12:10 (twelve ten)
symbols	• + (plus)
abbreviations	• Av. (avenue)
acronyms	• NY (New York), GPU (gee pee u)

Grapheme-to-Phoneme (G2P)

- **Dictionary**
  - character-to-sound correspondences
- **Prediction**
  - words not in the dictionary
- **Disambiguation**
  - homographs non-homophones
- **Sandhi**
  - changes in word boundaries

Homographs Non-Homophones (GenAm)

The nurse wound the bandage around my wound. (/wund/, /waʊnd/)

I did not object after being asked to carry the large object. (/əbˈd͡ʒɛkt/, /ˈɑb.d͡ʒɛkt/)

Sheldon and Amy weren’t close enough to the car door to be able to close it so Leonard had to do it himself. (/kləʊs/, /kloʊz/)

The mouth of a huge bass was painted on the bass drum (/bæs/, /beɪs/)

I shed a tear when I saw the tear in my shirt. (/tɪəɹ/, /tɛəɹ/)

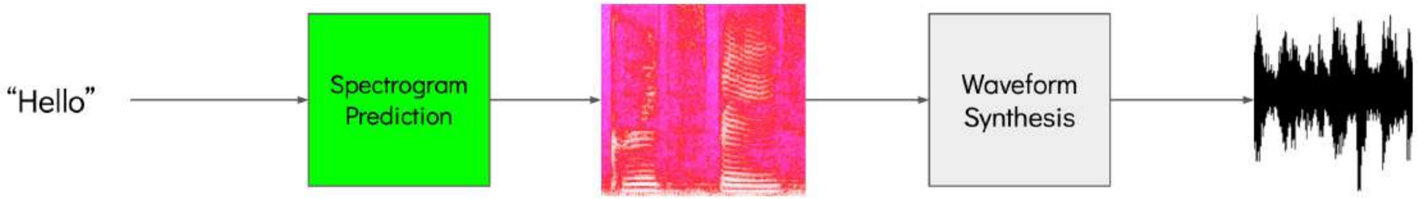
Homographs Non-Homophones (PT)

- Eu jogo nesse jogo (V/N)
- Ele foi colher flores com uma colher (V/N)
- Pisou um prego enquanto estava a pregar (N/V)
- O barco seguiu a sua rota mesmo com a vela rota (N/A)
- Estava na sede do clube e fiquei com sede (N/N)

Acoustic Model

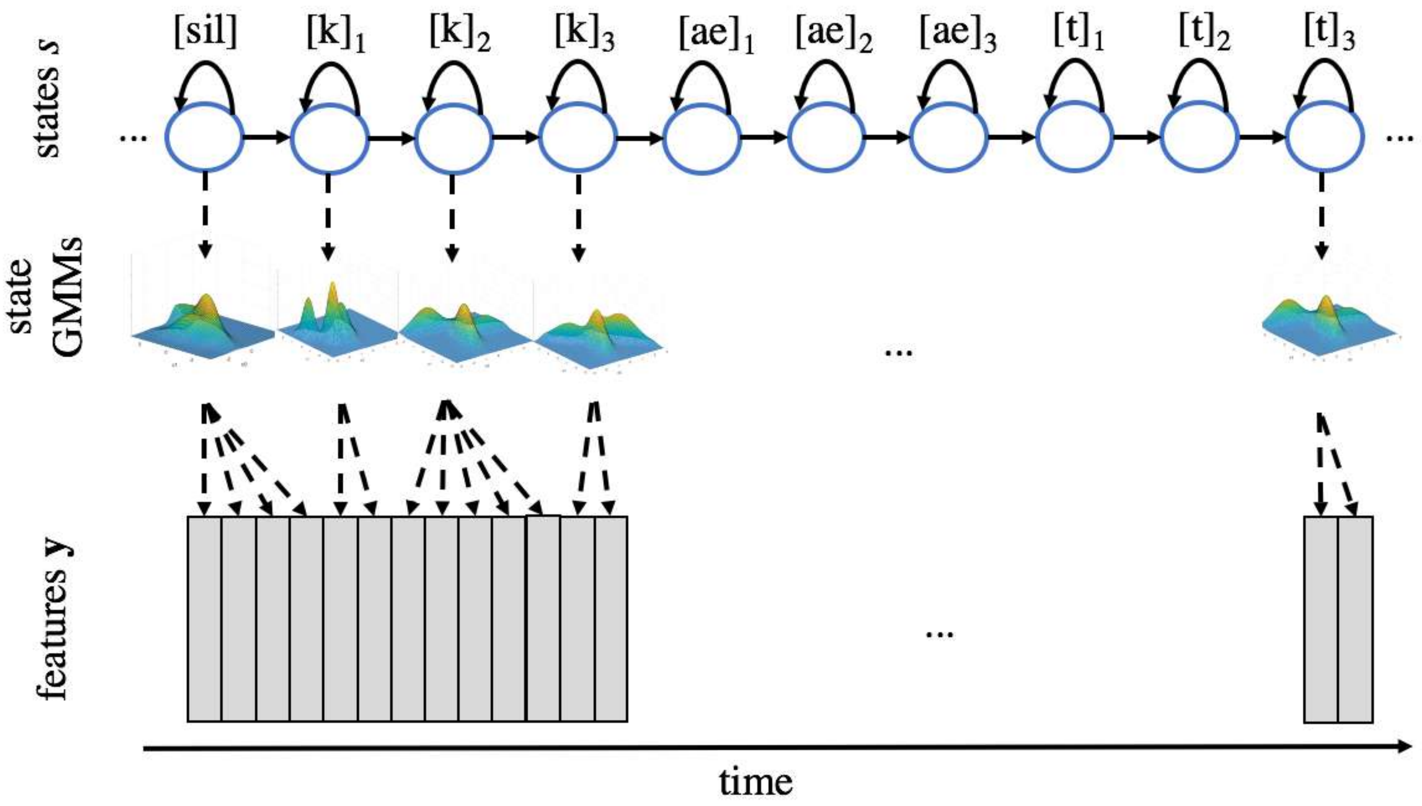
Intermediate Spectrogram

- Phones and prosody can be modeled with the magnitude of the spectrum
- Perceptually based (e.g. mel spectrogram)
- Adequate resolution in time with short-time analysis
- FFT provides efficient computation



SPSS: HMM Acoustic Model

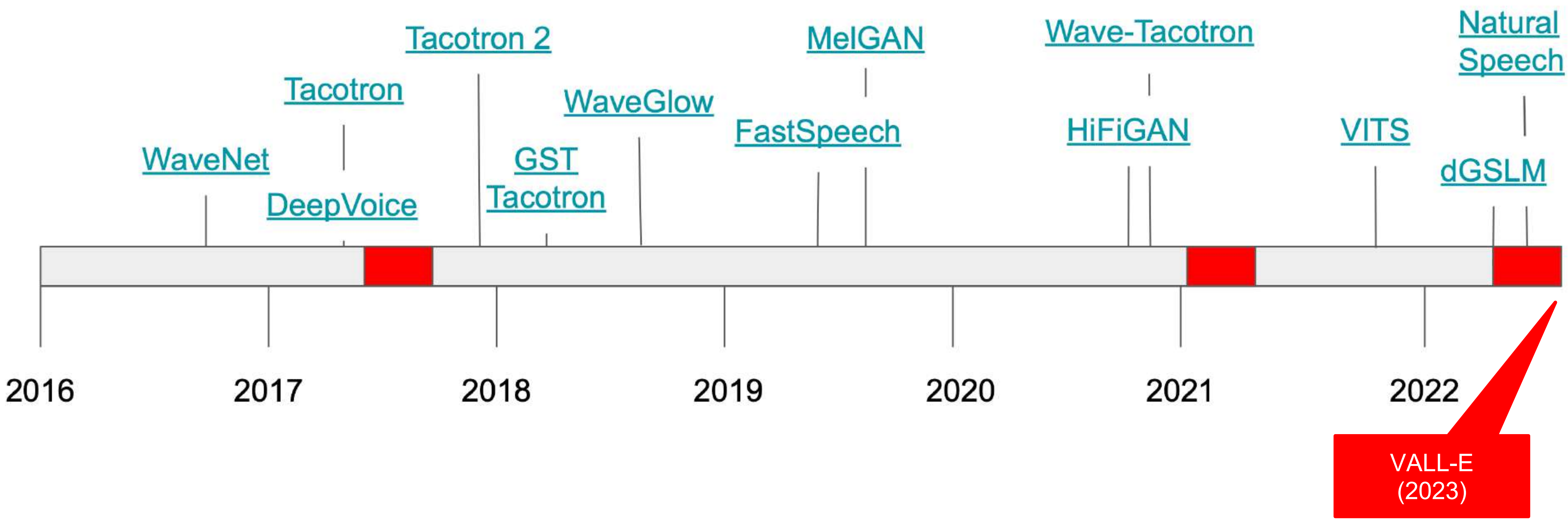
- Observation vectors are spectral parameters and f0 (acoustic features)
- Context-dependent modeling



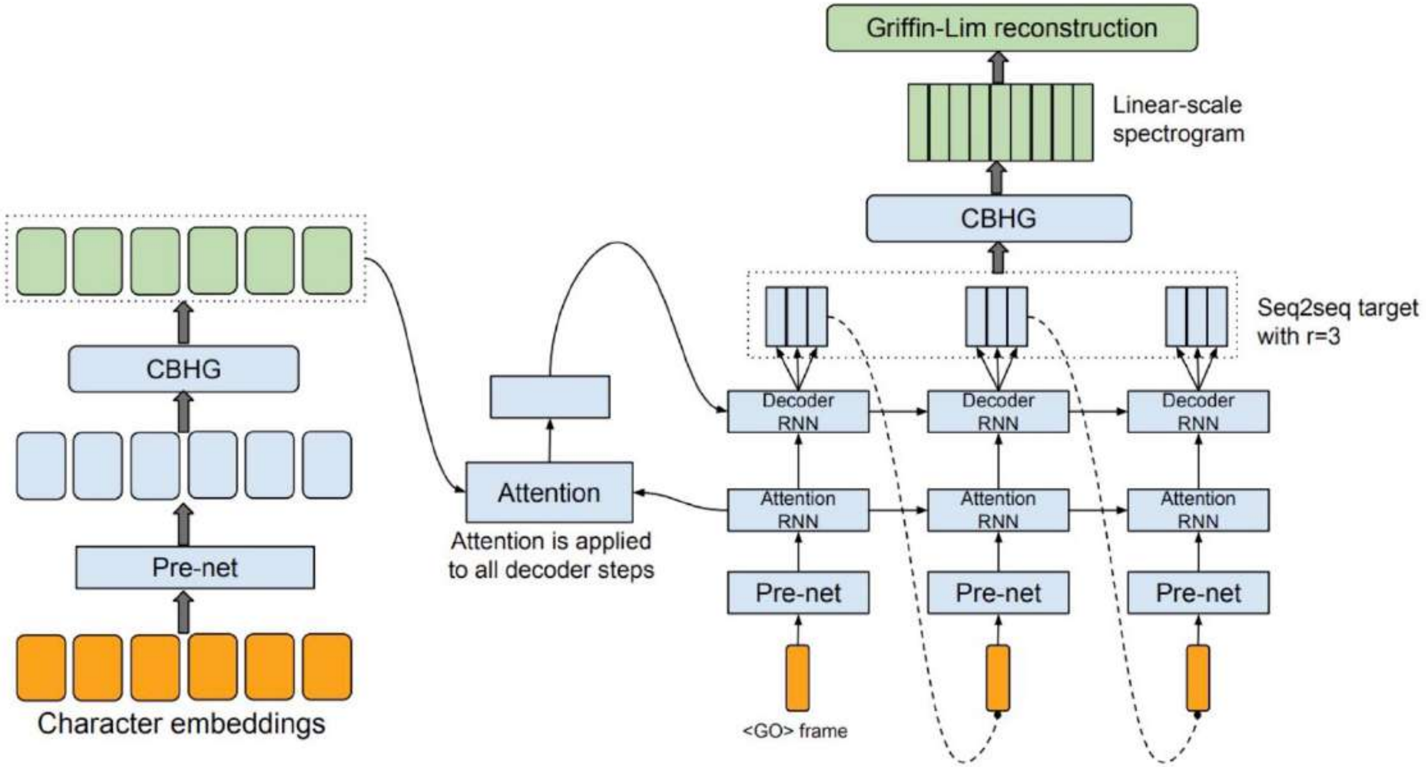
State chain: $P(s_t | s_{t-1})$

State emissions: $P(\mathbf{y} \mid s) = \sum_{k=1}^{K} \phi_{k,s} N(\mathbf{y} \mid \boldsymbol{\mu}_{k,s}, \Sigma_{k,s})$

Neural Speech Synthesis Models



RNN-based: Tacotron



- Encoder-decoder with attention
- Predicts mel spectrograms
- Pre-net provides information bottleneck for regularization
- CBHG is a 1-D convolutional filters, followed by highway networks and a bidirectional gated recurrent unit

Tacotron2

- Replaces CBHG and GRU by LSTM
- Bidirectional LSTM
- Location-sensitive attention instead of additive attention
- Training with the accurate spectrum, not the predicted
- WaveNet instead of Griffin-Lim

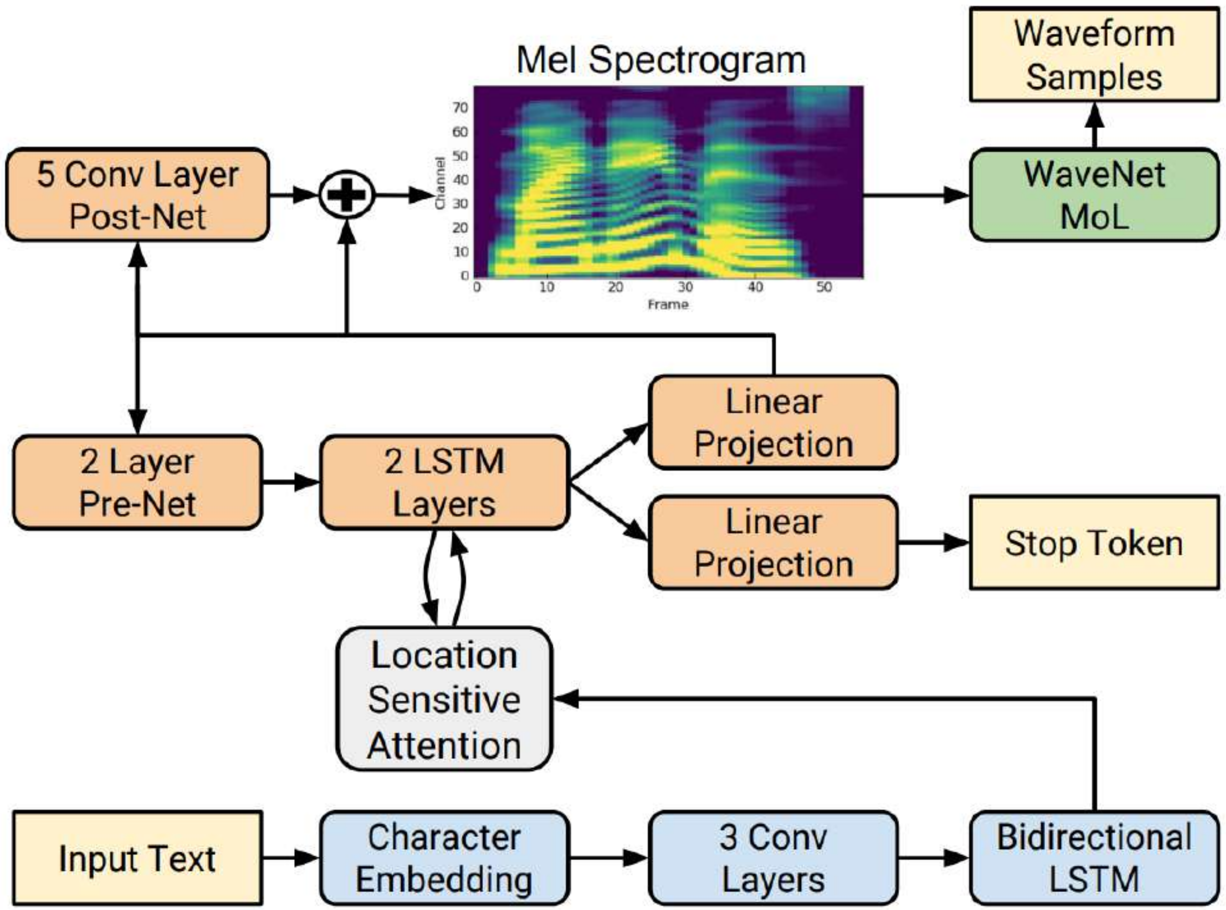


Fig. 1. Block diagram of the Tacotron 2 system architecture.

Attention vs Duration-Based S2S Models

Attention-based

- No alignments needed
- Adaptable to diverse or noisy datasets
- Capable of more natural prosody

Uses attention mechanism to align input and and output sequences

Duration-based

- Fast parallel inference
- Less chance of alignment problems
- Easier to train if alignments are available
- More robust to silence in training data

Uses an explicit duration model that predicts the duration of each phone

Waveform Generation

The Vocoder

- Initially conceived to reduce the bandwidth necessary to transmit intelligible voice
- Splits speech in source and frequency bands (acoustic features)
- Generates a waveform from the acoustic features
- Needs to reconstruct the phase information

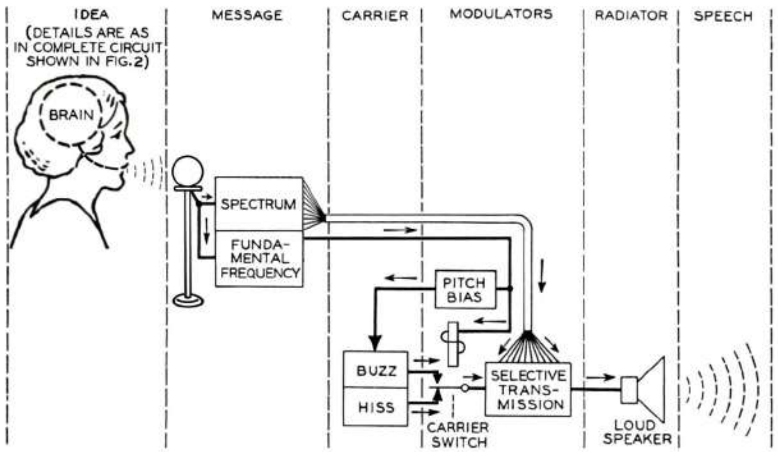
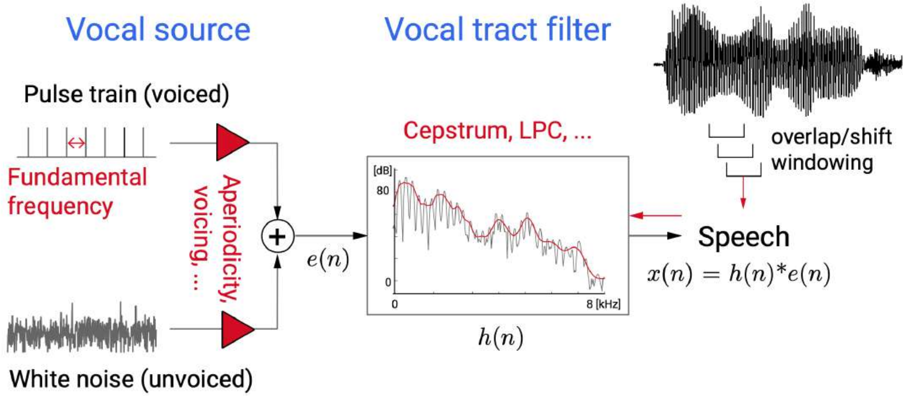
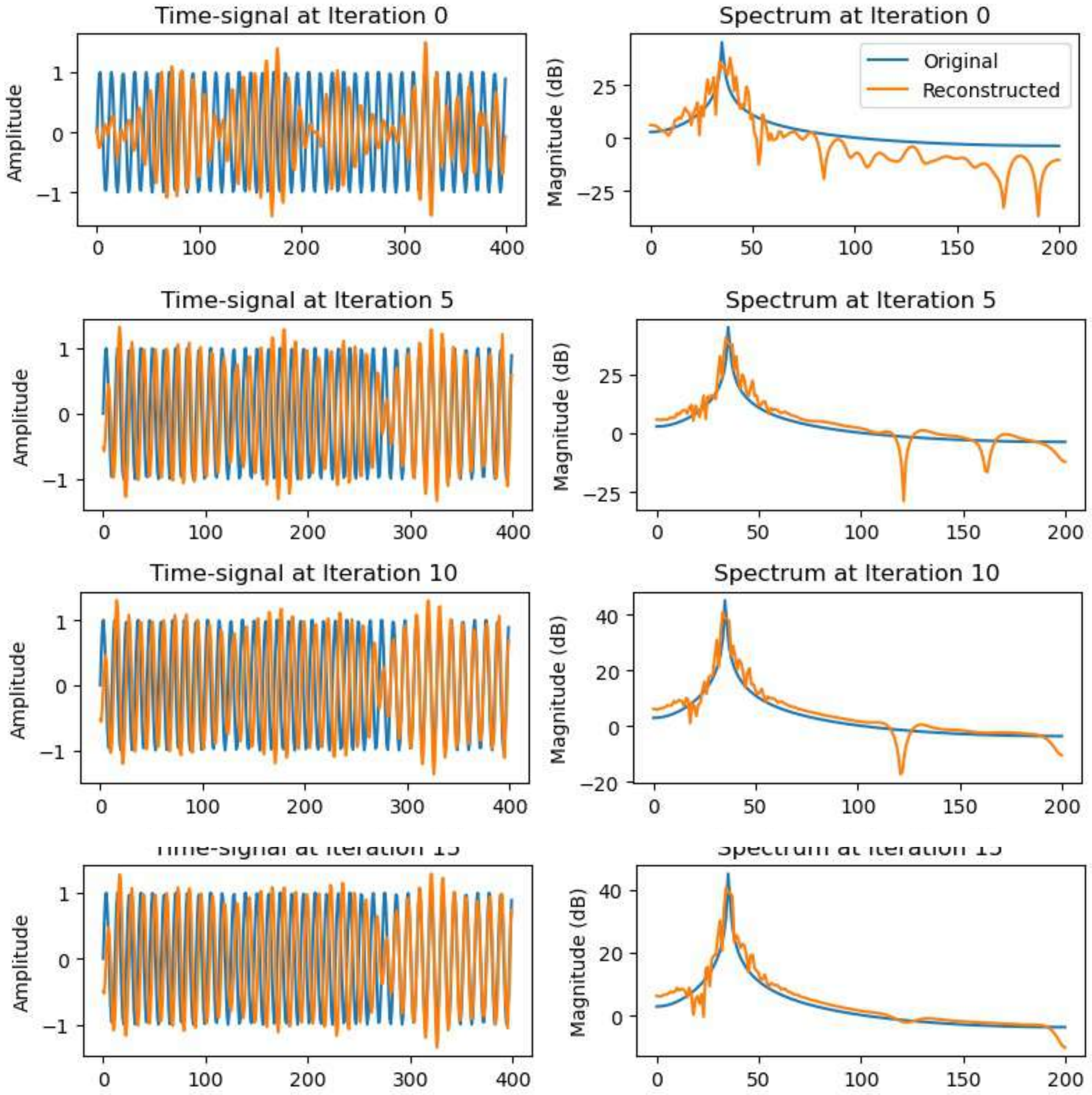


Fig. 7—Schematic circuit of the vocoder.



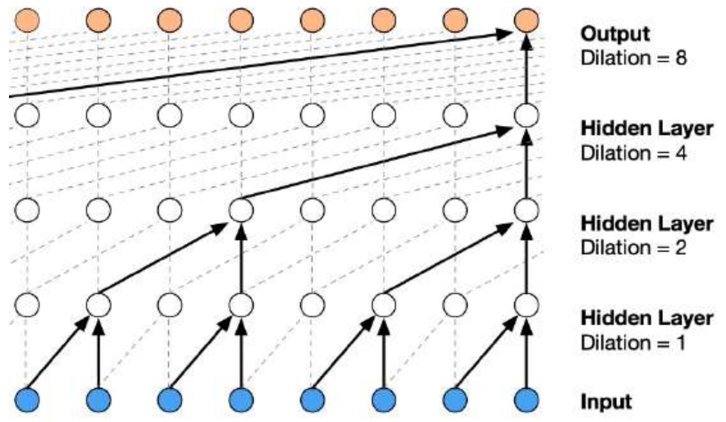


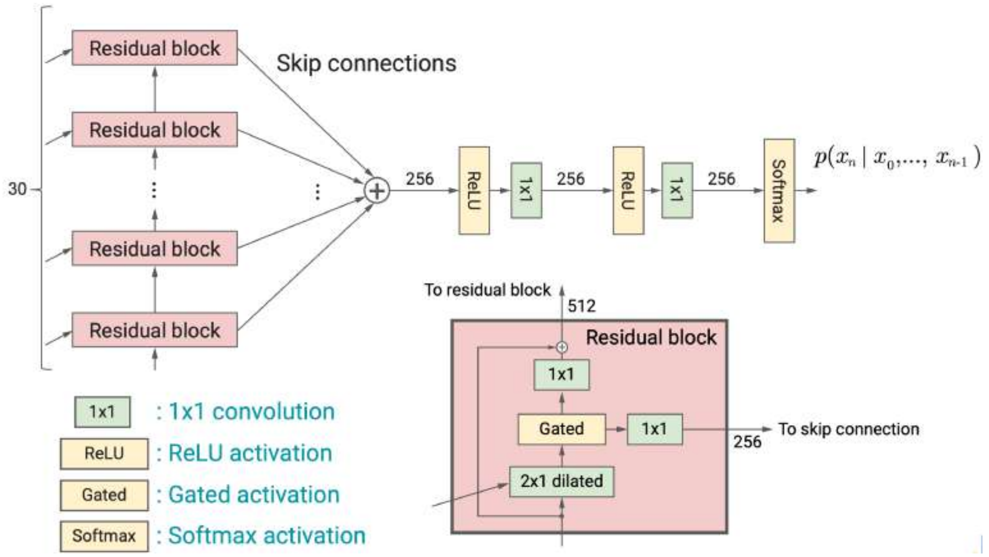
Grifin-Lim Algorithm

- Initialization: original magnitude + random phase
- Reconstruction: time-domain using ISTFT
- New Phase: extracted from STFT
- Phase Update: original magnitude + new phase
- Iteration: repeat reconstruction until convergence

WaveNet Vocoder

- Autoregressive model
- Predict the next sample with a stack of convolution layers
- Extent range by using dilated convolutions
- Softmax to produce discrete amplitude levels
- Extremely slow





Speaker and Style Embeddings

Speaker Characteristics

Speaking style

Personalized speech synthesis

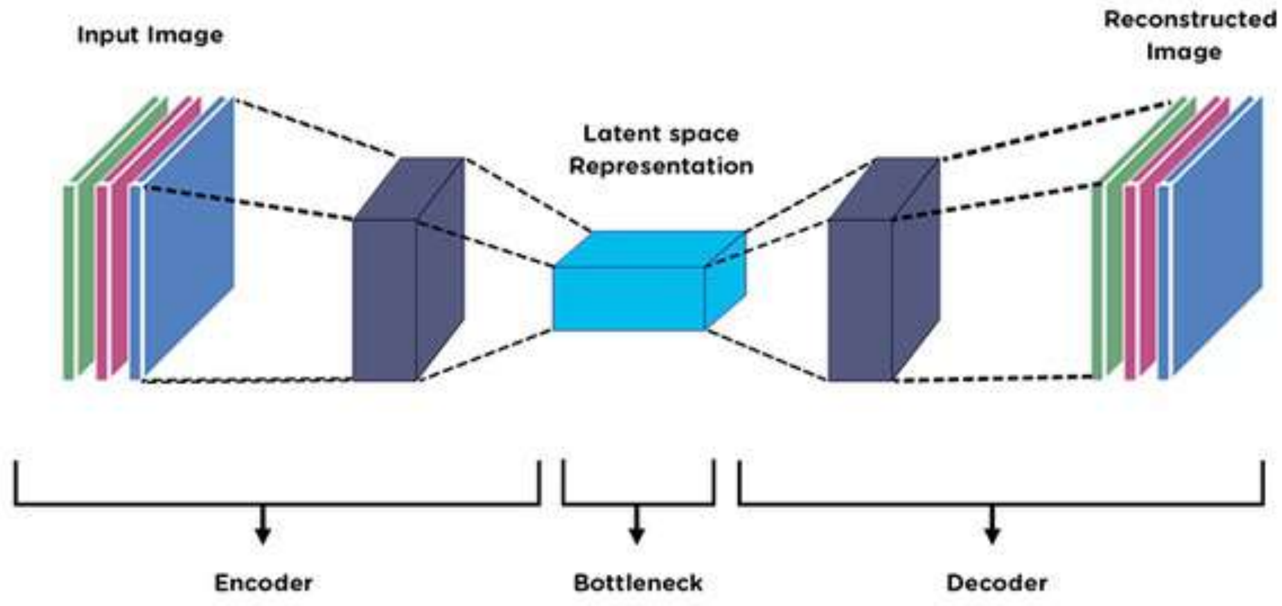
Voice-cloning

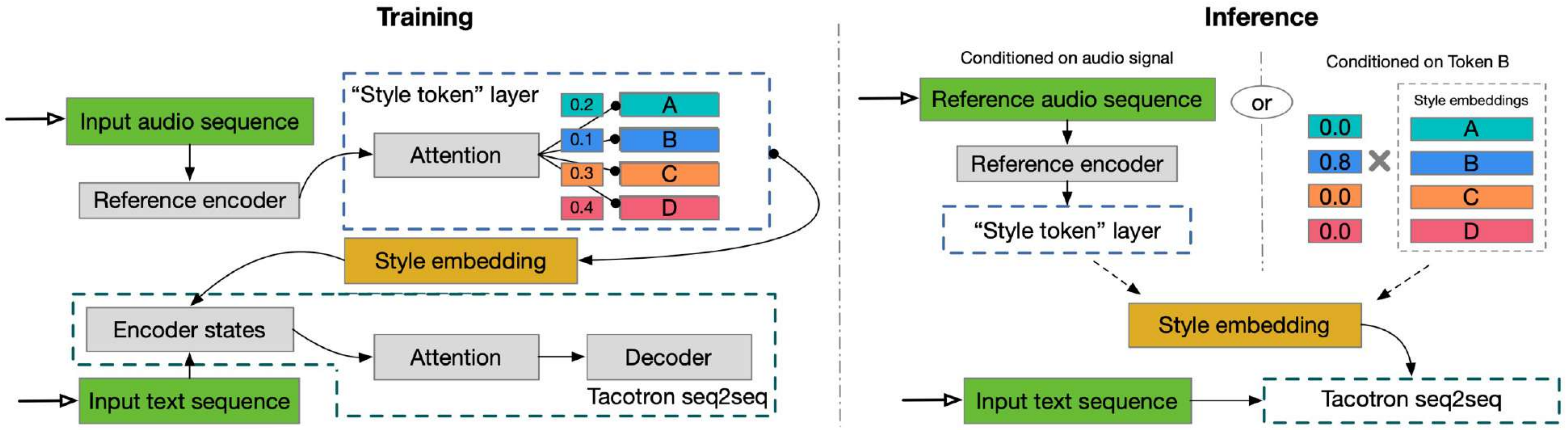
Cross-lingual voice cloning

Speech-to-speech translation

Latent Space

- Hidden state vector or bottleneck
- Module that contains the compressed knowledge representations



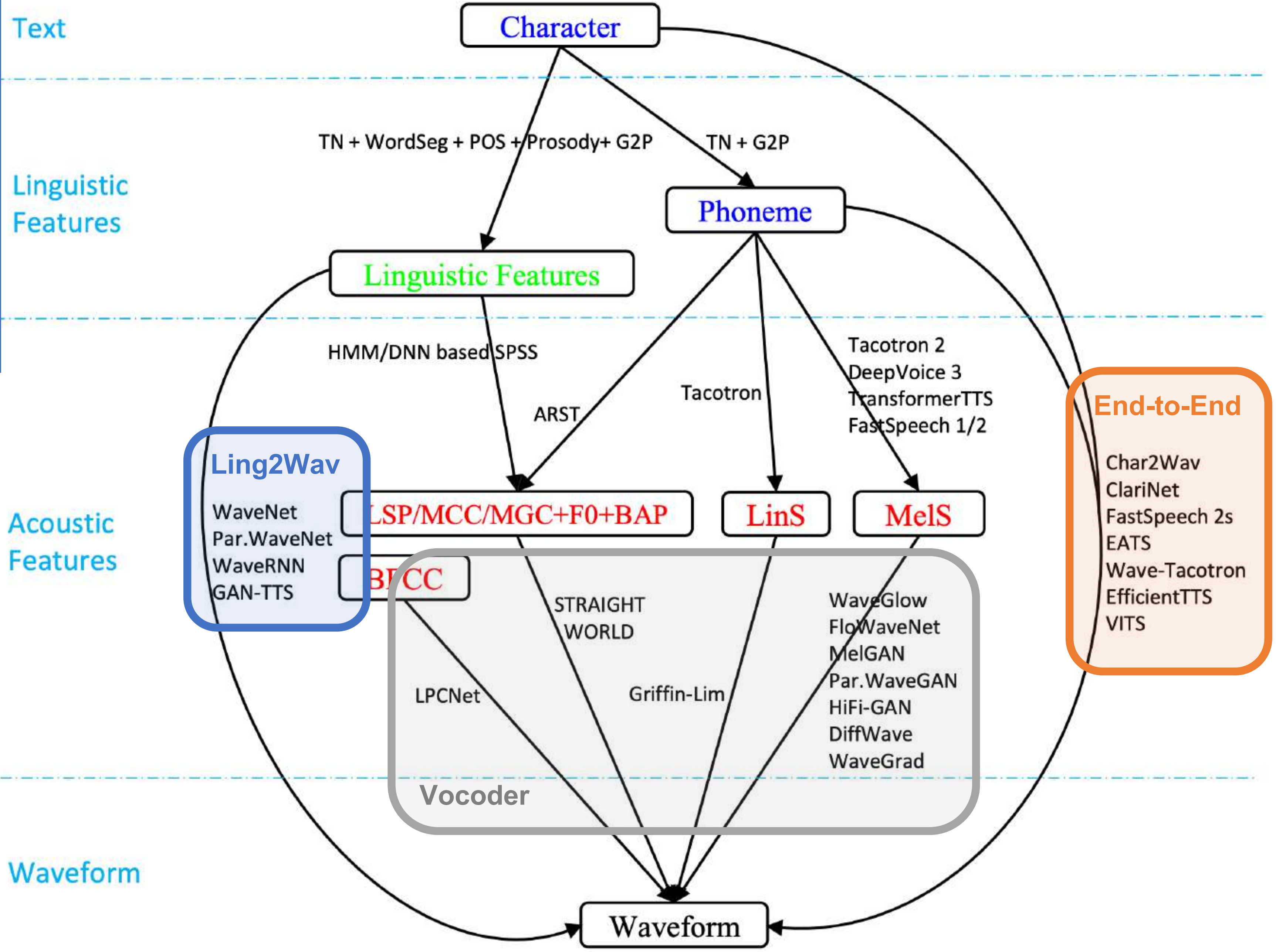


Global Style Tokens

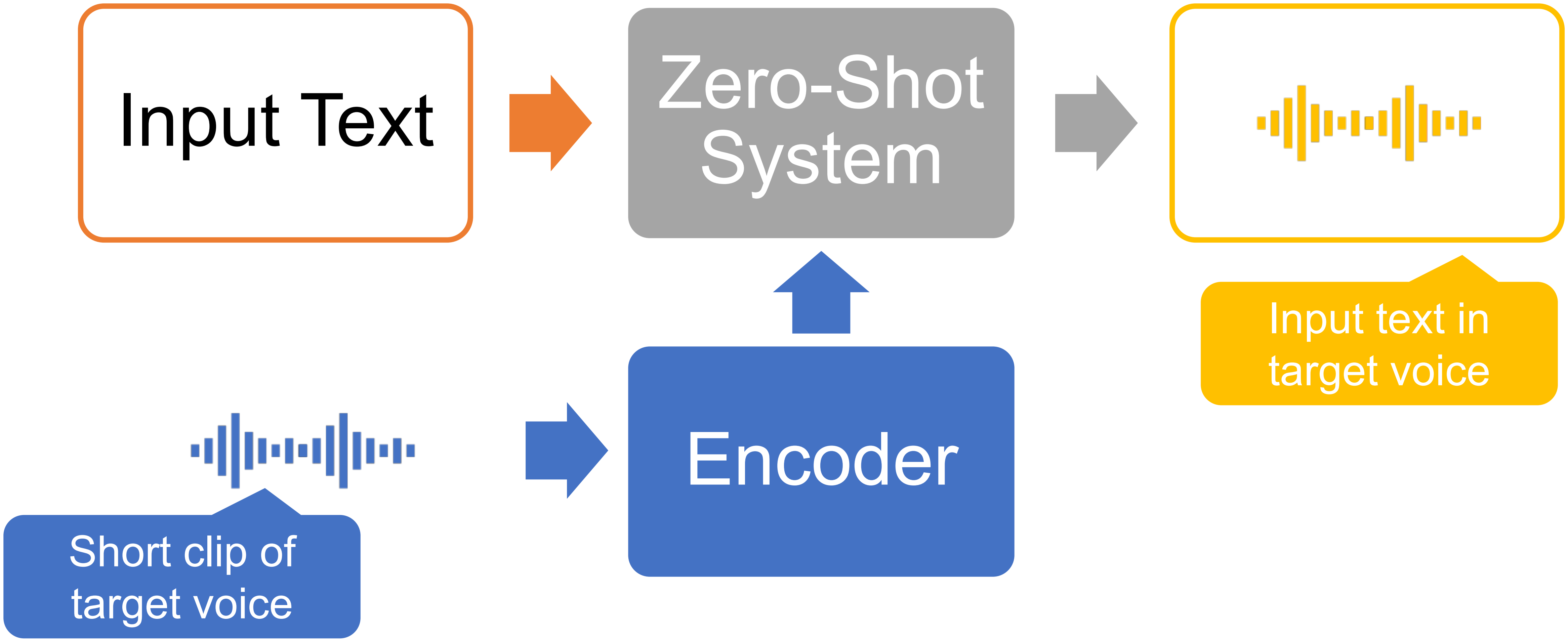
- Captures stylistic attributes or characteristics of speech
- Learned from large datasets with diverse speech styles
- Learned from the mel spectrogram by compressing the latent space
- Interpretable "labels" that can be used to modify the speaking style

End-to-End Models

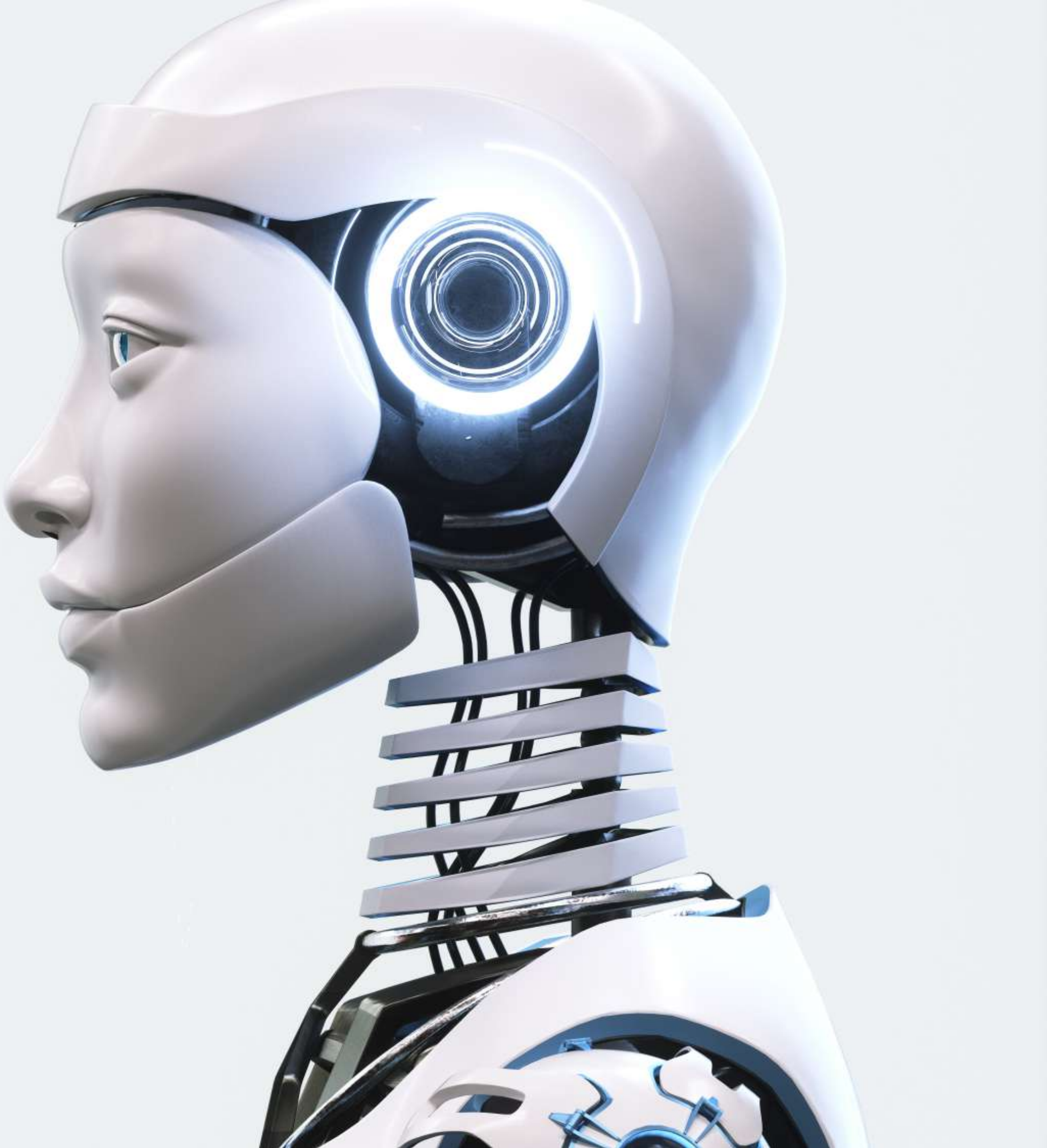
Neural TTS Systems



Zero-Shot TTS



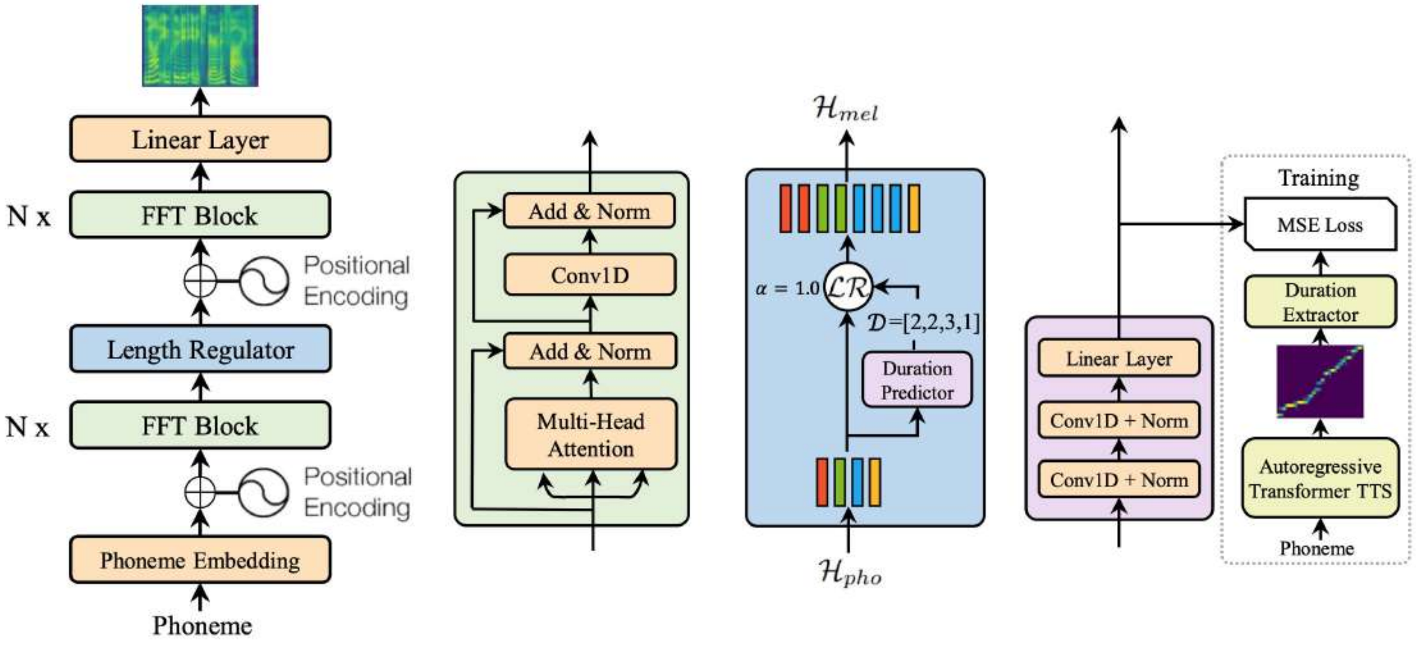
Part II: Advanced Topics



Acoustic Model

Transformer-based: FastSpeech

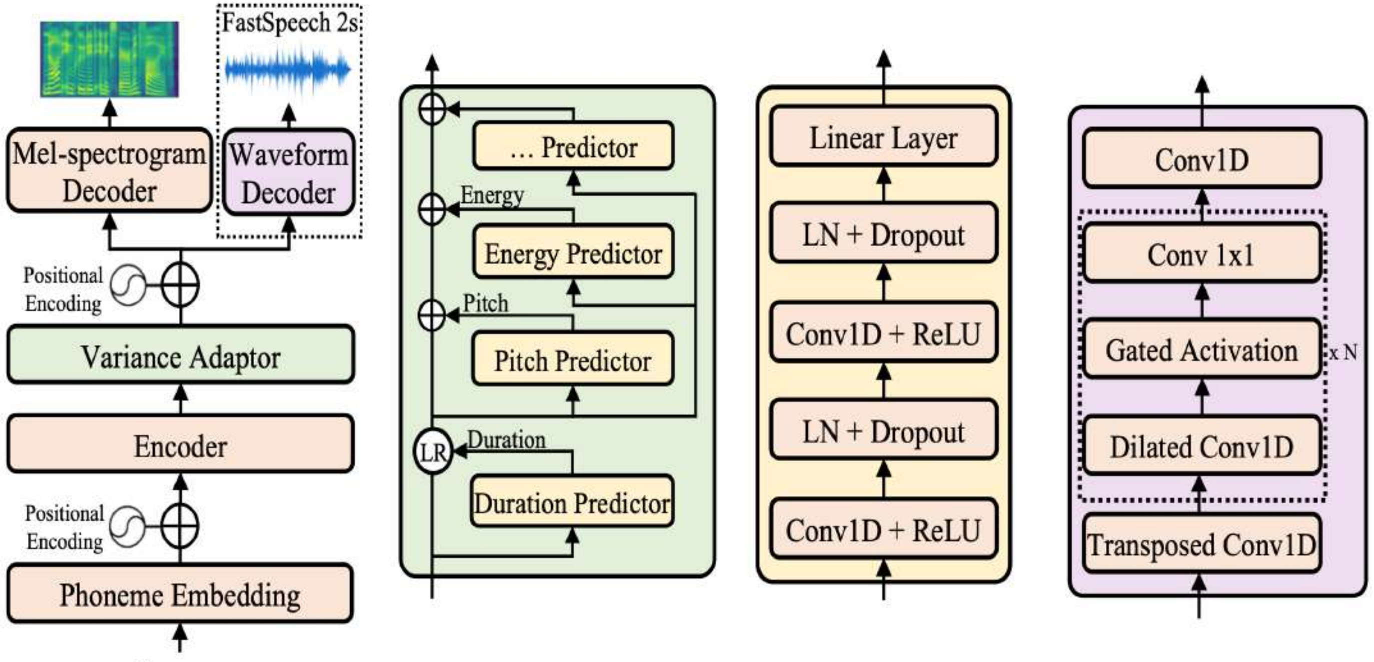
- Transformer predicts mel spectrograms in parallel
- No attention mechanism
- Explicitly predicts duration, energy and f0
- Exceptionally fast inference speed



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

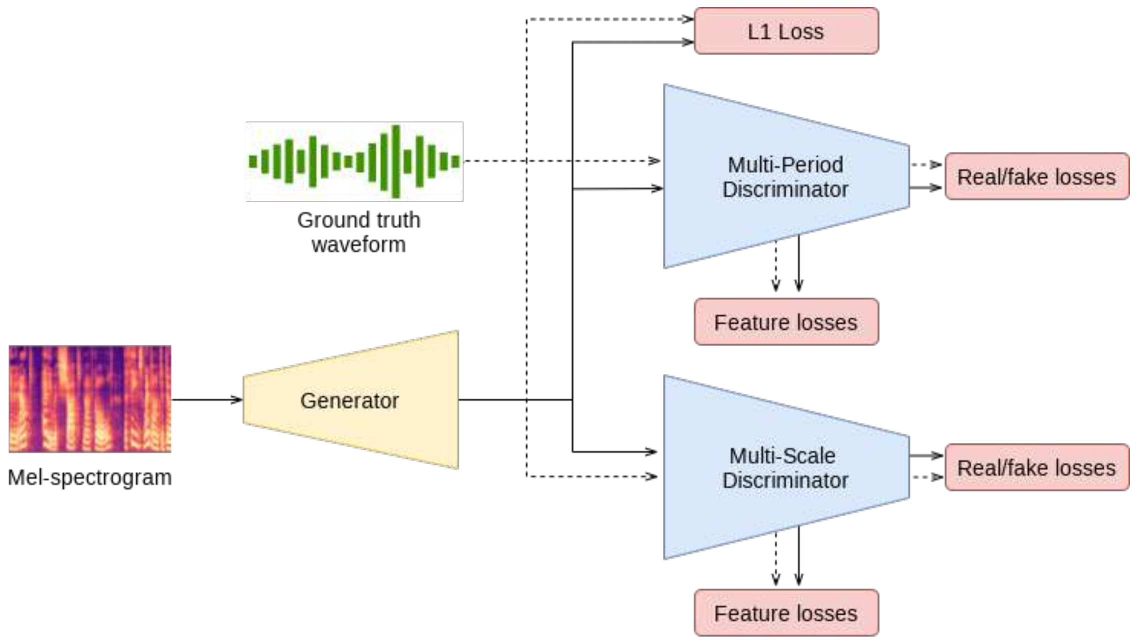
FastSpeech2

- Training with the accurate spectrum, not the predicted (AR)
- Variance information for f0, duration and energy
- Better voice quality



(a) FastSpeech 2 (b) Variance adaptor (c) Duration/pitch/energy predictor (d) Waveform decoder

Waveform Generation



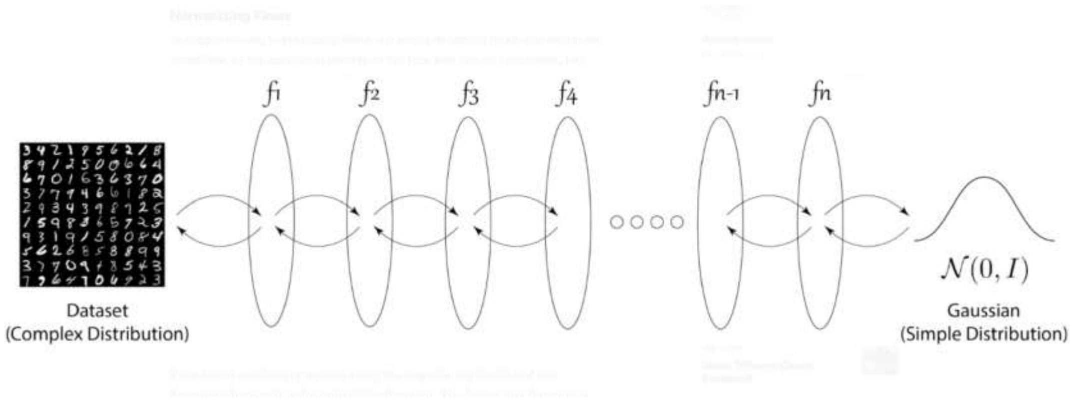
HifiGAN Vocoder

- Uses a Generative Adversarial Network (GAN)
- Generator: fully convolutional network
- Up-samples spectrogram to waveform temporal resolution
- Multi-period discriminator: periodic component
- Multi-scale discriminator: consecutive and long-term patterns

End-to-End Models

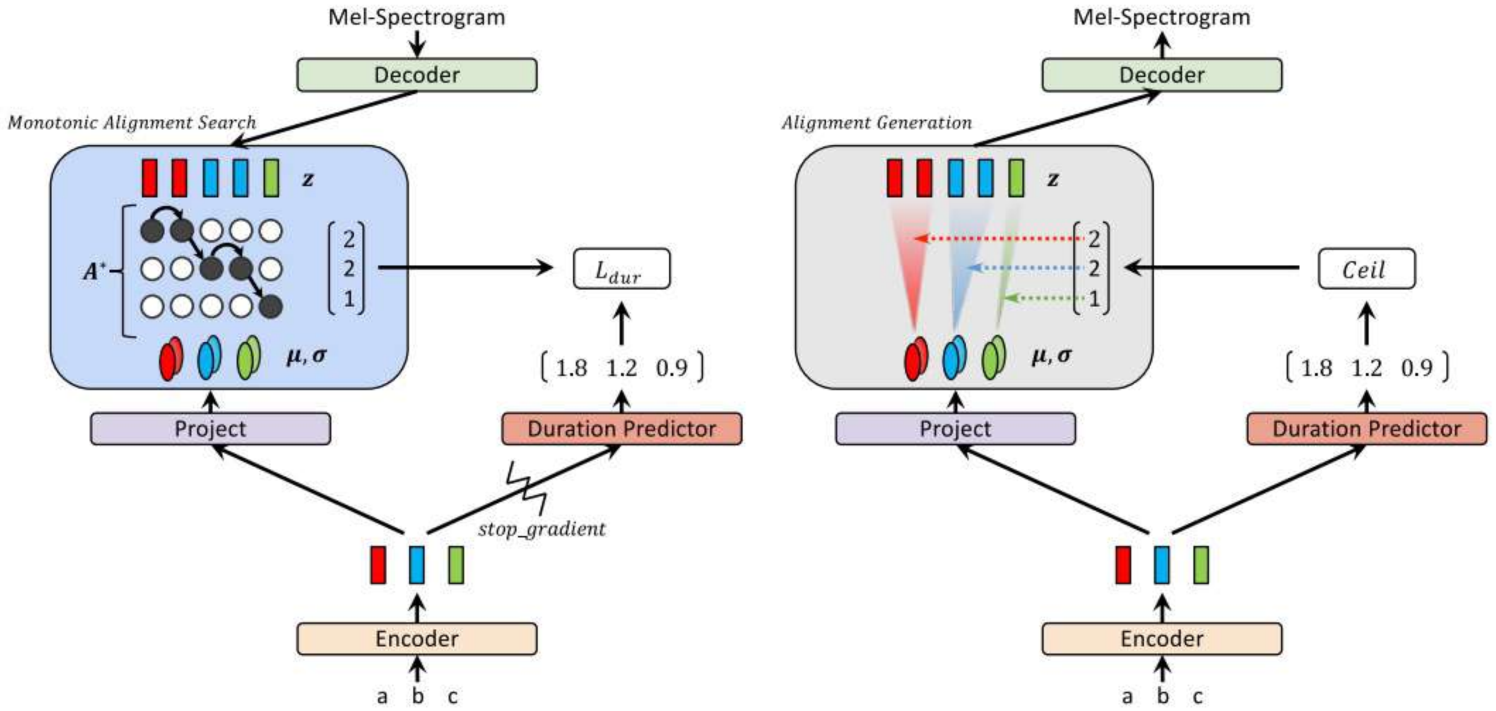
Flow-based Models

- Normalizing flows models the target distribution by transforming a simple distribution through a sequence of invertible mappings
- Both the forward and inverse transformations can be easily computed
- Flow-based models learn a mapping between a low-dimensional latent space and the high-dimensional space of speech waveforms



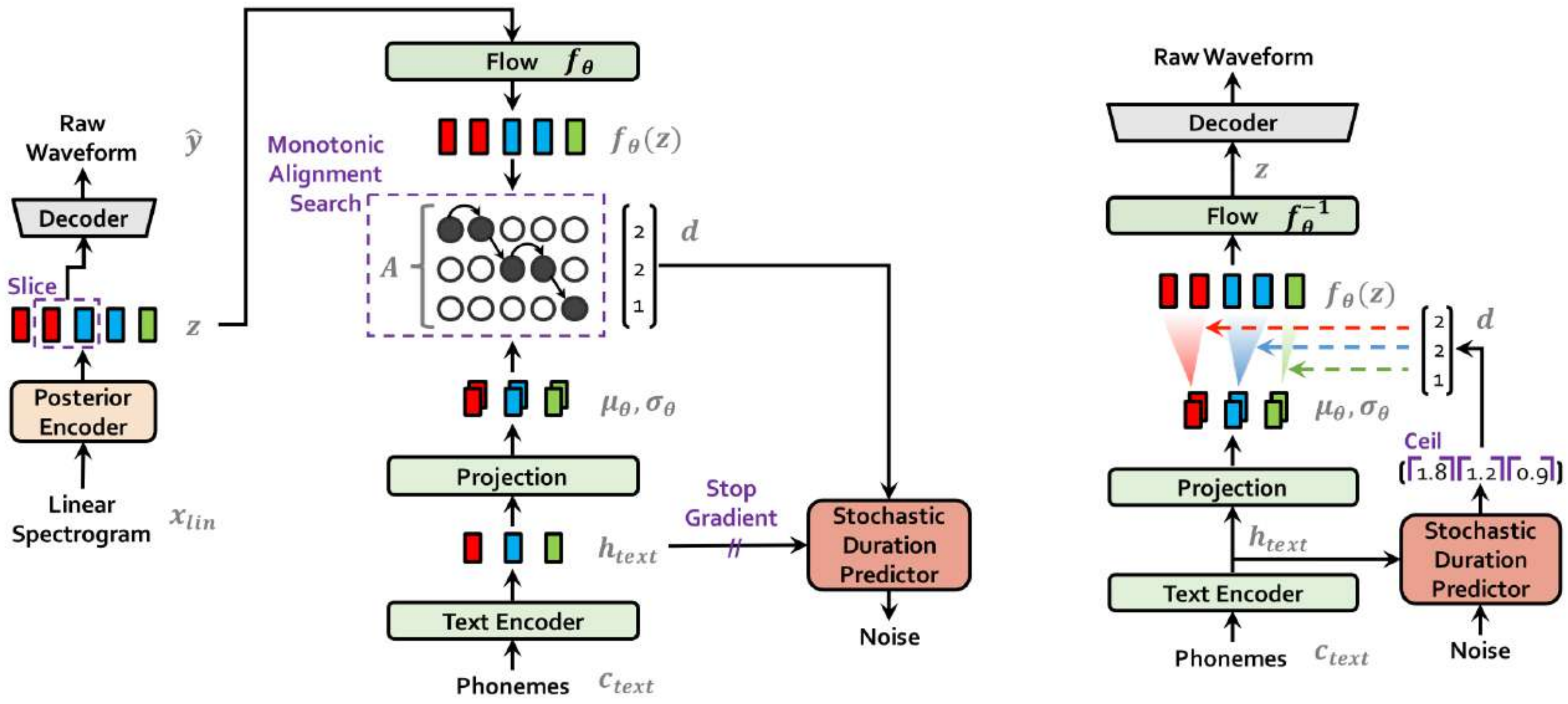
GlowTTS

- Searches for the most probable monotonic alignment between text and the latent representation of speech
- Monotonic alignments provides robustness and generalizes to long utterances
- Flows enable fast, diverse, and controllable speech synthesis.



(a) An abstract diagram of the training procedure. (b) An abstract diagram of the inference procedure.

Figure 1: Training and inference procedures of Glow-TTS.



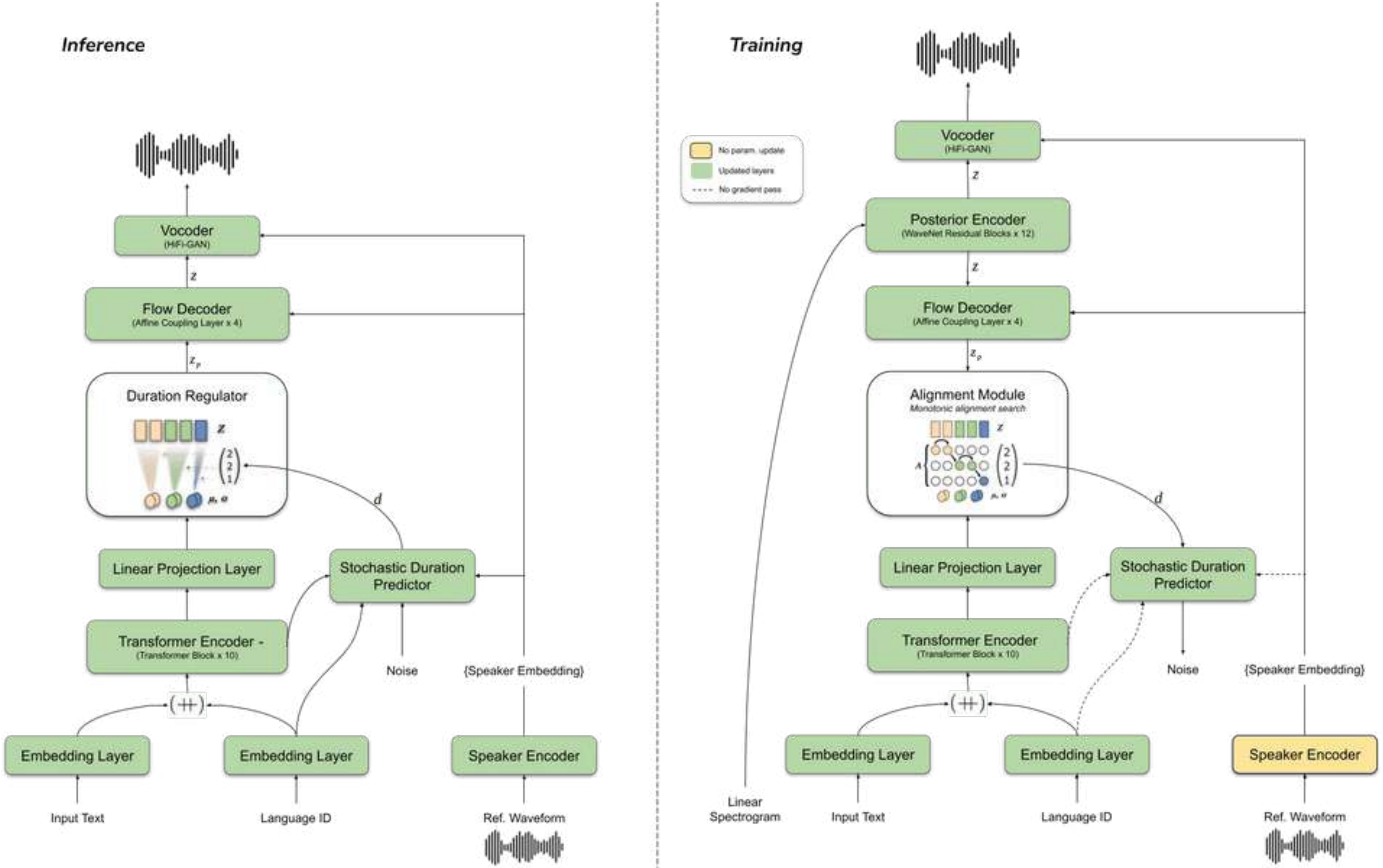
(a) Training procedure

(b) Inference procedure

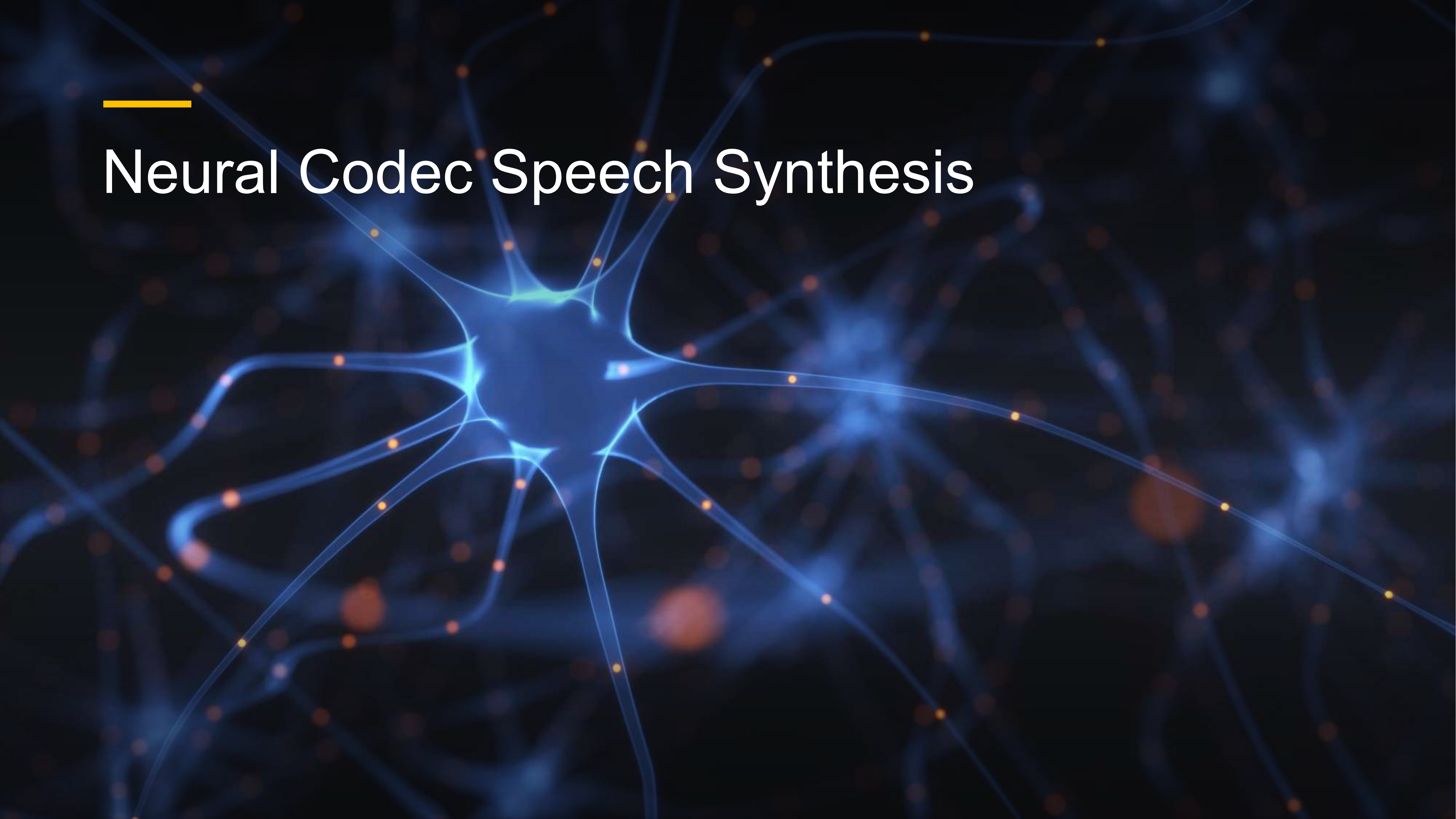
VITS

- Combination of GlowTTS with HiFiGAN vocoder
- Monotonic alignment search
- Inference runs x67 real-time

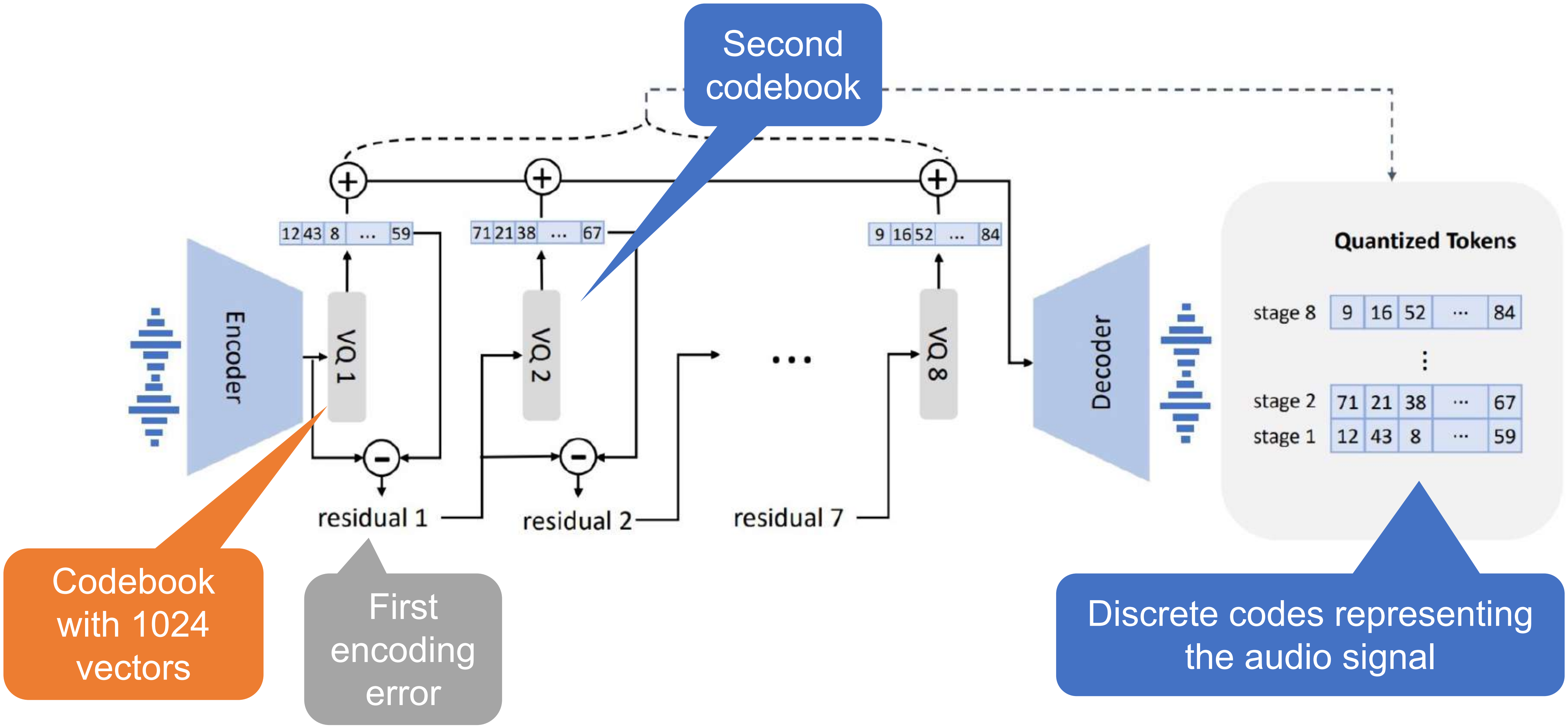
YourTTS Architecture



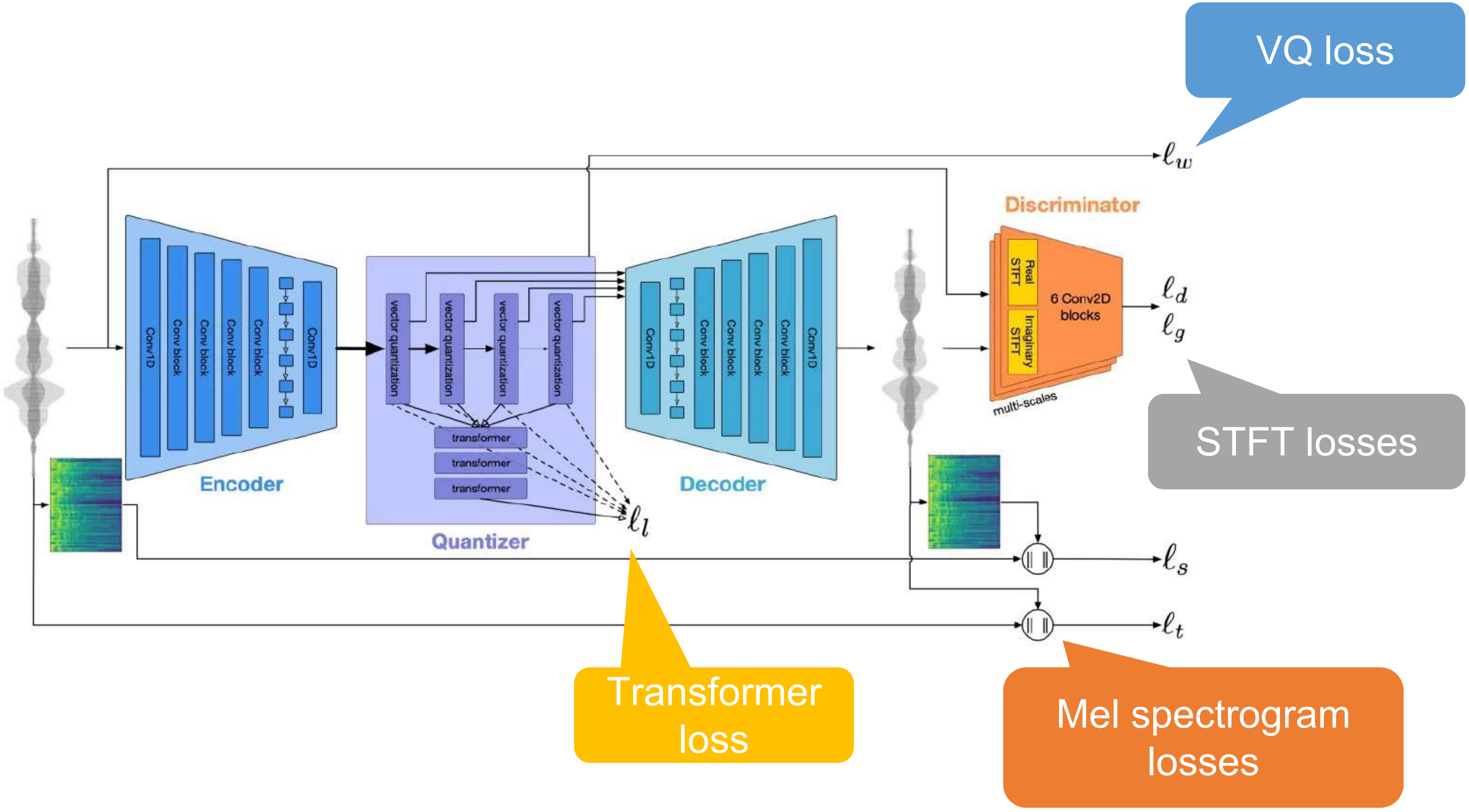
Neural Codec Speech Synthesis



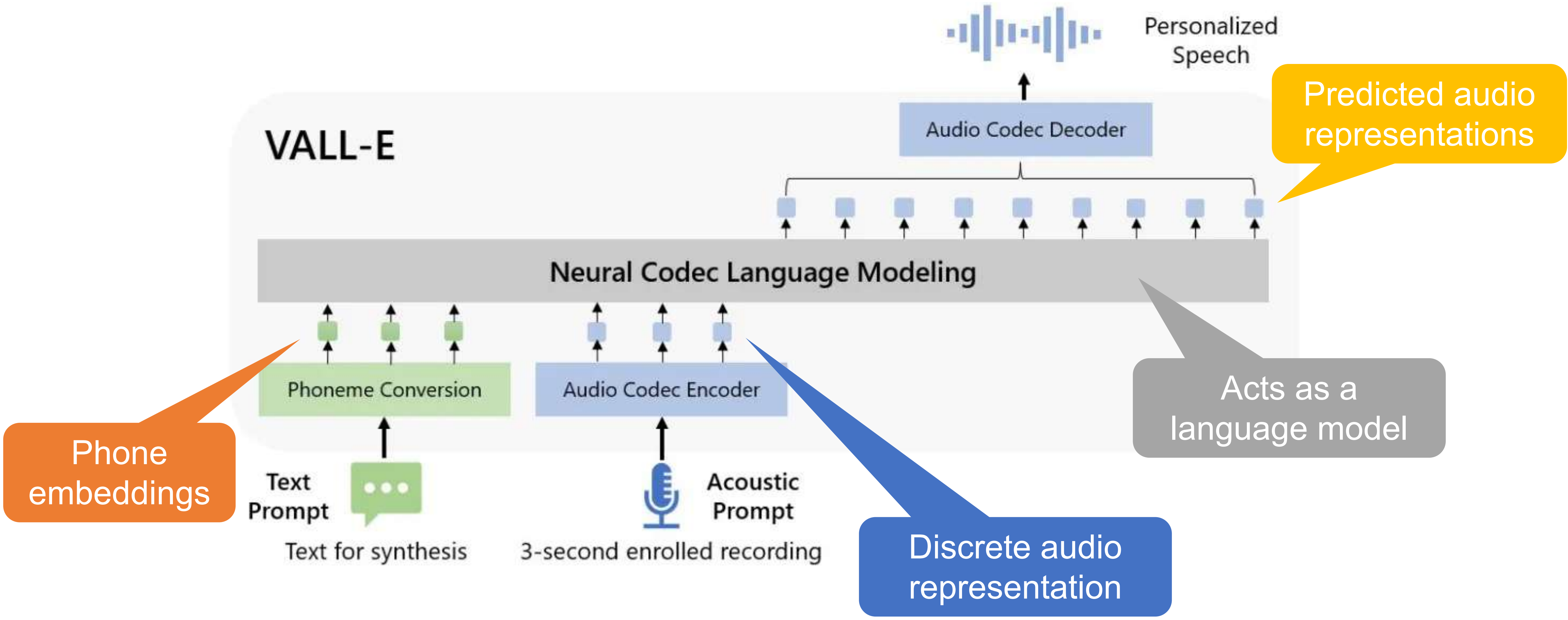
Encodec Model



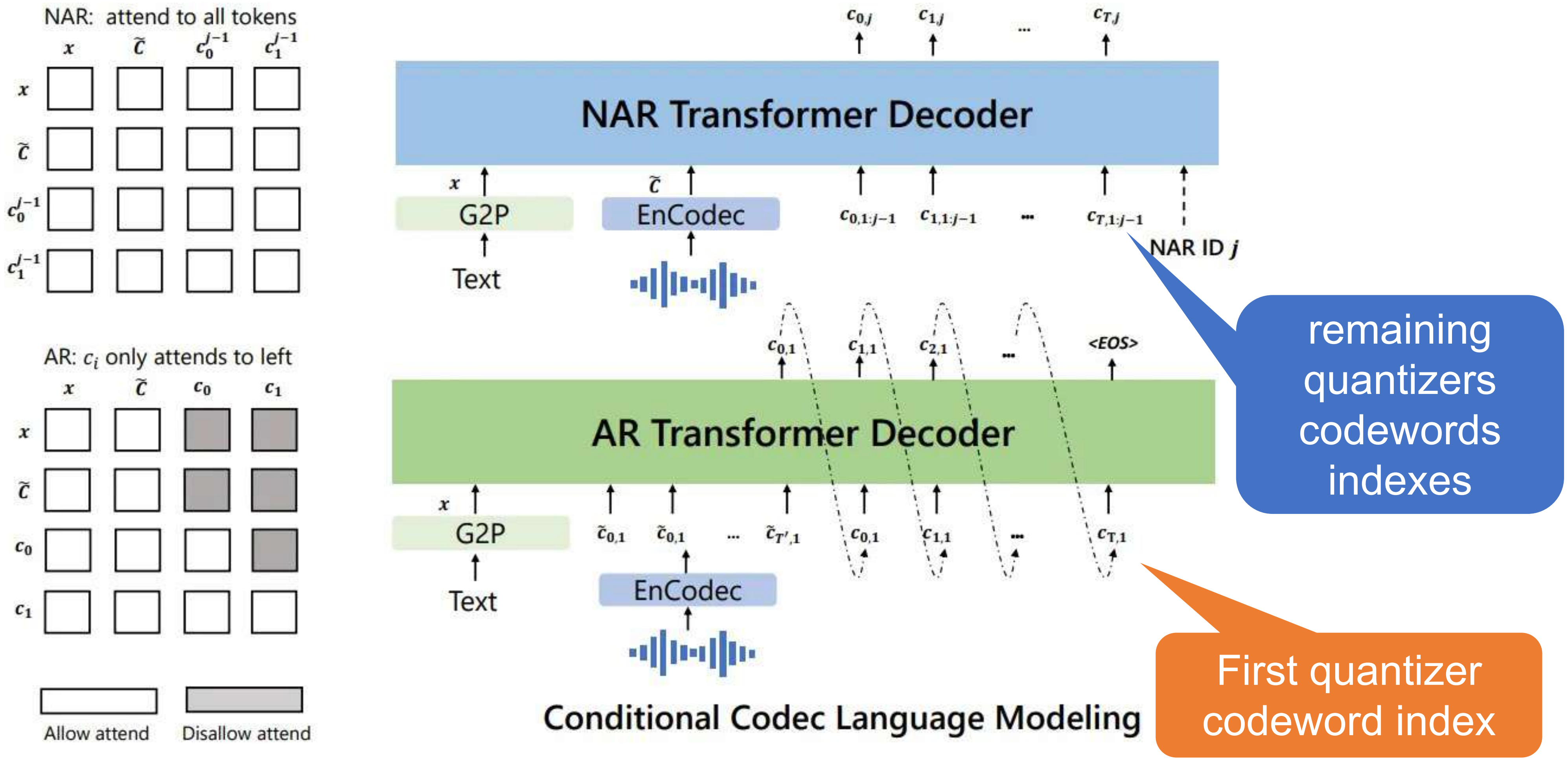
Encodec Model Training



VALL-E Architecture



VALL-E Neural Codec Language Model



VALL-E Equations

first audio codeword index

phone embeddings

auto regressive (AR)

$$p(\mathbf{c}_{:,1}|\mathbf{x}, \tilde{\mathbf{C}}_{:,1}; \theta_{AR}) = \prod_{t=0}^{T} p(\mathbf{c}_{t,1}|\mathbf{c}_{<t,1}, \tilde{\mathbf{c}}_{:,1}, \mathbf{x}; \theta_{AR})$$

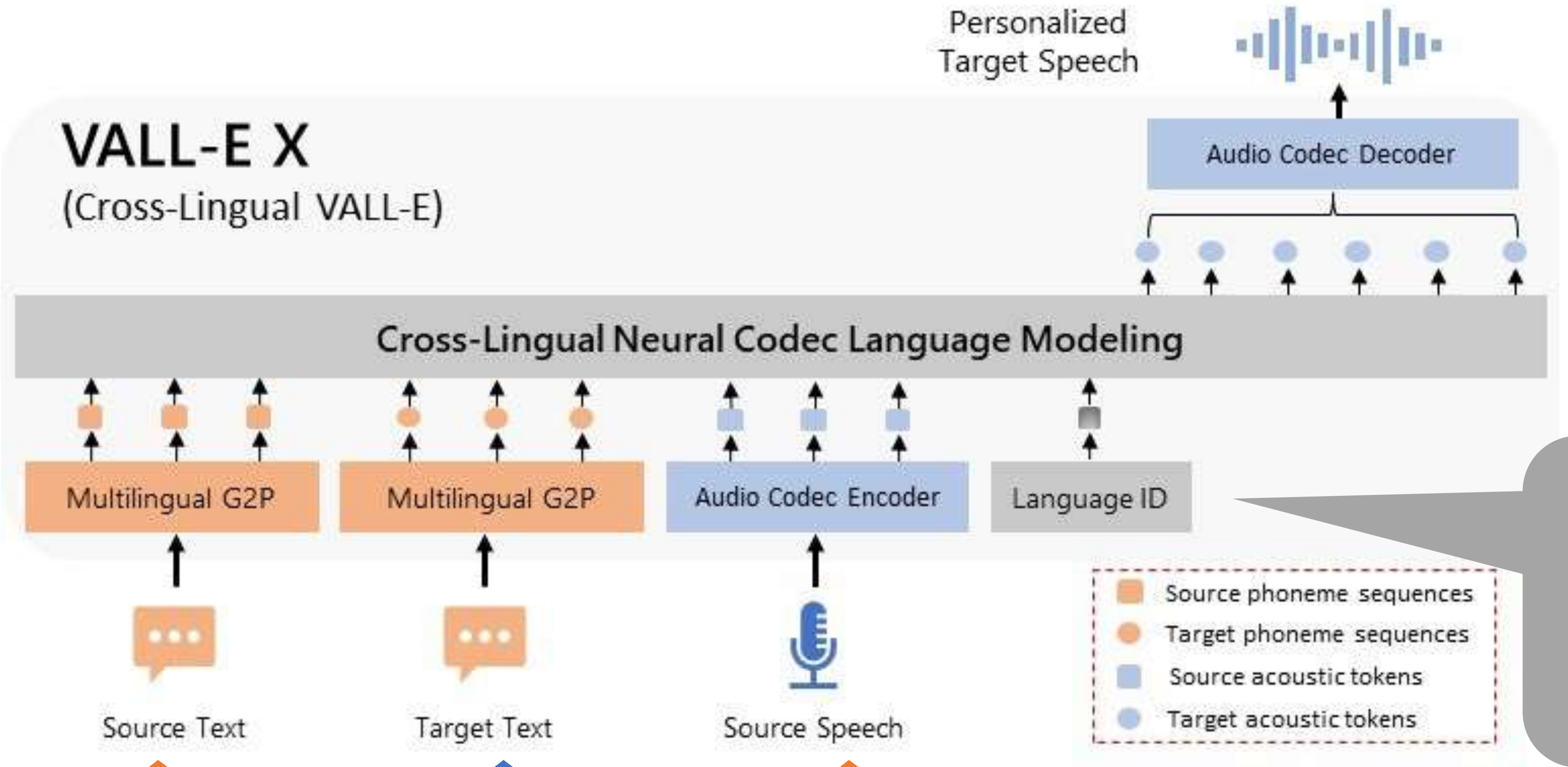
acoustic prompt

$$p(\mathbf{C}|\mathbf{x}, \tilde{\mathbf{C}}; \theta) = p(\mathbf{c}_{:,1}|\tilde{\mathbf{C}}_{:,1}, \mathbf{X}; \theta_{AR}) \prod_{j=2}^{8} p(\mathbf{c}_{:,\mathbf{j}}|\mathbf{c}_{:,<j}, \mathbf{x}, \tilde{\mathbf{C}}; \theta_{NAR})$$

remaining indexes

non auto regressive (NAR) combines the 7 indexes

VALL-E X



source language

target language

source language

controls the speaking style of generated speech

Summary

Technologies

- Articulatory, formants, concatenative, statistical parametric SS, neural SS

Evaluation

- Subjective and objective tests

Probabilistic Formulation

- Acoustic model, acoustic features, linguistic features, pipeline

Front End

- Text normalization, POS tagging, prosody prediction, G2P conversion

Summary (cont.)

Acoustic Model

- Generates the intermediate spectrogram with SPSS, RNN or transformers. Attention vs duration.

Waveform Generation

- LPC, Griffin-Lim and WaveNet vocoders

Speaker and Style Embeddings

- Latent space and global style tokens

End-to-End Models

- Neural TTS Systems, Zero-shot TTS, VALL-E

Obrigado

TÉCNICO LISBOA