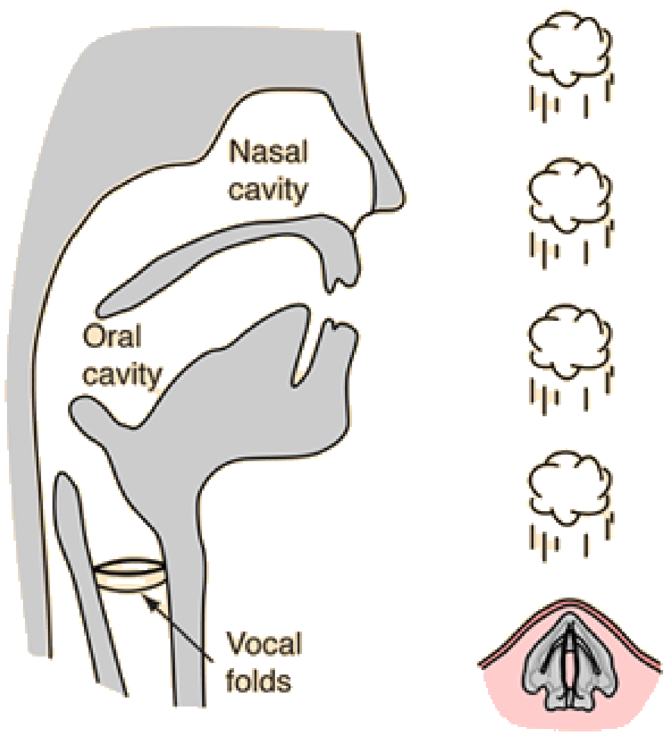
# Speech Synthesis

Instructor: Tom Ko

# Objectives

- Learn the basic mechanism of speech synthesis / text-to-speech (TTS)

# Speech Production System

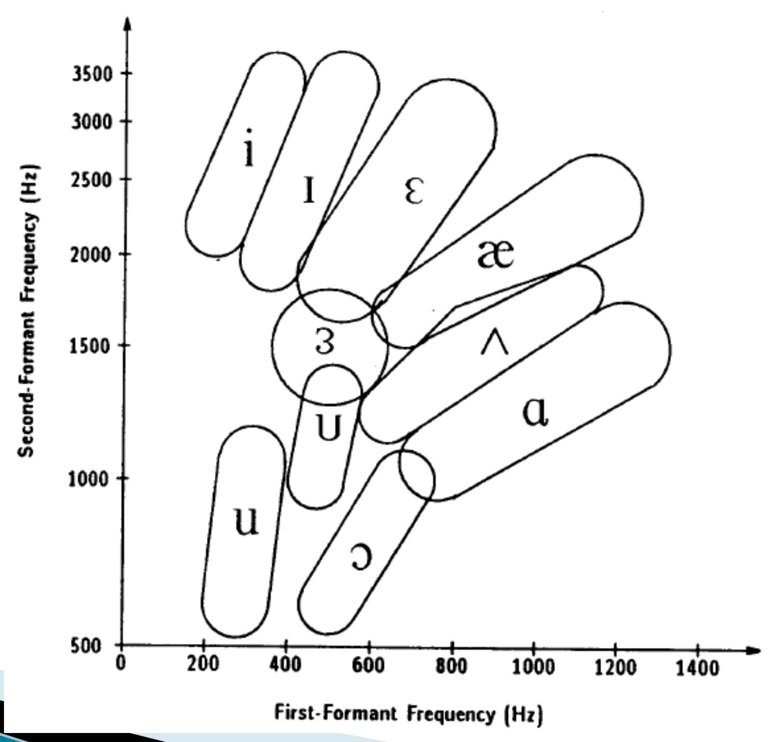



Schematic View of Vocal Tract

- The vocal folds generate periodic impulses.
- The vocal tract acts like a filter of which the impulse response convolutes with the impulses to form the sound.
- The impulse response changes with the shape of the tract.
- Production-based features encode the shape of vocal tract from the signal.

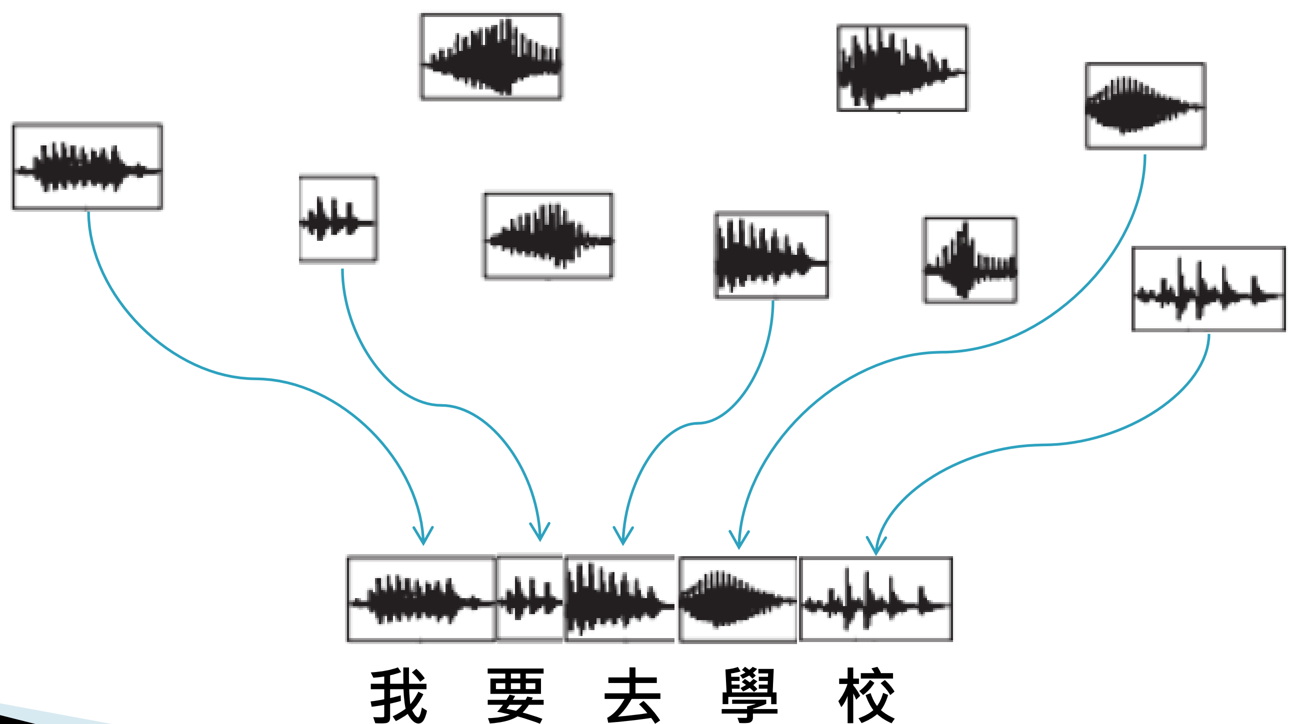
# Formant synthesis (~'90s)

- Rule-based
- Hand-crafting each phonetic units by rules
- Based on source-filter model



# Concatenation approach ('90s~)

- Data-based
- Concatenate speech units (waveform) from a database



# Parametric approach (mid–'90s)

- Data–based
- Need training
- Statistical acoustic model

- Training

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

- Synthesis

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

$\boldsymbol{\lambda}$ : model parameters

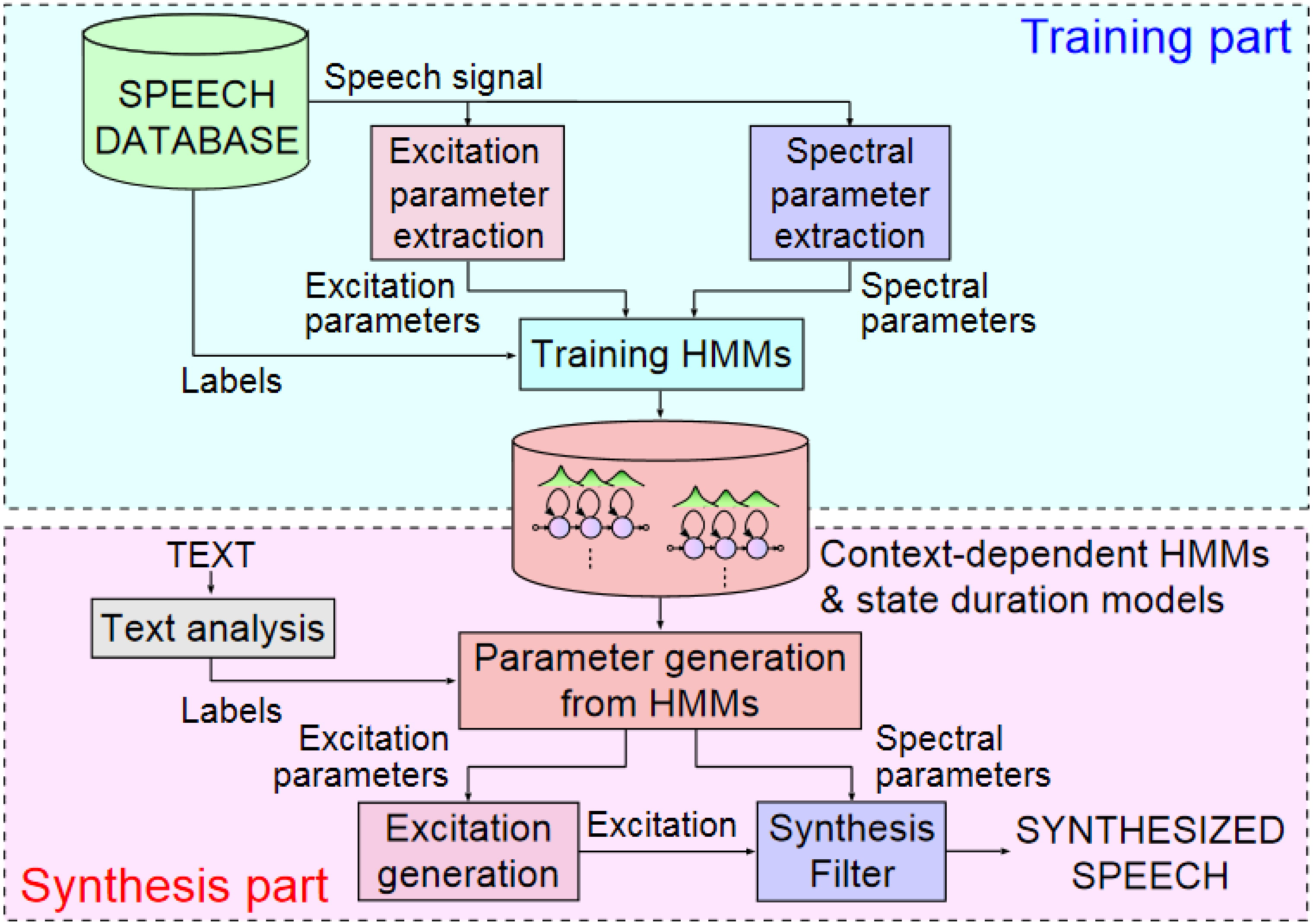
$\boldsymbol{O}$ : training data

$\mathcal{W}$ : transcriptions

$\boldsymbol{o}$ : synthesized speech

$w$ : input text

# HMM-based speech synthesis




Pictures from
http://www.sp.nitech.ac.jp/~tokuda/tokuda_interspeech09_tutorial.pdf

# DNN–HMM approach

- IDLAK toolkit (a reverse of KALDI)
- Frame–based
- It extracts several kinds of feature from the waveform
  - Mel–Cepstrum (MCEP)
  - F0 pitch
  - Band Aperiodicity feature (BNDAP)

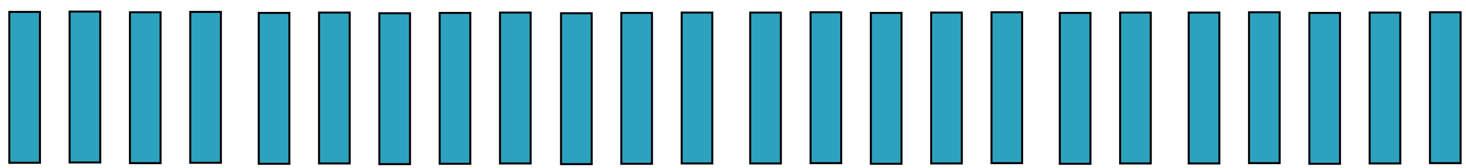
# DNN-HMM approach (training)

***ay m t aa m ay m hh ah ng g r iy***

I'm Tom, I'm hungry.

Feature Extraction

Features:



Alignment Estimation

***ay ay m m t t t aa aa m ay m m m hh ah ah ng ng g g r r iy***

Train a regression network

# DNN-HMM approach (testing)

Hello, How are you?

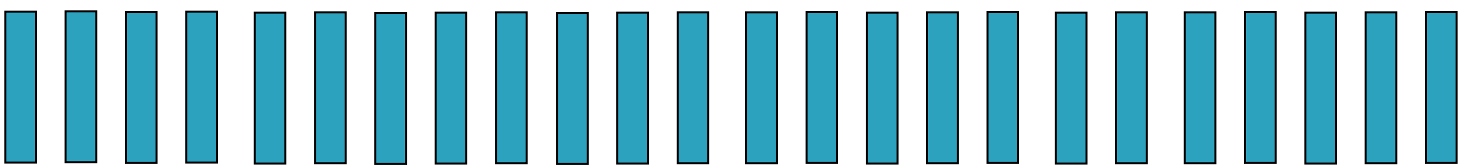
*hh ah l ow hh aw aa r y uw*

Duration Estimation

*hh hh hh ah ah l l l ow ow hh hh aw aw aa r r y y uw uw*

Feature Prediction

Features:



Vocoder

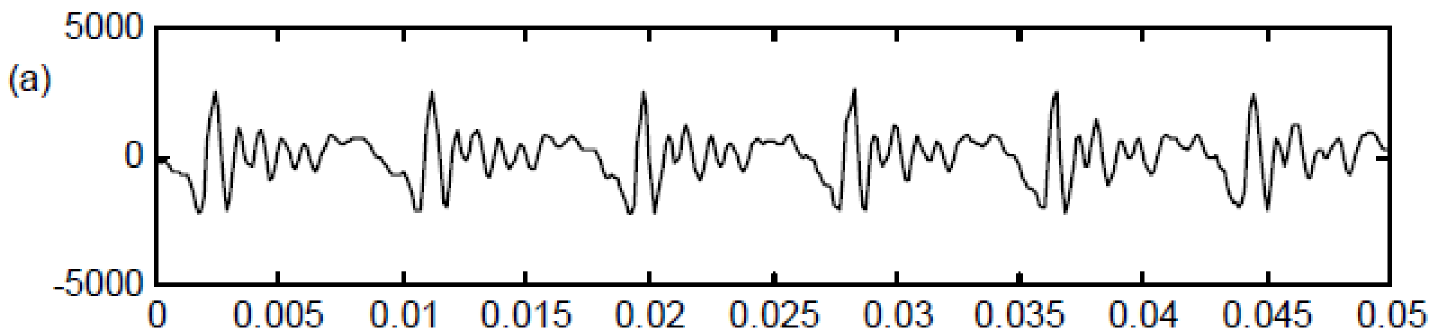
# Major components

- Two major components in TTS
- The acoustic model
  - Convert phoneme sequence into features
- Vocoder
  - Convert features into speech waveform

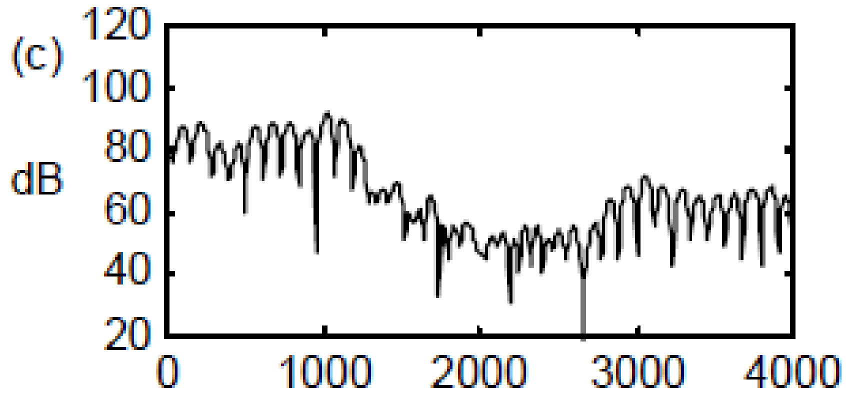
- In modern TTS systems, spectrogram is used as the features.

# Frequency spectrum

- A signal in time domain:



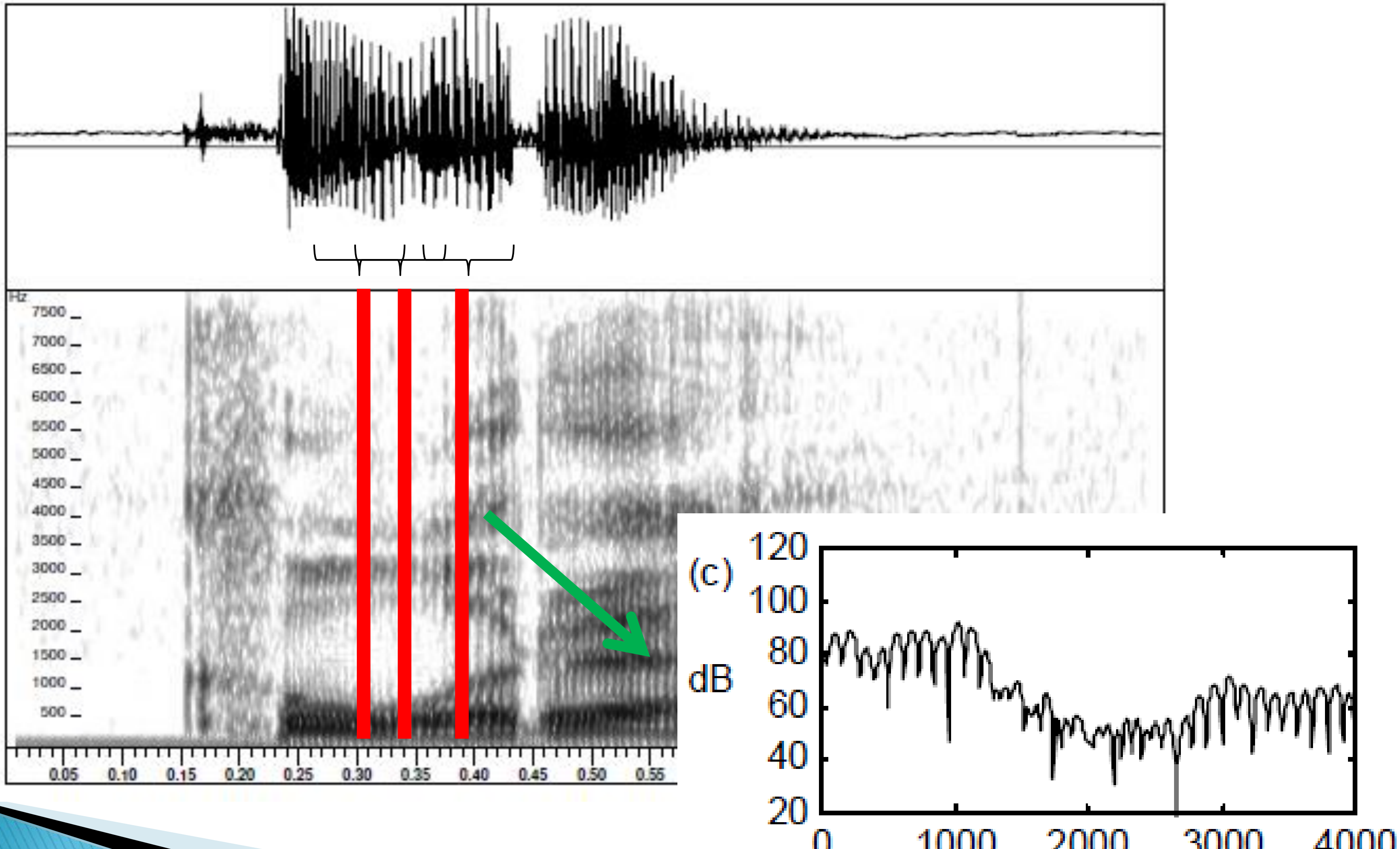

- Its form in frequency domain:



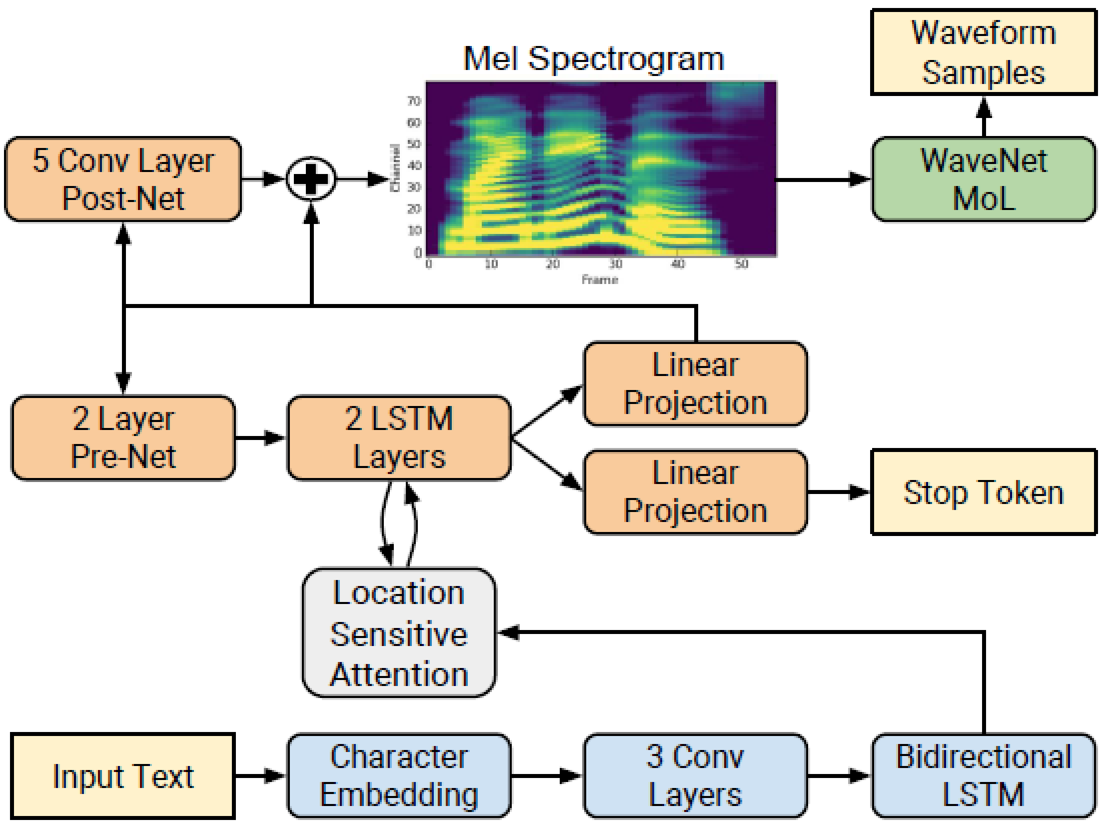
# Short-time Fourier Analysis

- We have dealt with periodic signals in our formulation, however, the signal is no longer periodic when longer segments are analyzed.
- **Short-time analysis**: a speech signal is decomposed into a series of short segments.
- In each segment, the signal is assumed to be *stationary.*
  - The region has to be short enough

# Short-time Fourier Analysis



# Tacotron 2

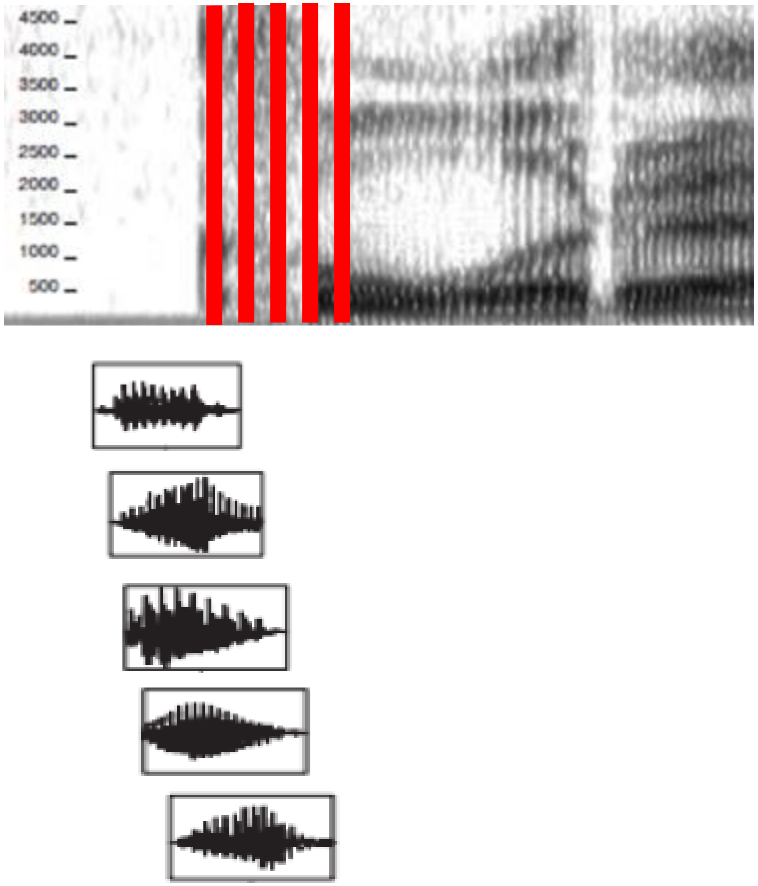


Mel Spectrogram
Waveform Samples
5 Conv Layer Post-Net
WaveNet MoL
Channel
Frame
Linear Projection
2 Layer Pre-Net
2 LSTM Layers
Linear Projection
Stop Token
Location Sensitive Attention
Input Text
Character Embedding
3 Conv Layers
Bidirectional LSTM

# Vocoder

- Algorithm-based
  - No need to learn
  - Griffin-Lim
- Neural-based
  - It treats the conversion from spectrogram to audio signal as a sequence to sequence problem
  - WaveNet

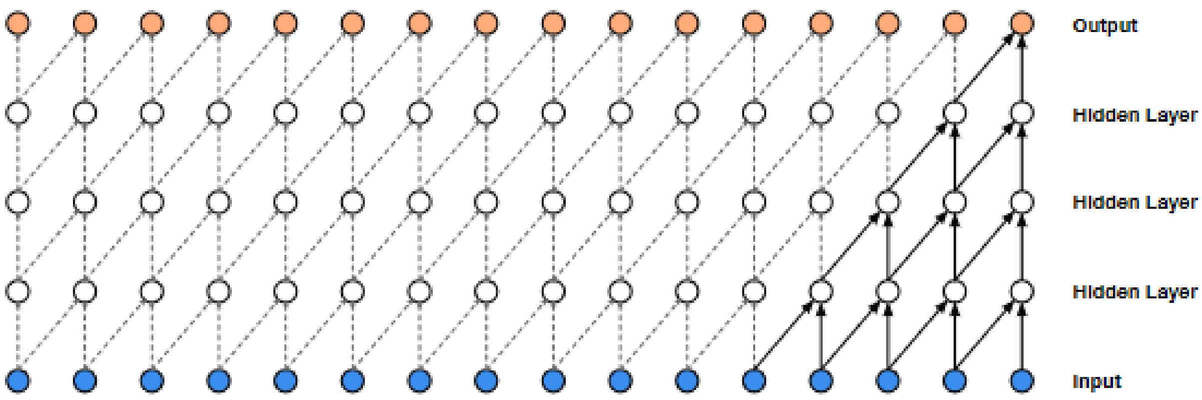
# Algorithm-based vocoder




- Introducing strong artifacts

# Wavenet

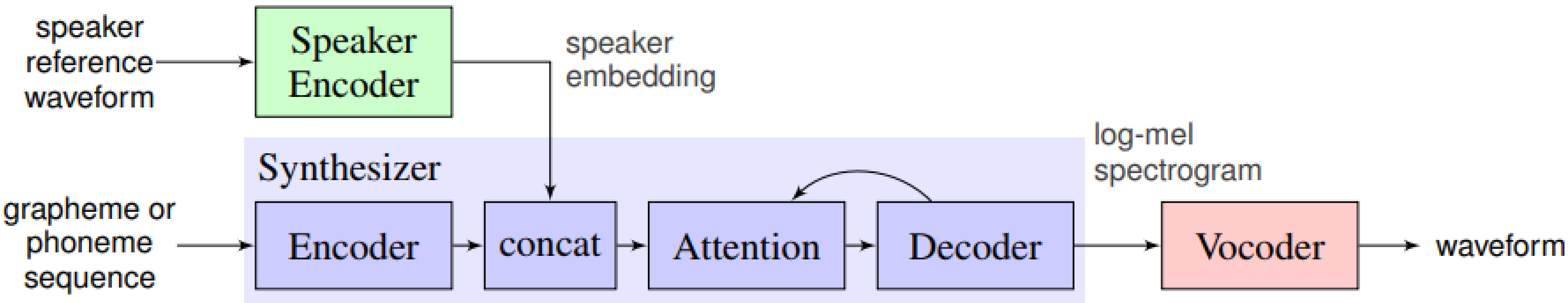
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$



# Research directions

- Single speaker TTS
  - Require 20-30 hours labeled data from a single speaker
- Multiple speaker TTS
  - Require data from ~100 speakers, 20 mins labeled data from each speaker
- Unseen speaker TTS
  - Require data from multiple speakers

# Speaker embedding approach



# Reading list

- Tacotron2 paper
  - "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions", https://arxiv.org/abs/1712.05884
- Wavenet paper
  - "WaveNet: A Generative Model for Raw Audio", https://arxiv.org/abs/1609.03499