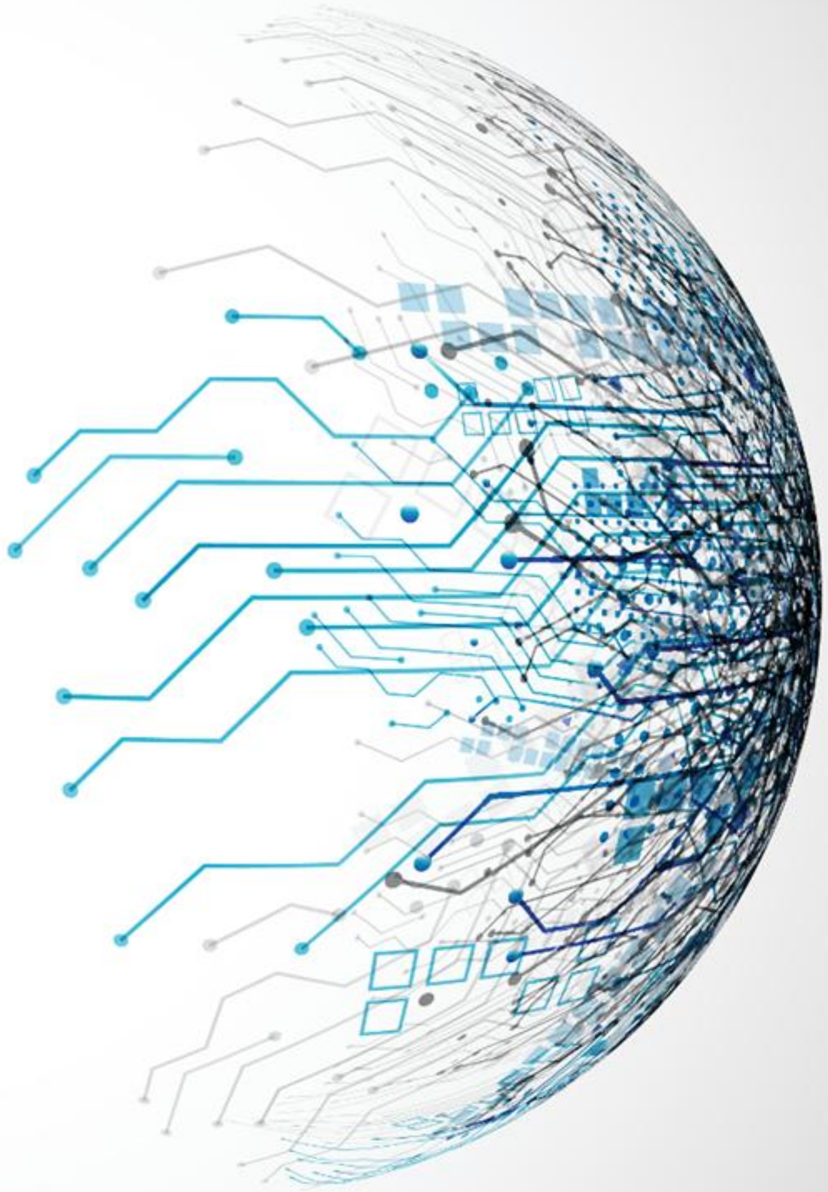
Deep Learning

Speech Technology and Text-to-Speech



Dr. Mohammed Salah Al-Radhi

BME-VIK-TMIT

12 November 2024

Copyright

Outline

1. Speech Technology
2. Speech Synthesis
3. Text-to-Speech (TTS)
4. Advances in TTS

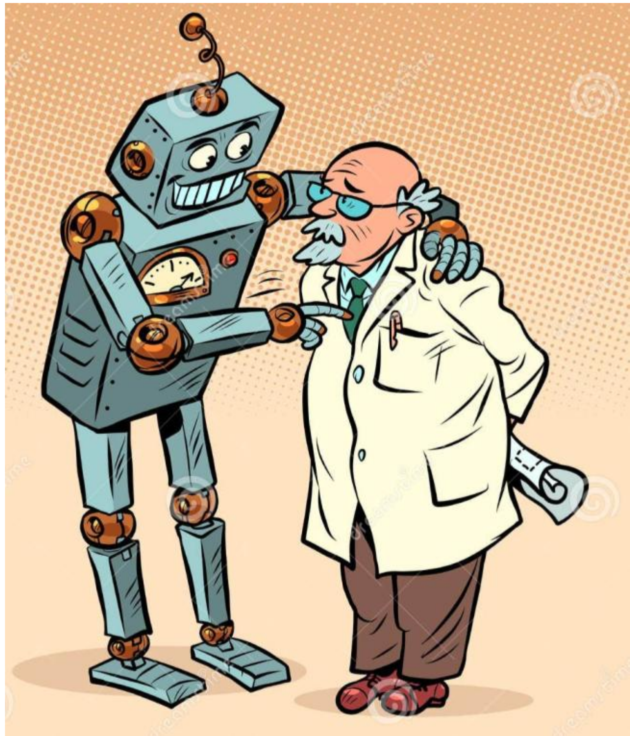
Speech Technology

Speech is great

- No baby learns from text
- No baby learns without communicative intent

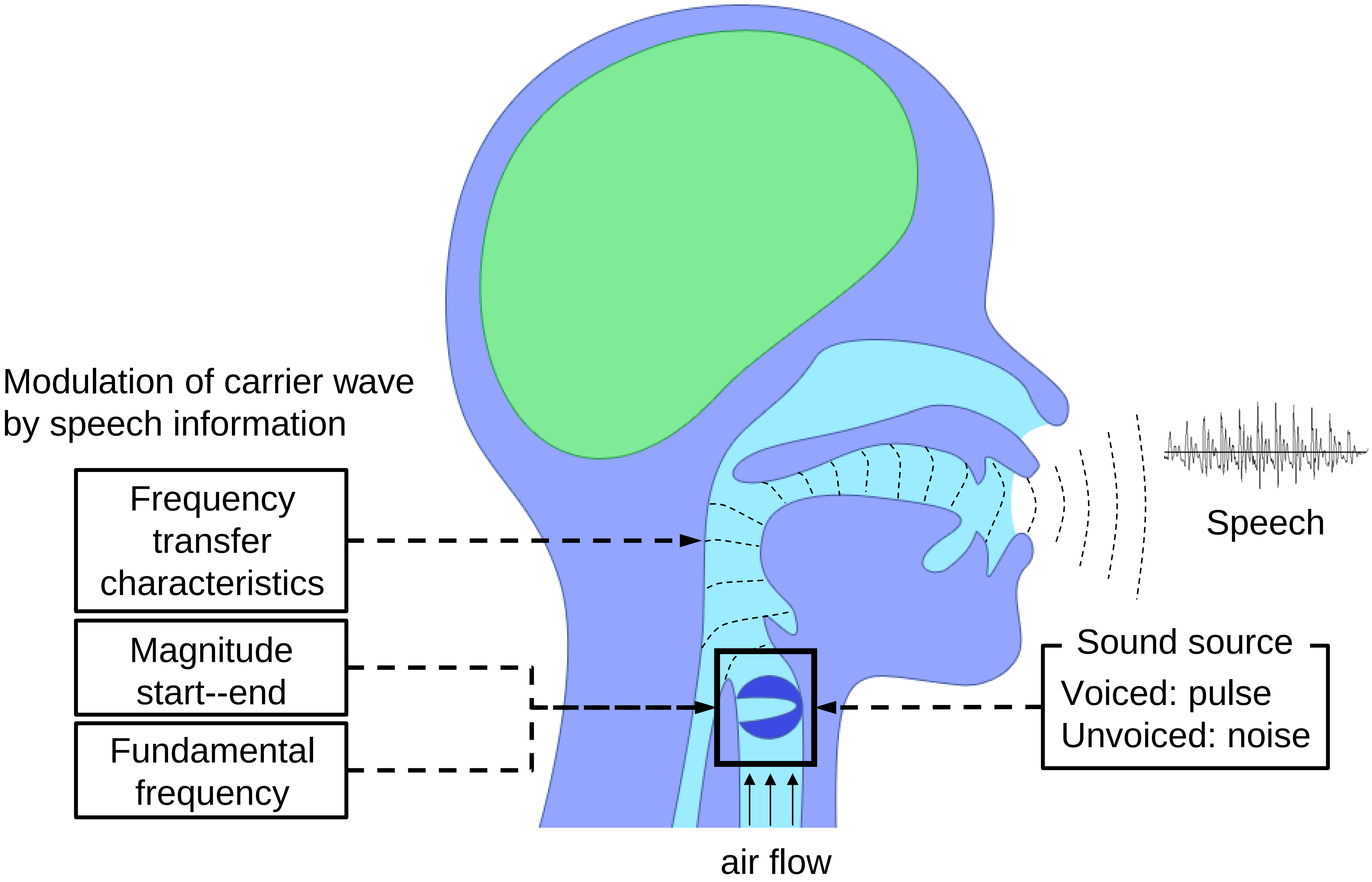
Speech is great

- Less complex than vision
- Continuous data (as opposed to image and text)



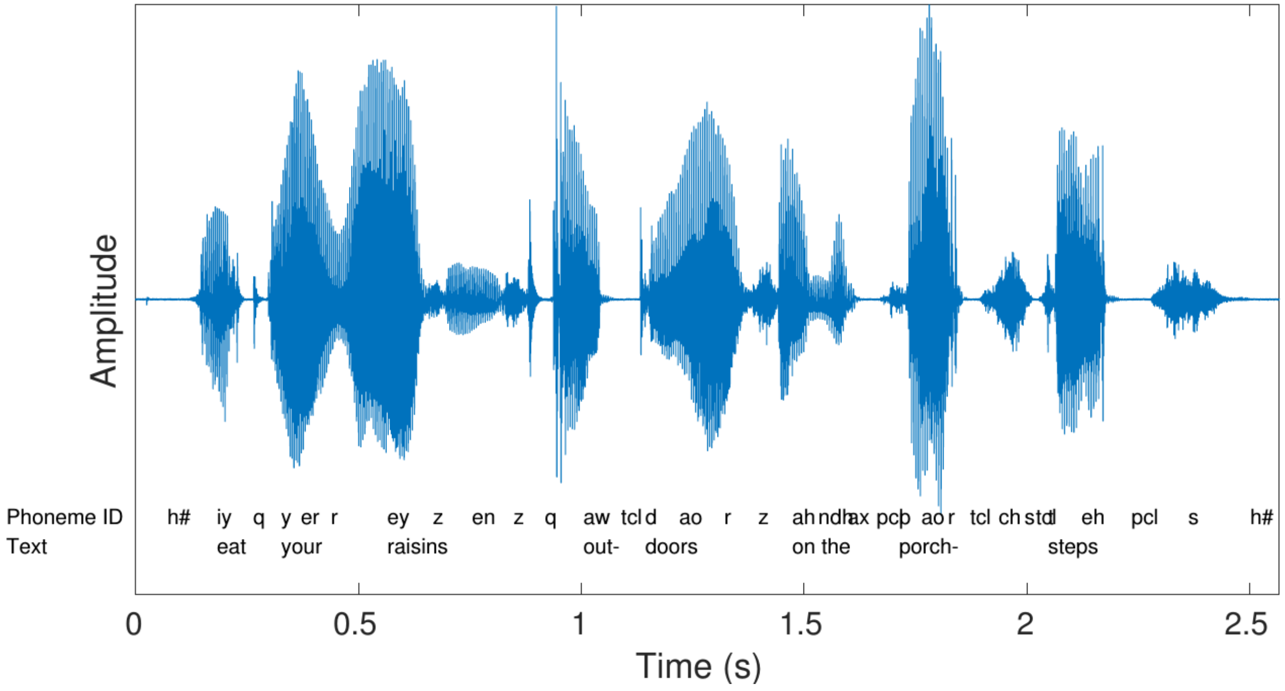
offers a more interaction with machines ☺

Speech Production Mechanism



Why Speech Processing?

- ❑ model and manipulate the speech signal to be able to:
  - transmit (code) speech efficiently
  - produce natural speech synthesis
  - recognize the spoken word



- ❑ speech is the natural form of communication between humans; it reflects a lot of the variability and complexity of humans!

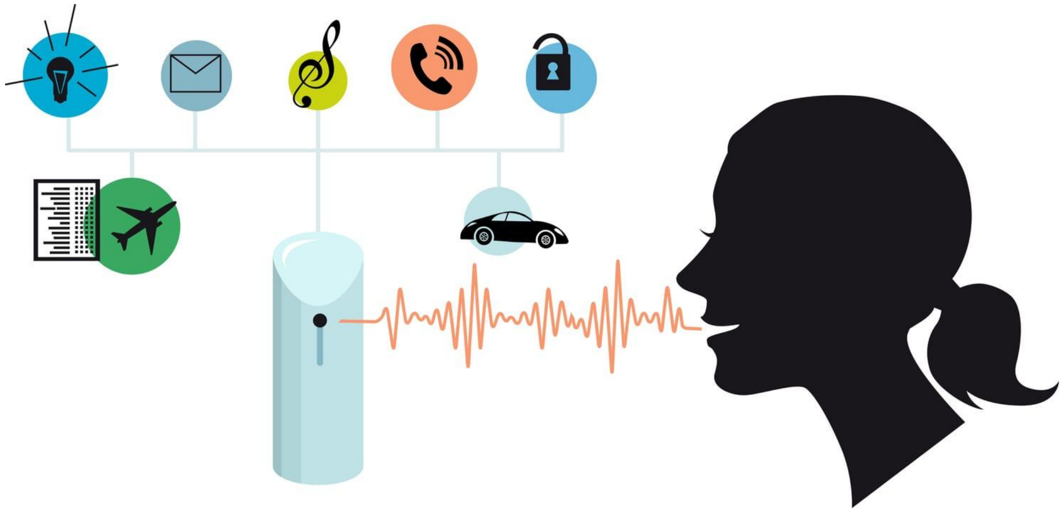
Intelligent Speech Technology

Enabling machines to ”listen & speak”

- **Speech Synthesis:** Converting text to speech → Installing artificial mouth for computers

- **Speech recognition:** Converting speech to text and recognize speech content, speaker, language and other information → Installing artificial ear for computers

- **Cognitive intelligence:** Understanding and Thinking Speech evaluation, Machine translation, Smart Customer Service



Speech Processing Applications

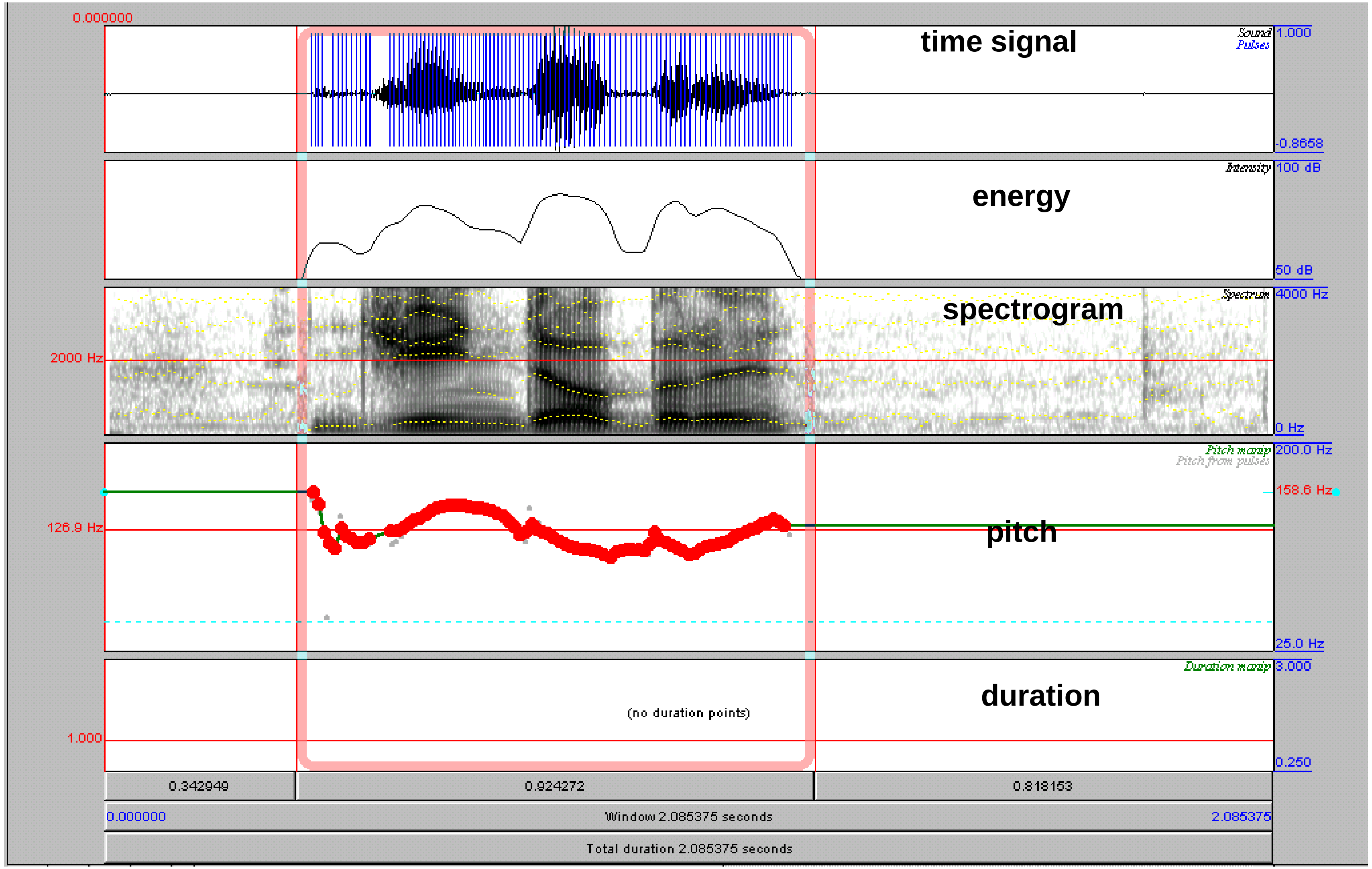
- ❑ **Human - Machine Communication**
  - Siri

- ❑ **Machine - Human Communication**
  - Toshiba / Cambridge Talking Head

- ❑ **Human - Human Communication**
  - speech coding (reduction in bit-rate/storage)
  - speech enhancement (removal of noise)
  - Voice Morphing, or voice transformation or voice conversion
  - speech translation aids for disabled

Speech Waveform

- non-stationary
- pseudo-periodic
- random components

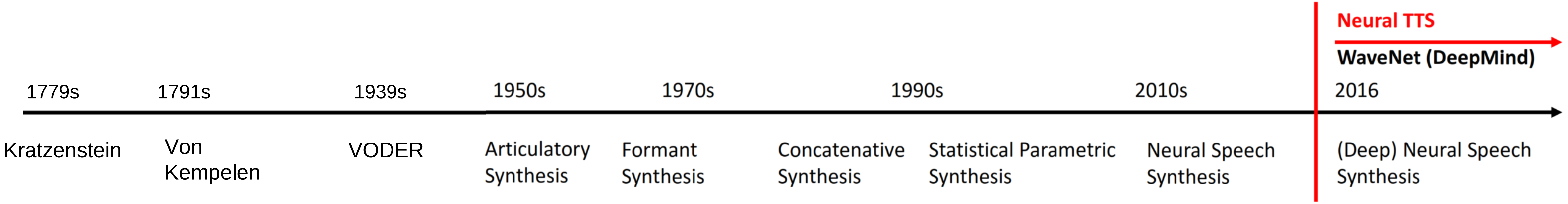


https://praat.en.softonic.com/

Speech Synthesis

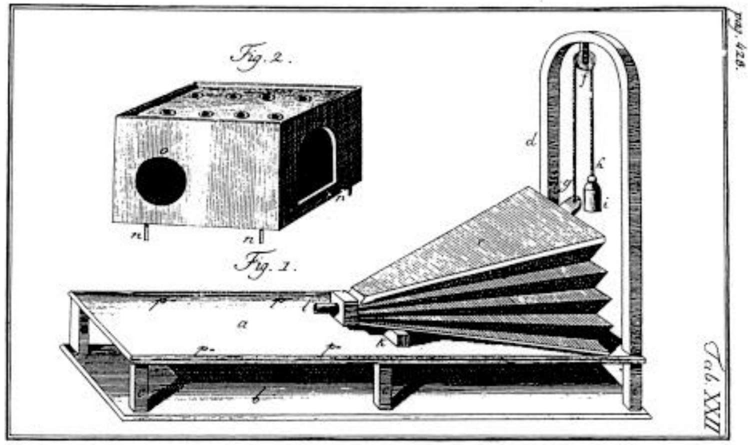
What is the Speech Synthesis?

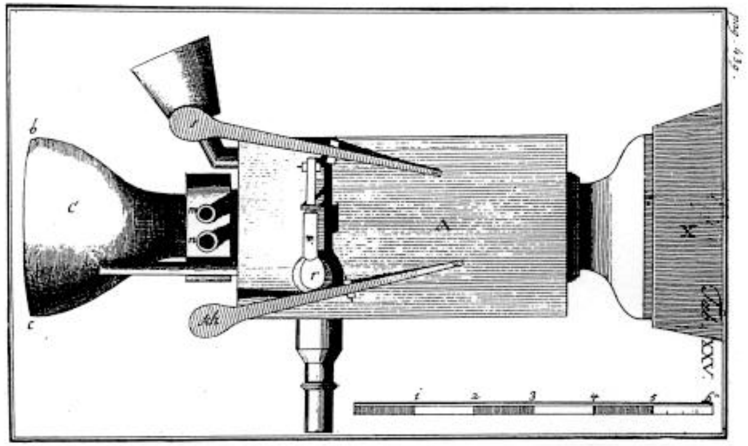
Speech synthesis is the artificial production of human speech that sounds almost like a human voice and is more precise with pitch, speech, and tone.

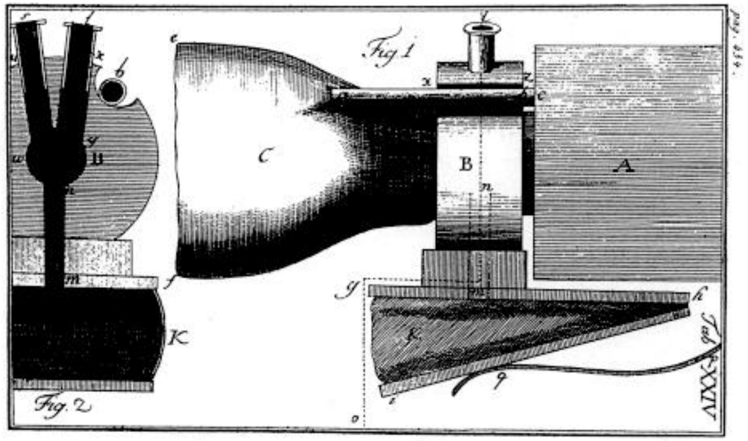


Von Kempelen: 1791

https://www.bme.hu/hirek/20200918/Az_informatika_koraban_is_el_Kempelen_Farkas_szellemisege

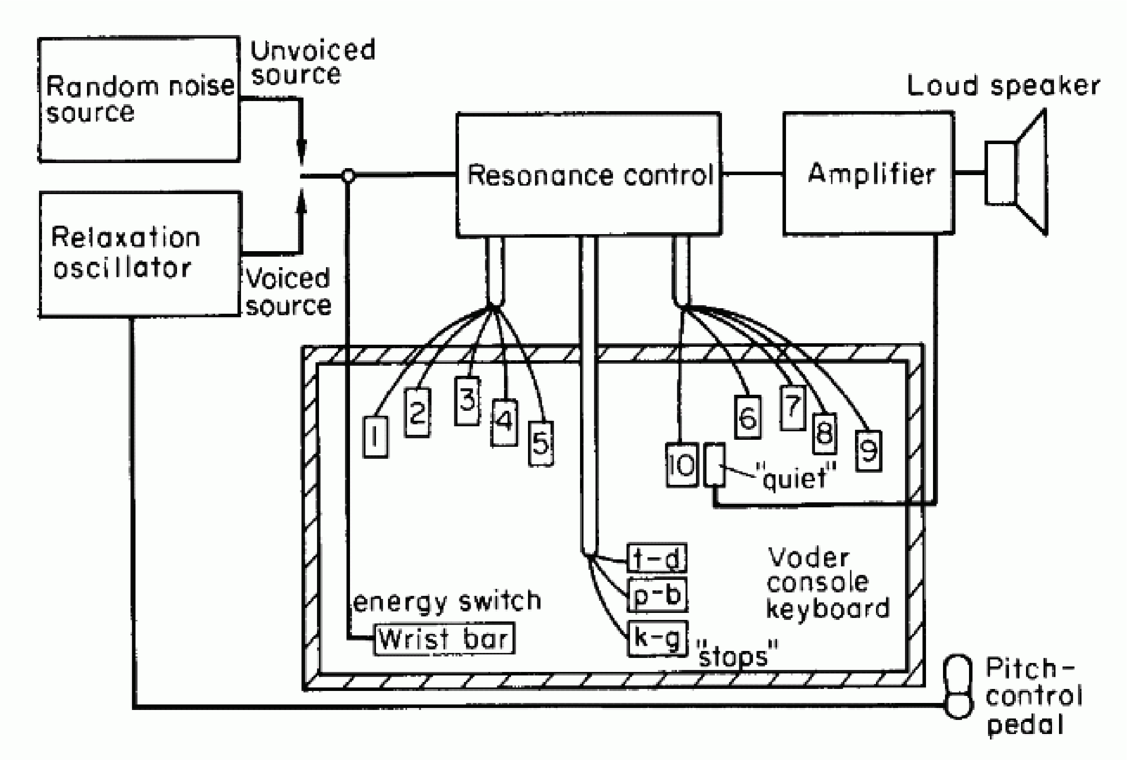




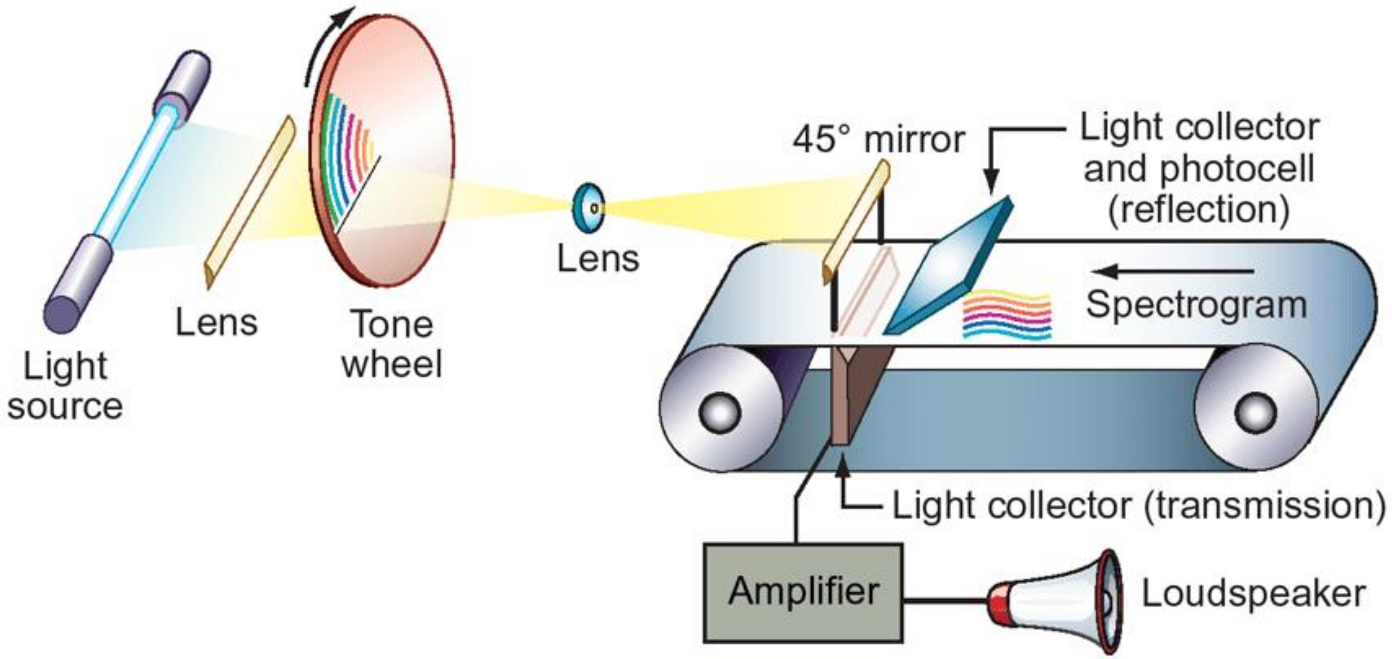


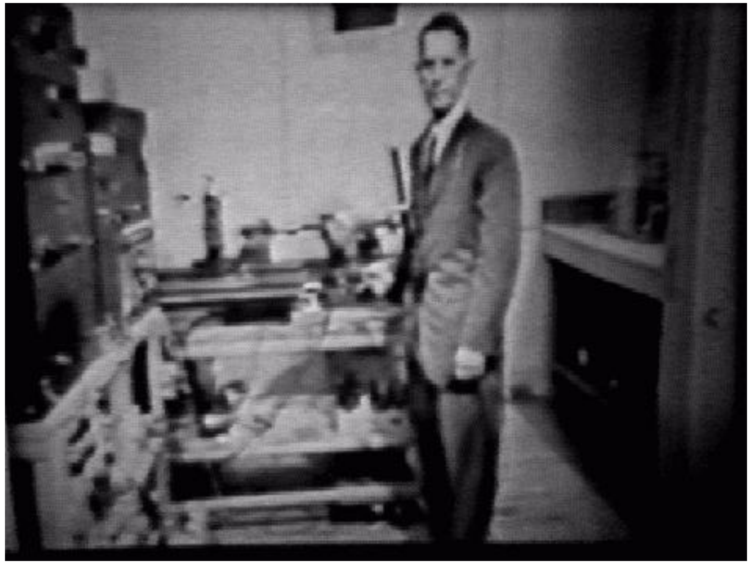
Homer Dudley's VODER: 1939

- World's Fair
- Manually controlled through complex keyboard



Cooper's Pattern Playback: 1949





https://120years.net/pattern-playback-franklin-s-cooper-usa-1949/

Gunnar Fant's OVE Synthesizer: 1953

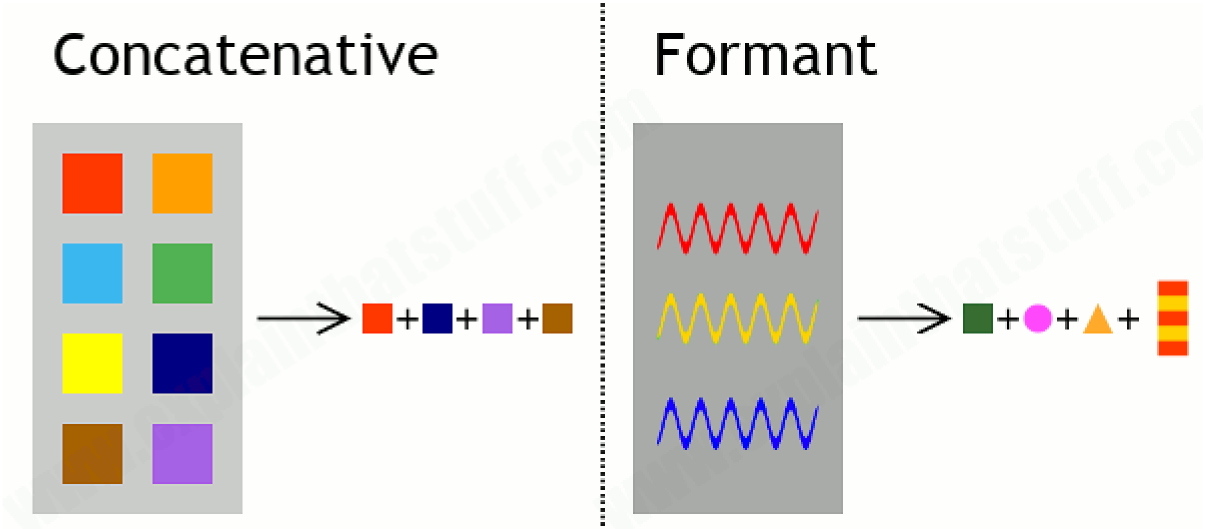
- Of the Royal Institute of Technology, Stockholm
- Formant Synthesizer for vowels
- F1 and F2 could be controlled

What Uses Does Speech Synthesis Have?

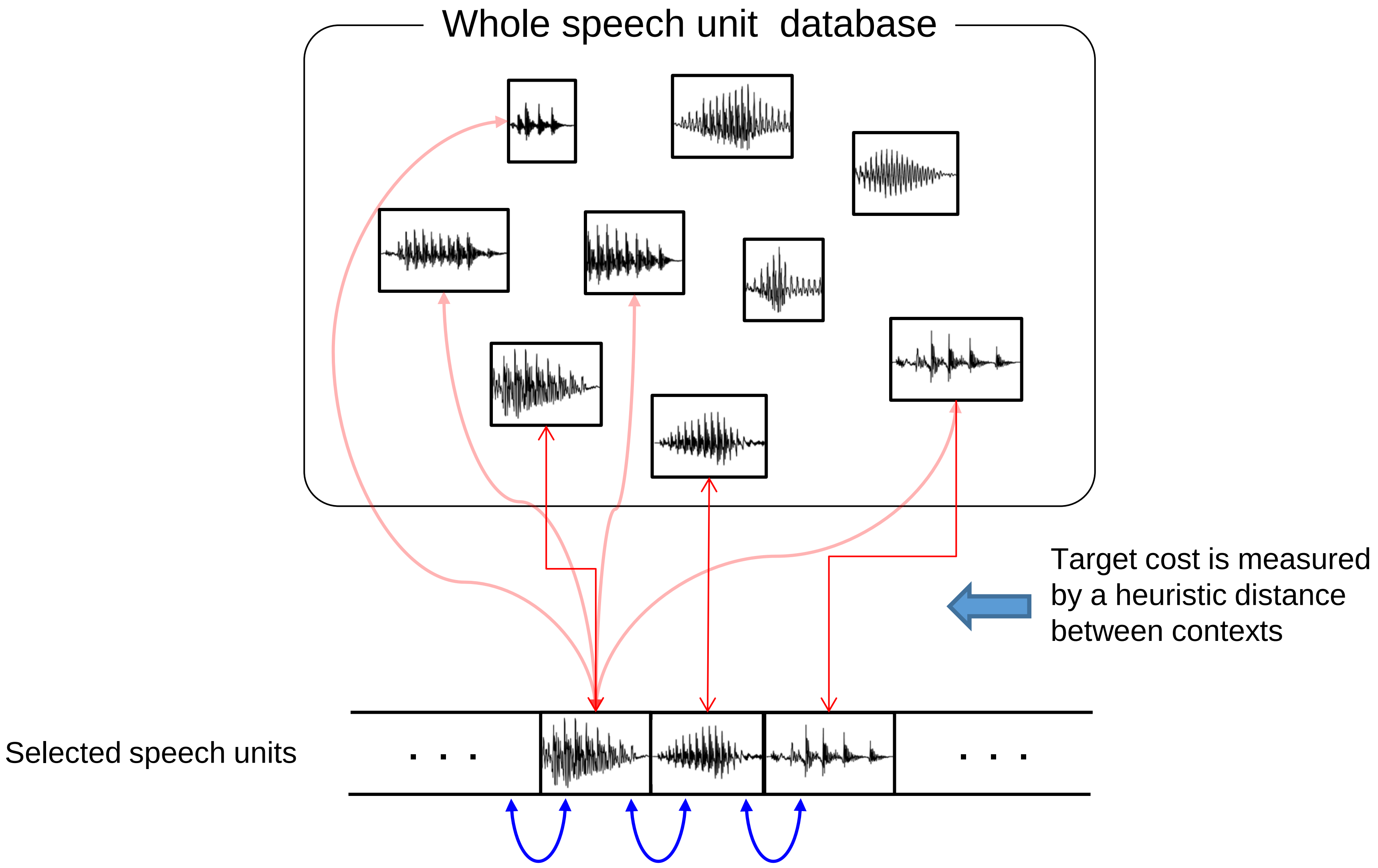
1. Assistive Technology for those with Speech Impairments
2. Navigation and Voice Commands—Enhancing GPS Navigation with Spoken Directions
3. Educational Materials and Language Learning
4. Audio Books
5. Entertainment Applications

Types of Speech Synthesis Systems

- rule-based:
  - formant synthesis
  - articulatory synthesis
- concatenation of units
  - monophone
  - diphone
  - micro-segmental
  - unit selection



Example: Concatenative synthesis



Experiment for yourself!

1970s

Formant synthesis

1980s

Diphone synthesis

1990s-2000s

Unit selection

2000s

HMM synthesis

2001

Microsoft XP synthesizer

2005

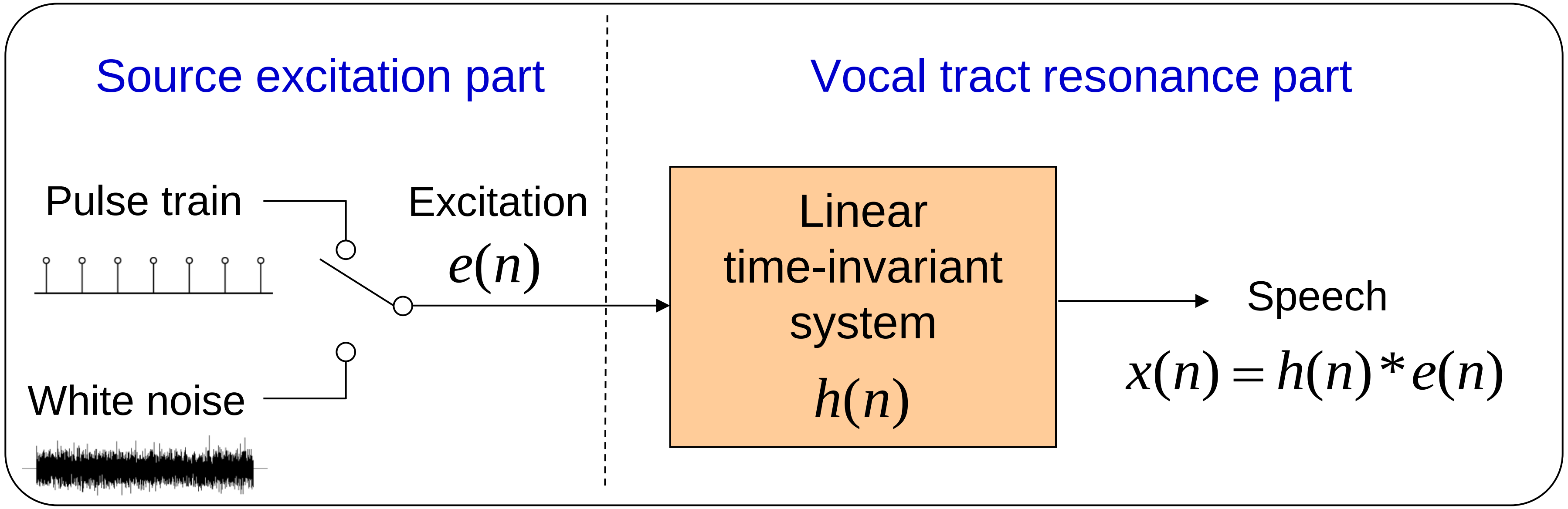
Microsoft 7 and Windows Vista synthesizer

2020

IBM's Watson neural synthesizer

Many of these make use the source-filter model for speech production

Overview of speech vocoding

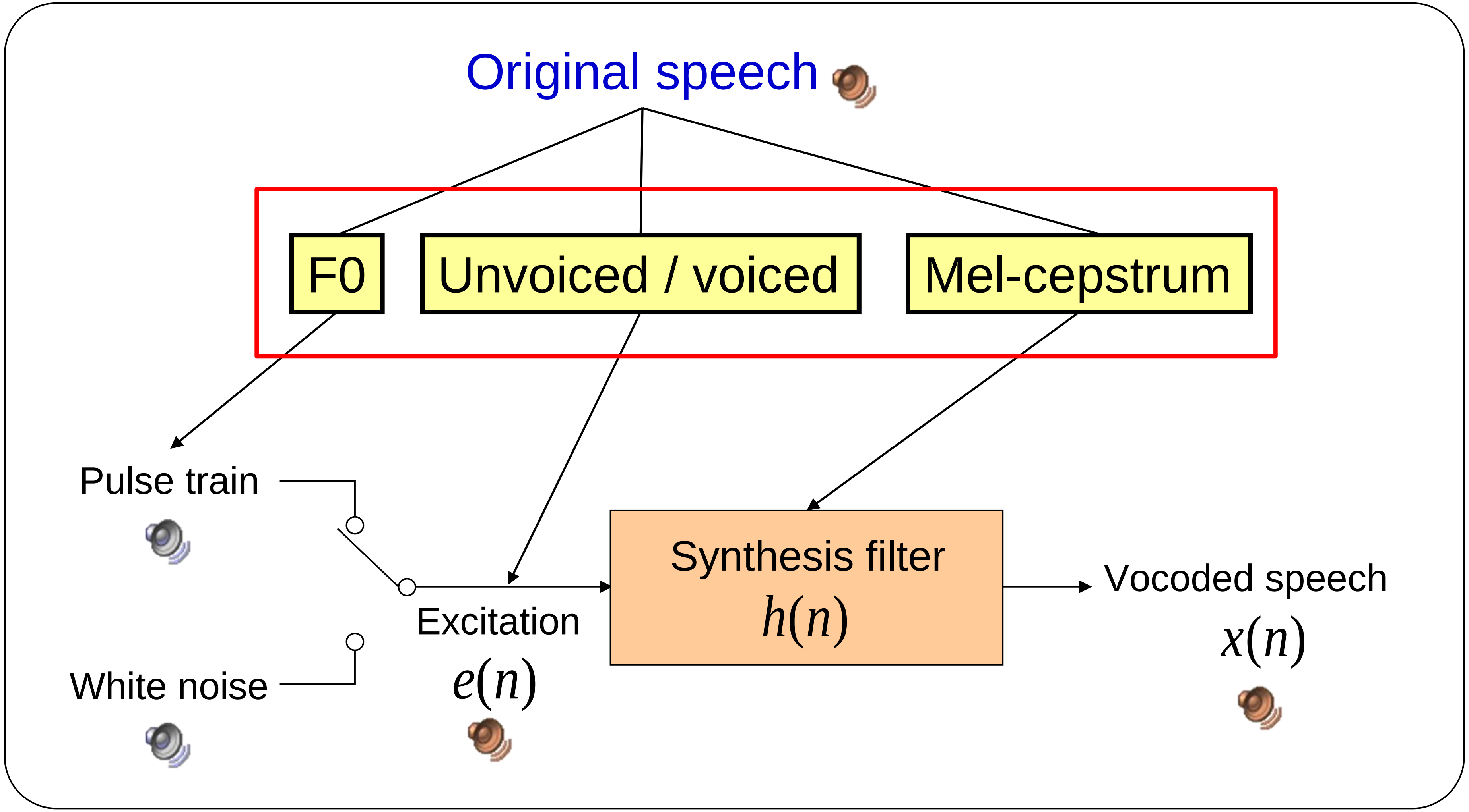


$$x(n) = h(n) * e(n)$$

↓ Fourier transform

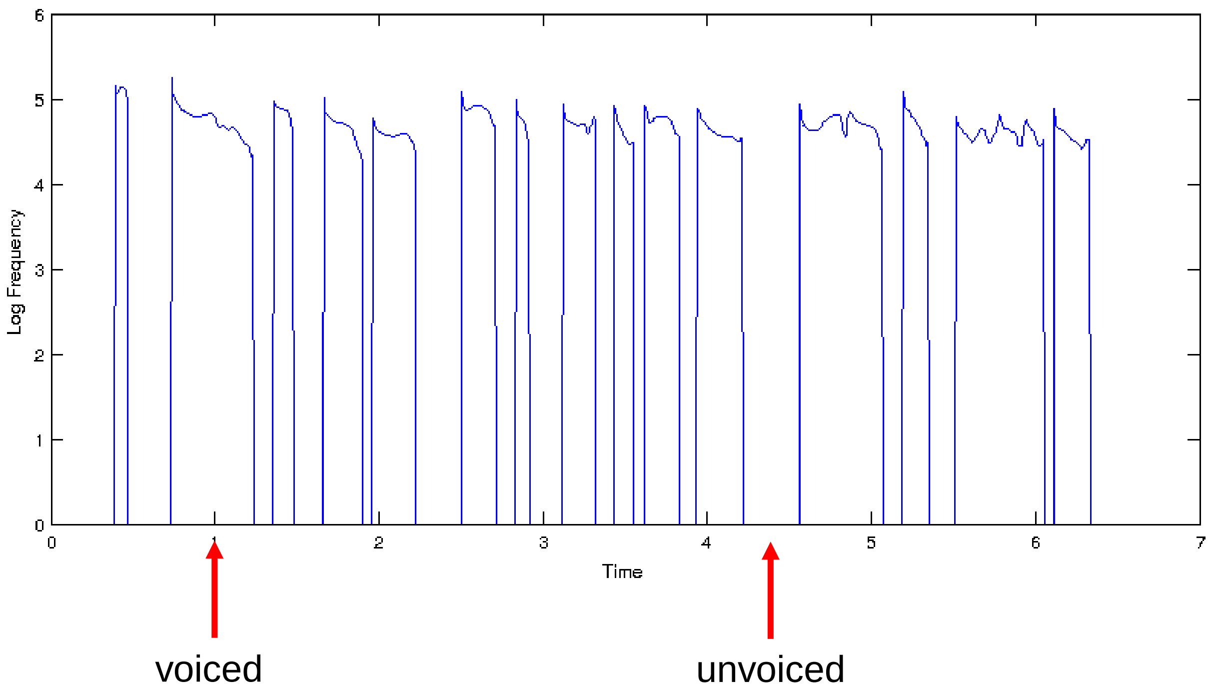
$$X(e^{j\omega}) = H(e^{j\omega})E(e^{j\omega})$$

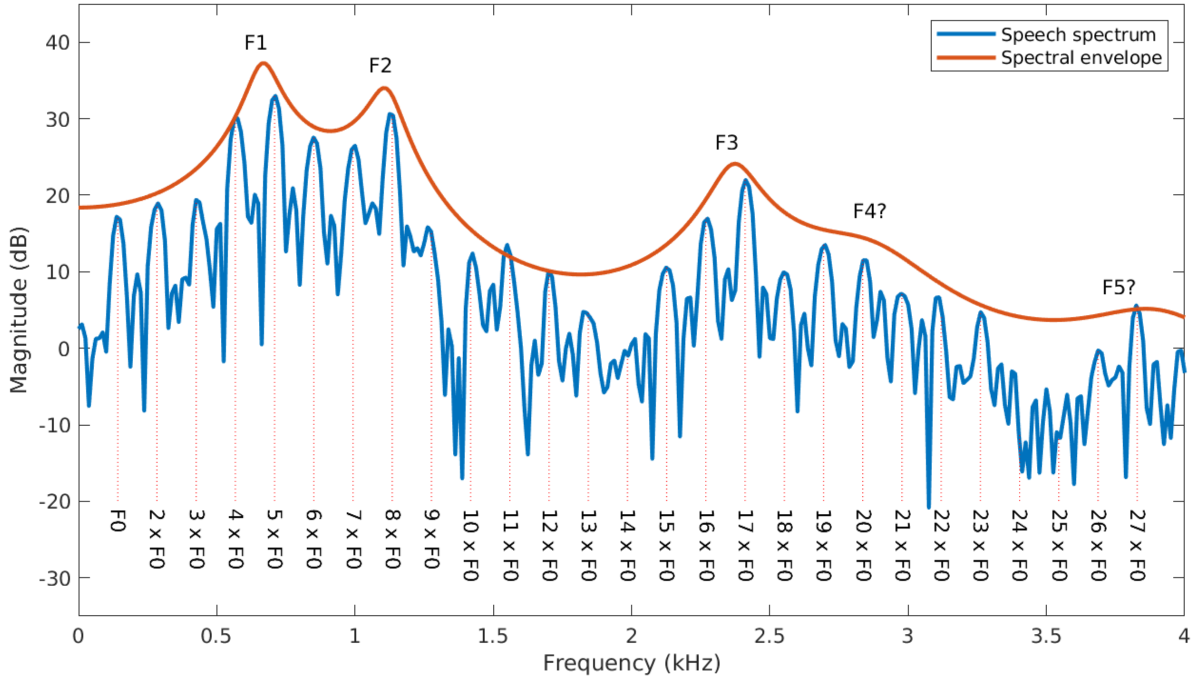
Source-Filter Model



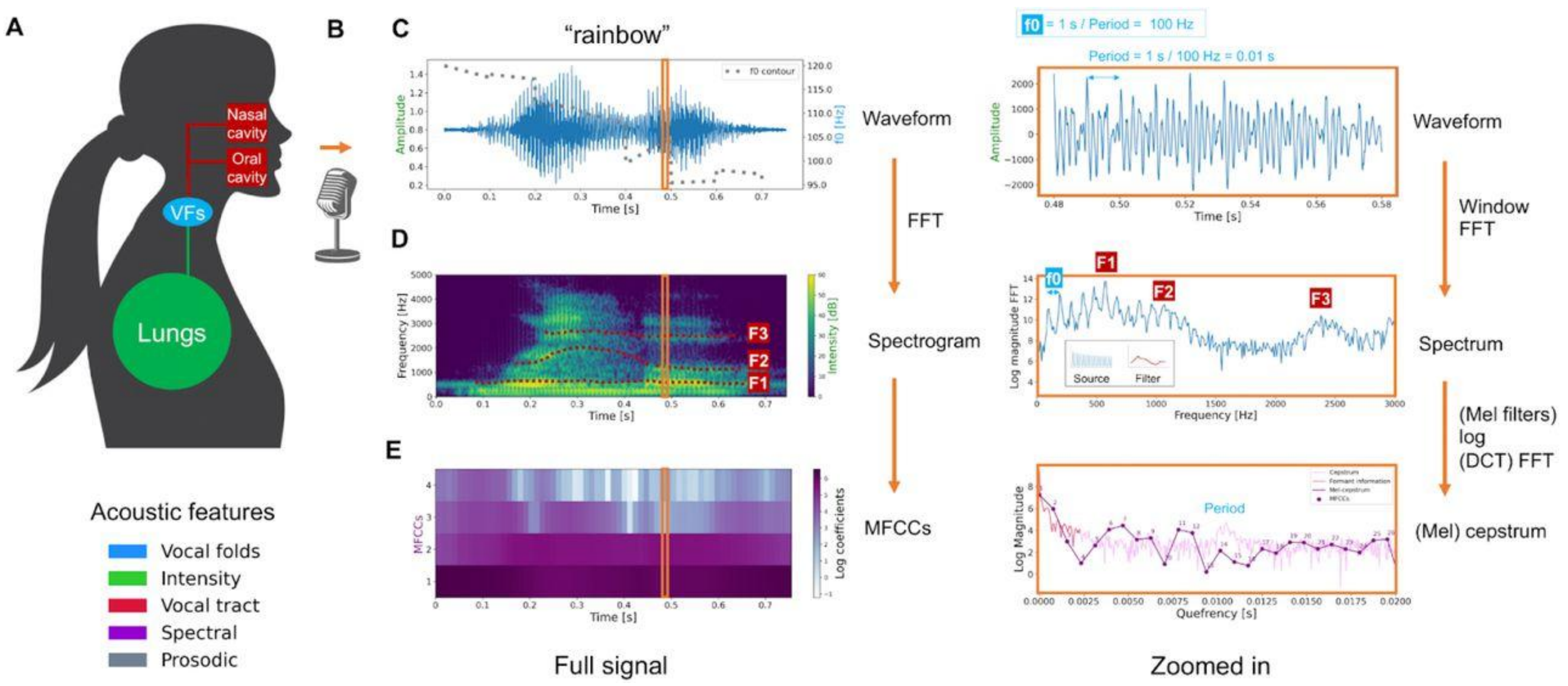
Speech parameters

F0





Speech parameters



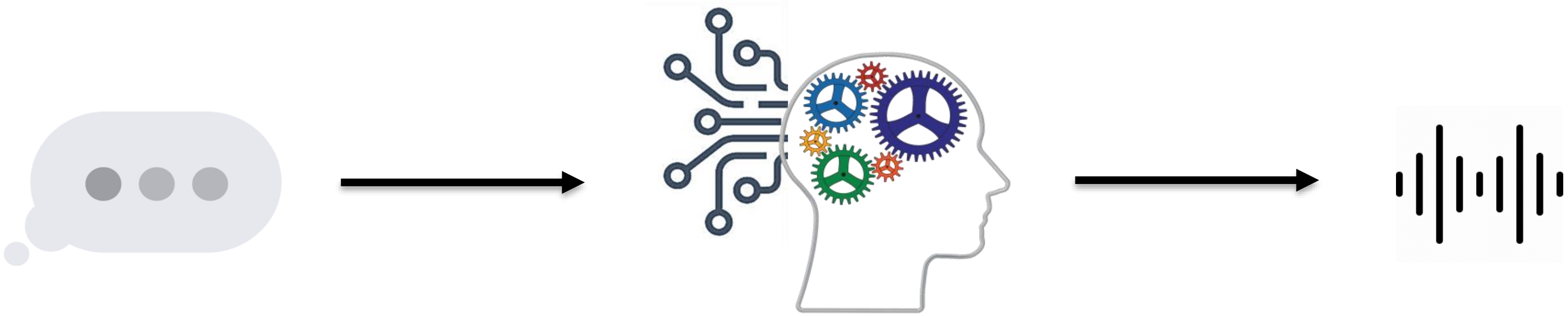
(A) Speech production
(B) recording characteristics
(C) Waveform (f0)
(D) spectrogram (F1-F3, intensity)
(E) mel-frequency cepstral coefficients (MFCCs)

https://www.medrxiv.org/content/10.1101/2020.11.23.20235945v5.full

Text-to-Speech (TTS)

What is TTS Synthesis?

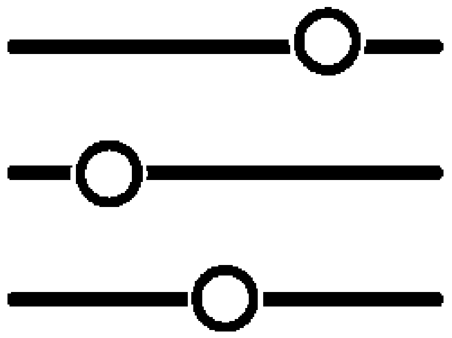
- It is a technology that converts written text into spoken words.
- TTS systems analyze input text and generate corresponding synthesized speech output, allowing computers or devices to "speak" the text aloud.



What is Parametric TTS

How does it work?

- using learning based parametric models, e.g., HMM
- all the information required to generate speech is stored in the parameters of the model
- also called statistical parametric synthesis (SPSS)



Advantages:

- lower data cost and more flexible

Limitations:

- less intelligible than concatenative TTS

What is Neural TTS

How does it work?

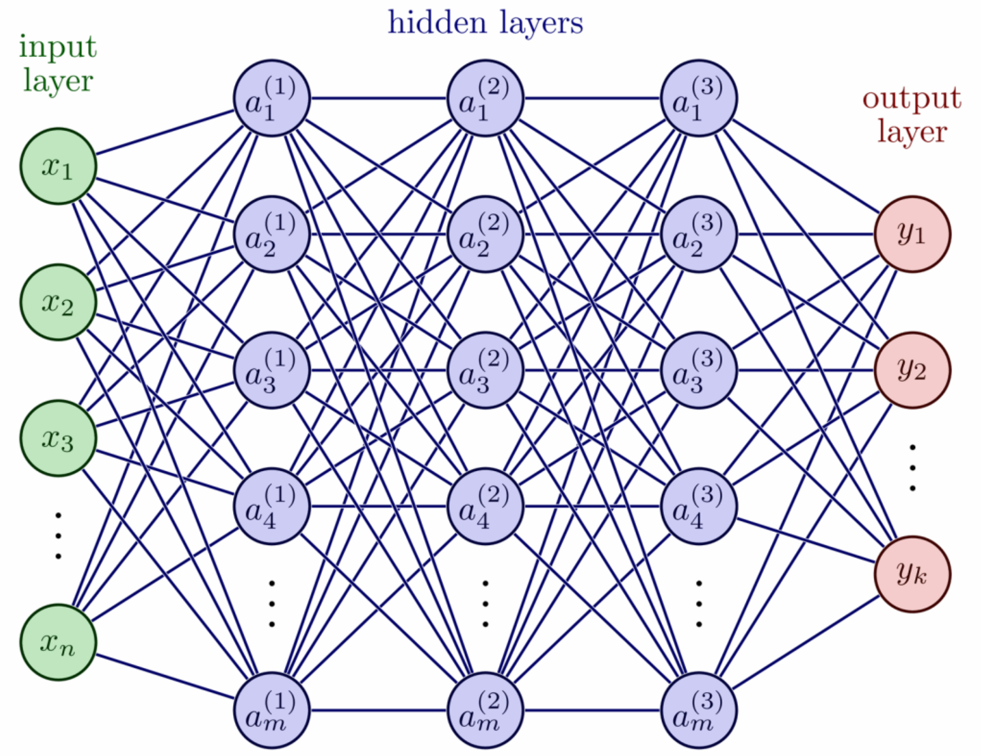
- special kind of parametric models
- text to waveform mapping is modeled by deep neural networks

Advantages:

- huge quality improvement (intelligibility and naturalness)
- less human preprocessing and feature engineering

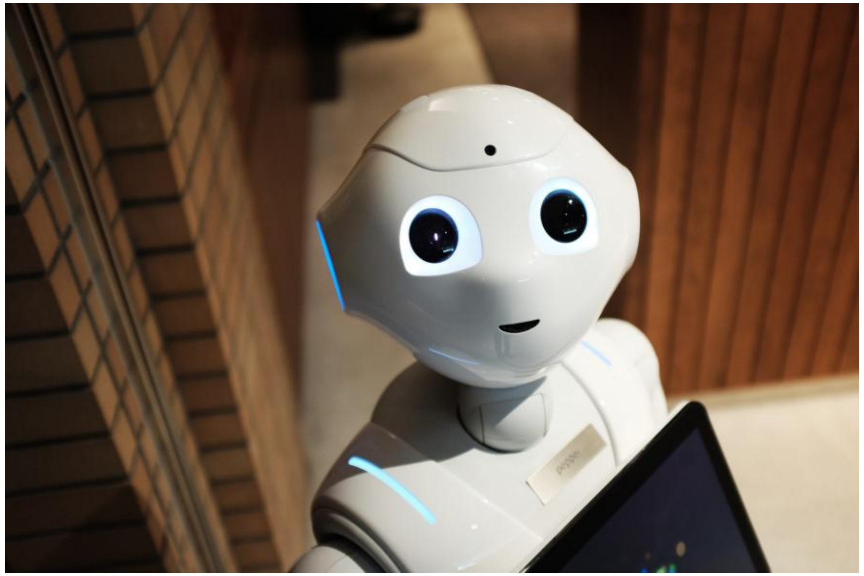
Disadvantages:

- Training/inference costly

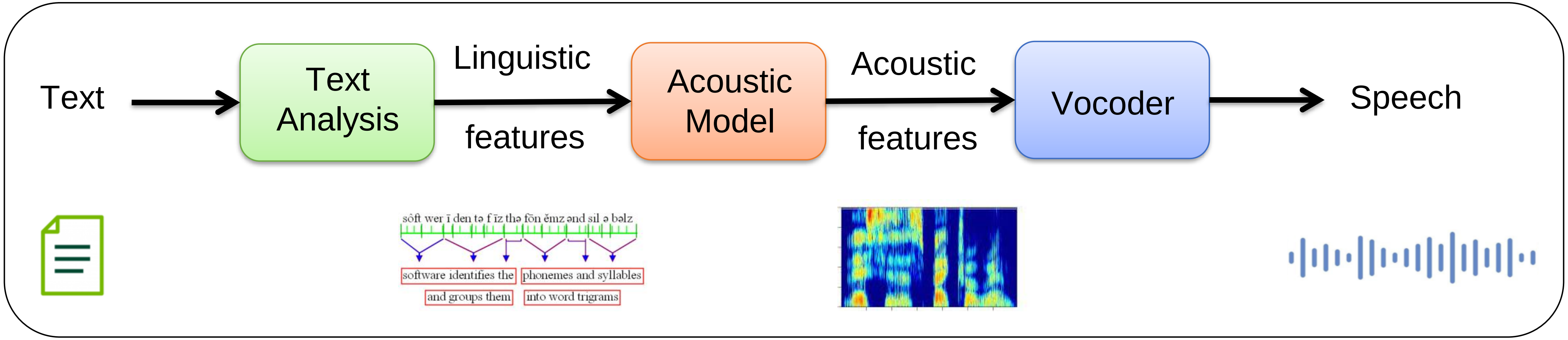


Applications of TTS

- learning disabilities
- proof-reading in word-processors
- language tutoring systems
- navigation and location services
- information access over telephone
- aid to the handicapped
- e-books and audiobooks
- voice generation for content creation
- games, simulators, toys
- etc.



Key components of TTS systems



- Text analysis: text → linguistic features
- Acoustic model: linguistic features → acoustic features
- Vocoder: acoustic features → speech

Text analysis

Transforms input text into linguistic features

Text normalization
- 1989 → nineteen eighty-nine, Jan. 24th→ January twenty-fourth

Phrase/word/syllable segmentation
- synthesis → syn-the-sis

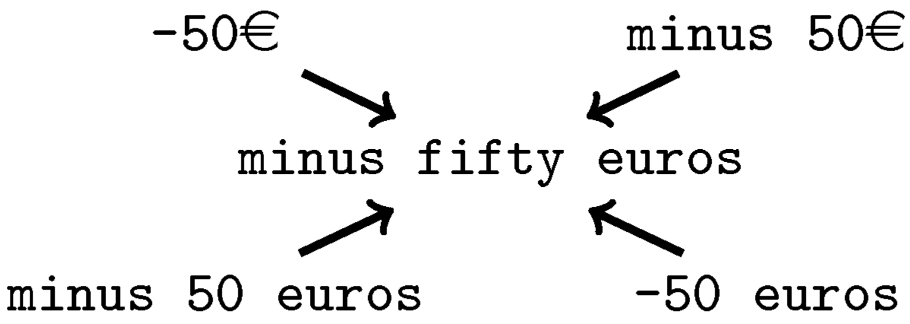
Part of speech (POS) tagging
- Mary went to the store → noun, verb, prep, noun,

Grapheme-to-phoneme conversion
- Speech → s p iy ch

Text normalization

- process of transforming text into a standard, consistent format:
    - **Lowercasing:** convert all characters to lowercase for uniformity
    - **Tokenization:** break down the text into individual words or tokens.
    - **Stemming/Lemmatization:** reduce words to their base or root form.
    - **Stop Words Removal:** eliminate common words that don't contribute significantly to meaning.
    - **Handling Numbers/Symbols:** standardize the representation of numbers, dates, and special characters.

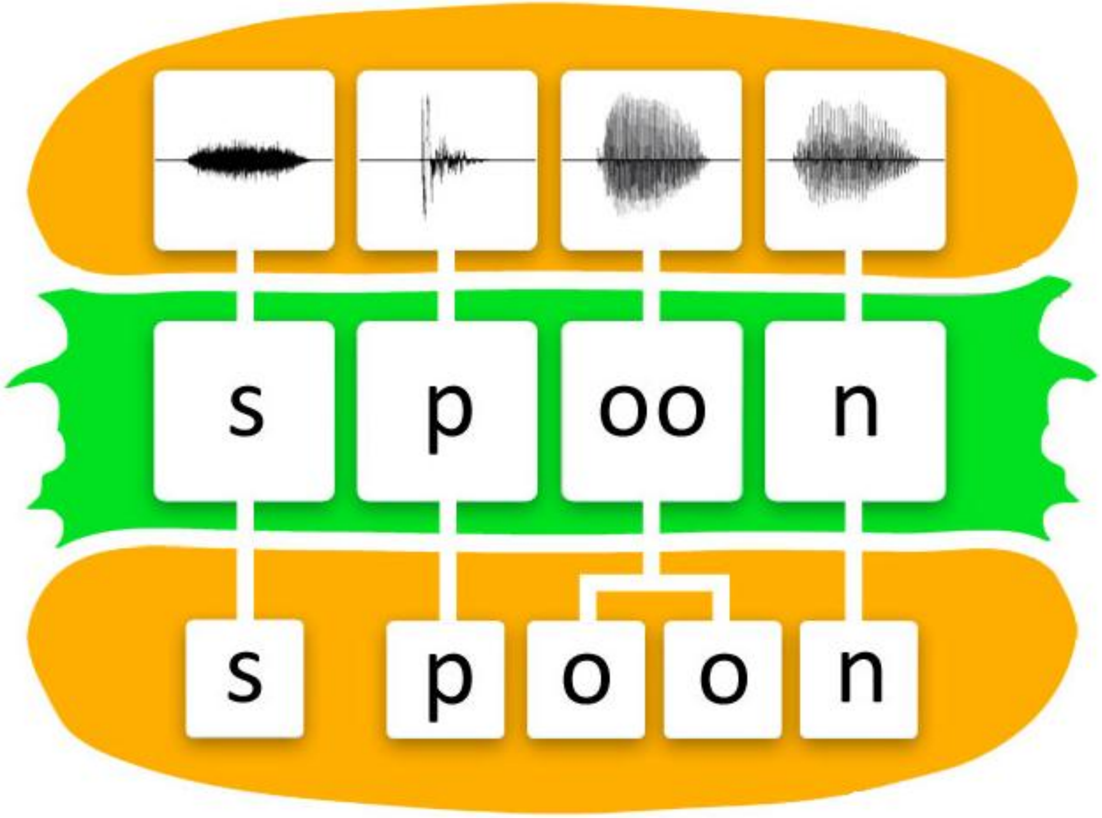
"The meeting is scheduled for 3:30 PM.“ → "meeting schedule 3:30 pm."



Raw	**Normalized**
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

Grapheme-to-Phoneme conversion

- **Phonemes**: smallest units of sound in a language
- **Graphemes:** smallest units of a writing system
- **Letters:** visual building blocks of written words.



Phonemes

Graphemes

Letters

G2P: process of converting written language into spoken language.

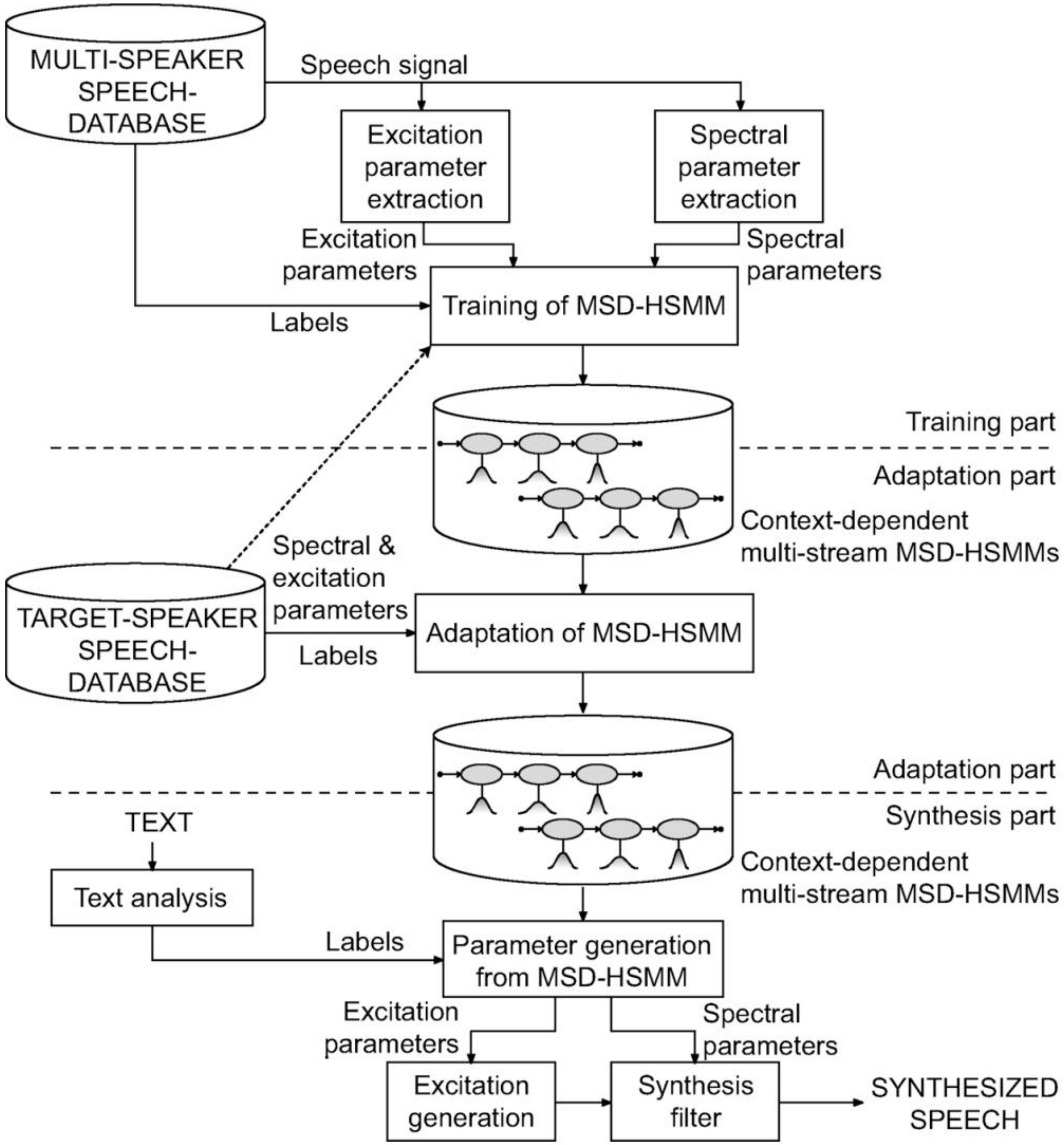
Acoustic model

➤ **Generate/Predict acoustic features from linguistic features**

- F0, V/UV, energy
- Mel-scale Frequency Cepstral Coefficients (MFCC), Bark-Frequency Cepstral Coefficients (BFCC)
- Mel-generalized coefficients (MGC), band aperiodicity (BAP),
- Linear prediction coefficients (LPC),
- Mel-spectrograms
- Pre-emphasis, Framing, Windowing, Short-Time Fourier Transform (STFT), Mel filter

Acoustic model —— HMM

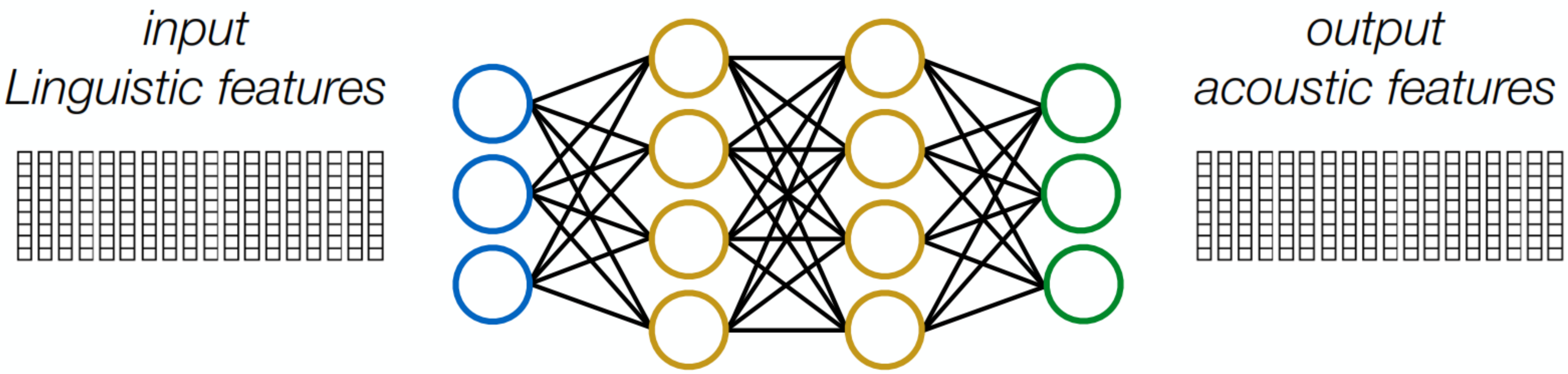
Robust Speaker-Adaptive HMM-Based TTS Synthesis



https://era.ed.ac.uk/bitstream/handle/1842/3962/yamagishi-taslp09.pdf?sequence=1&isAllowed=y

Acoustic model —— FF-DNN

Feed Forward Deep Neural Network

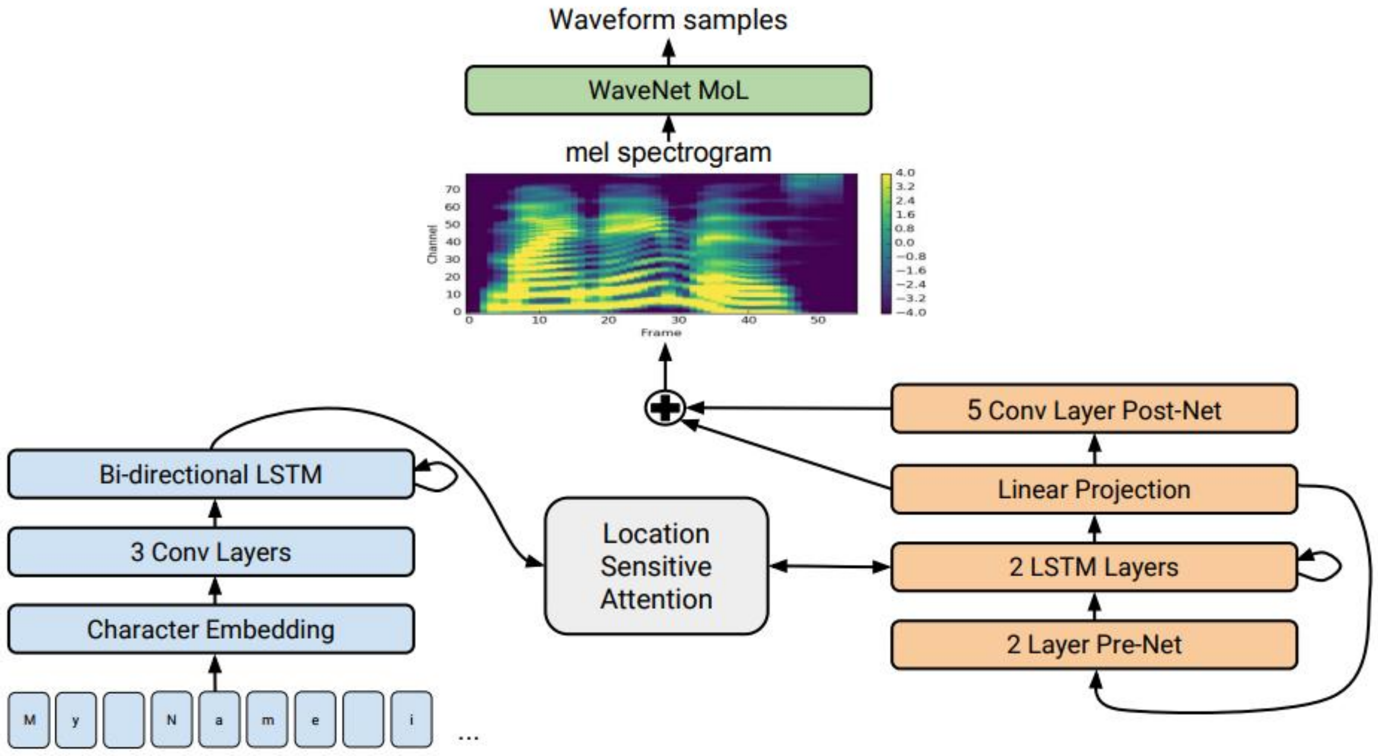


https://www.isca-speech.org/archive/pdfs/ssw_2016/wu16_ssw.pdf

Acoustic model —— RNN

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

- Text to mel-spectrogram generation
- LSTM based encoder and decoder
- Location sensitive attention
- WaveNet as the vocoder

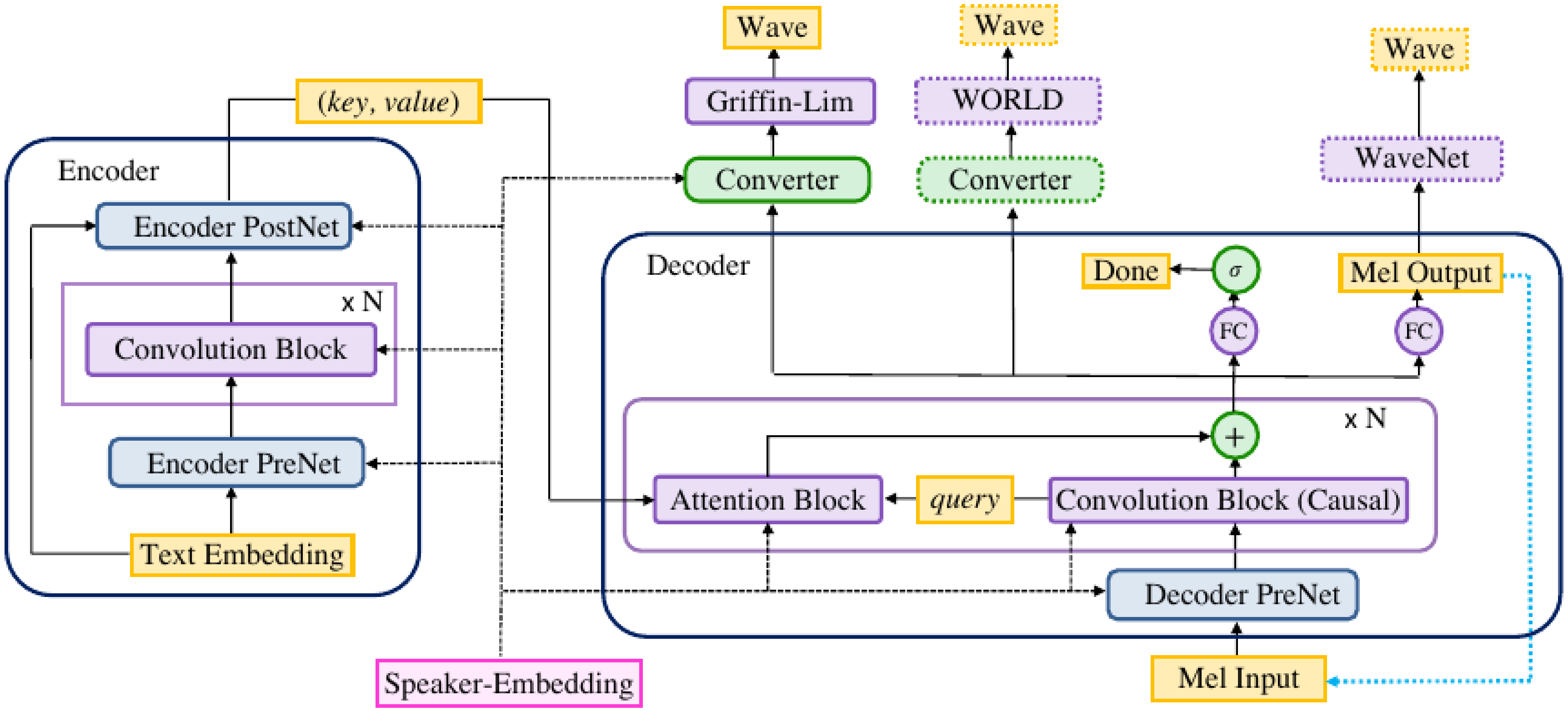


https://arxiv.org/abs/1712.05884

Acoustic model —— CNN

Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning

- Enhanced with purely CNN based structure
- Support different acoustic features as output
- Support multi-speakers

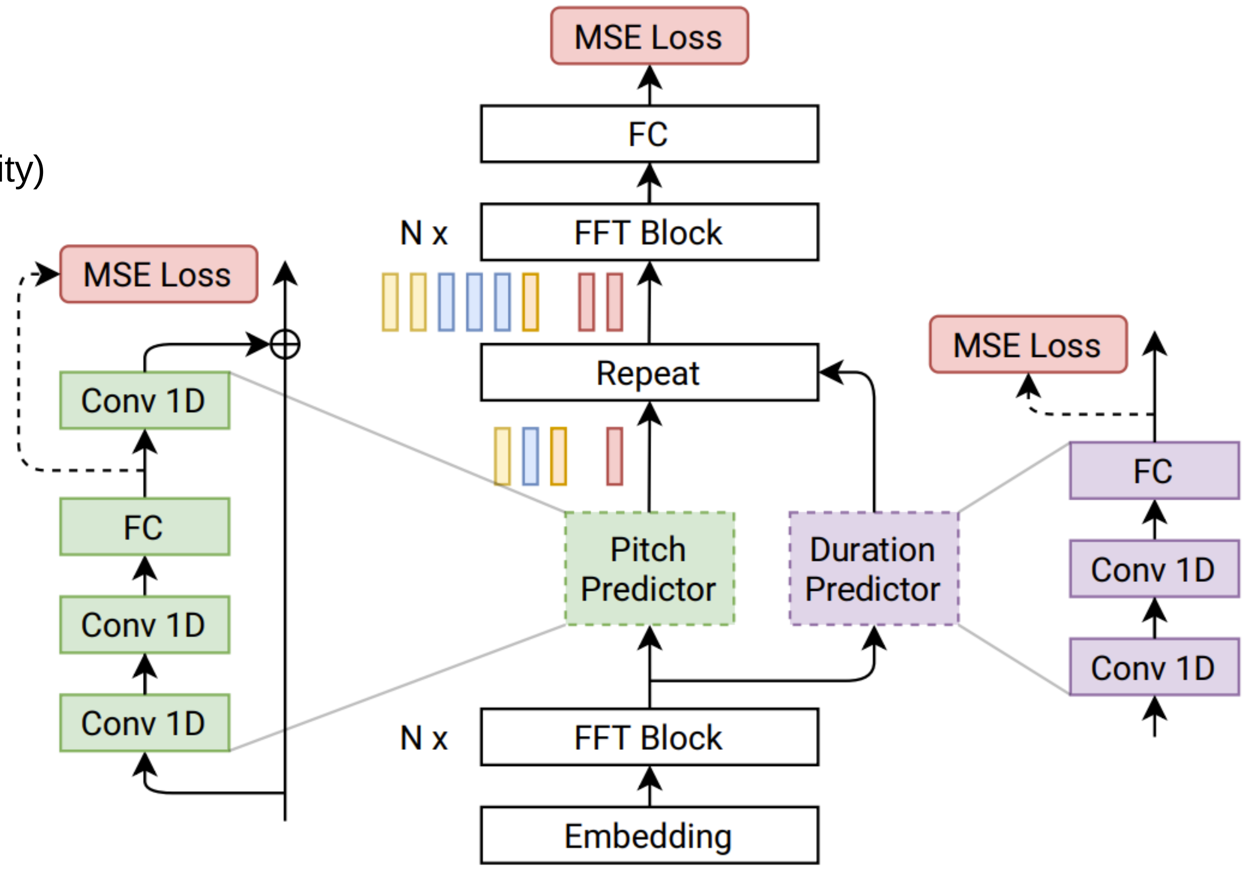


https://arxiv.org/abs/1710.07654

Acoustic model —— Transformer

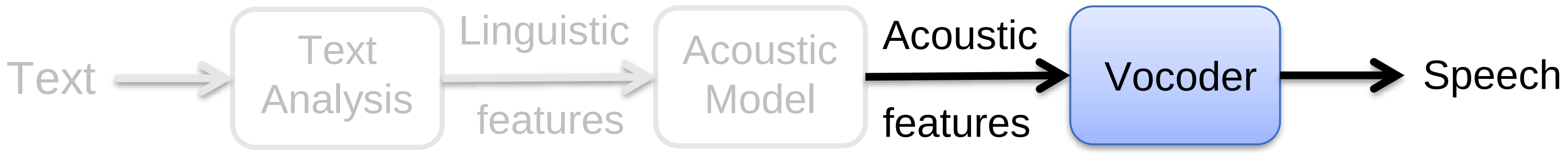
FastPitch: Parallel Text-to-speech with Pitch Prediction

- conditioned on fundamental frequency contours
- generate mel-spectrogram in parallel (for speedup)
- feed-forward transformer with length regulator (for controllability)
- predicts pitch contours during inference



https://arxiv.org/pdf/2006.06873.pdf

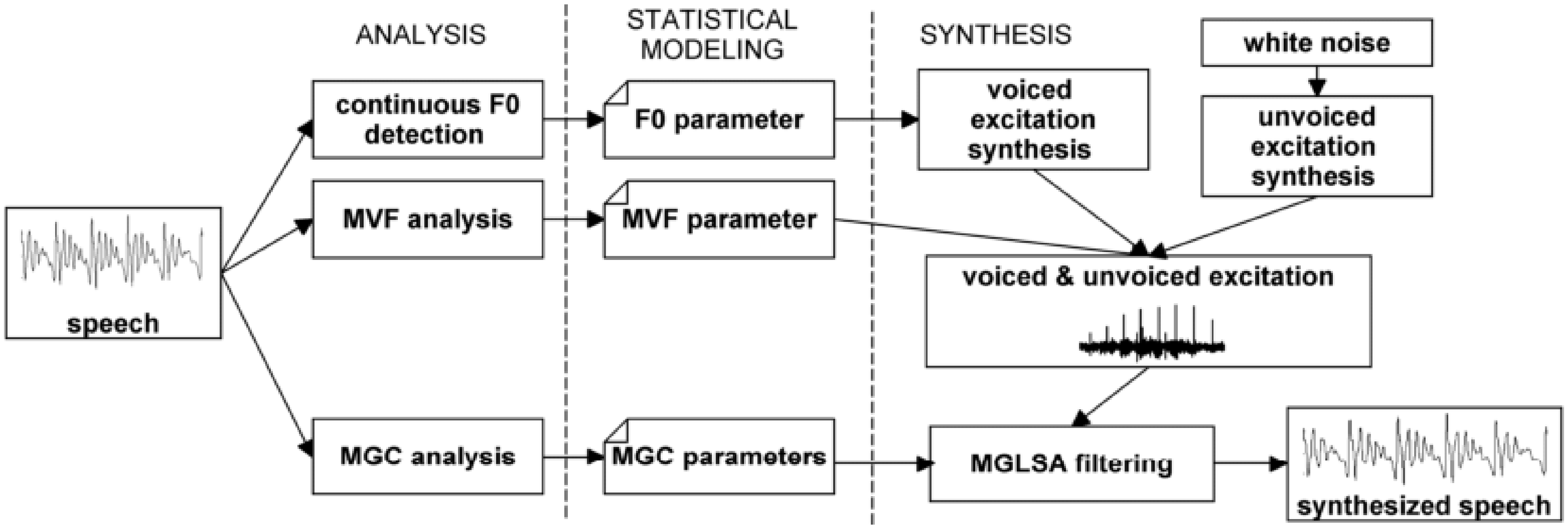
Vocoder



Model	Vocoder
Autoregressive	WaveNet, LPCNet, WaveRNN, FFTNet
Flow	WaveGlow, WaveFlow
GAN	WaveGAN, MelGAN, Hifi-GAN,
VAE	Wave-VAE
Diffusion	WaveGrad, DiffWave

Vocoder —— SPSS

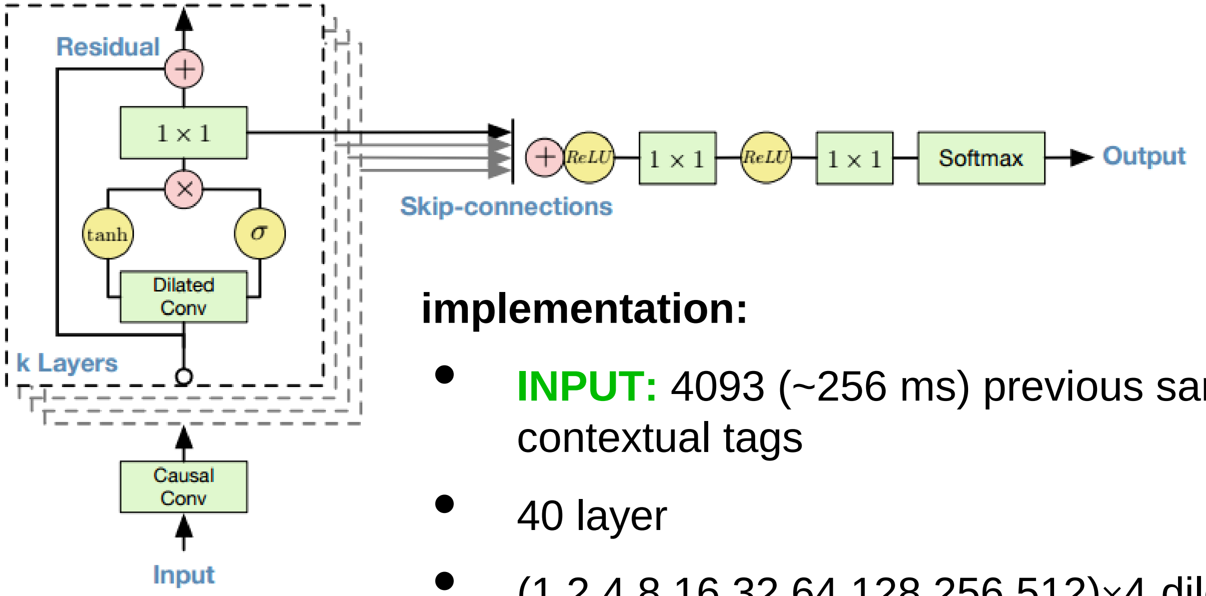
Continuous vocoder



https://www.isca-speech.org/archive/pdfs/interspeech_2019/csapo19_interspeech.pdf

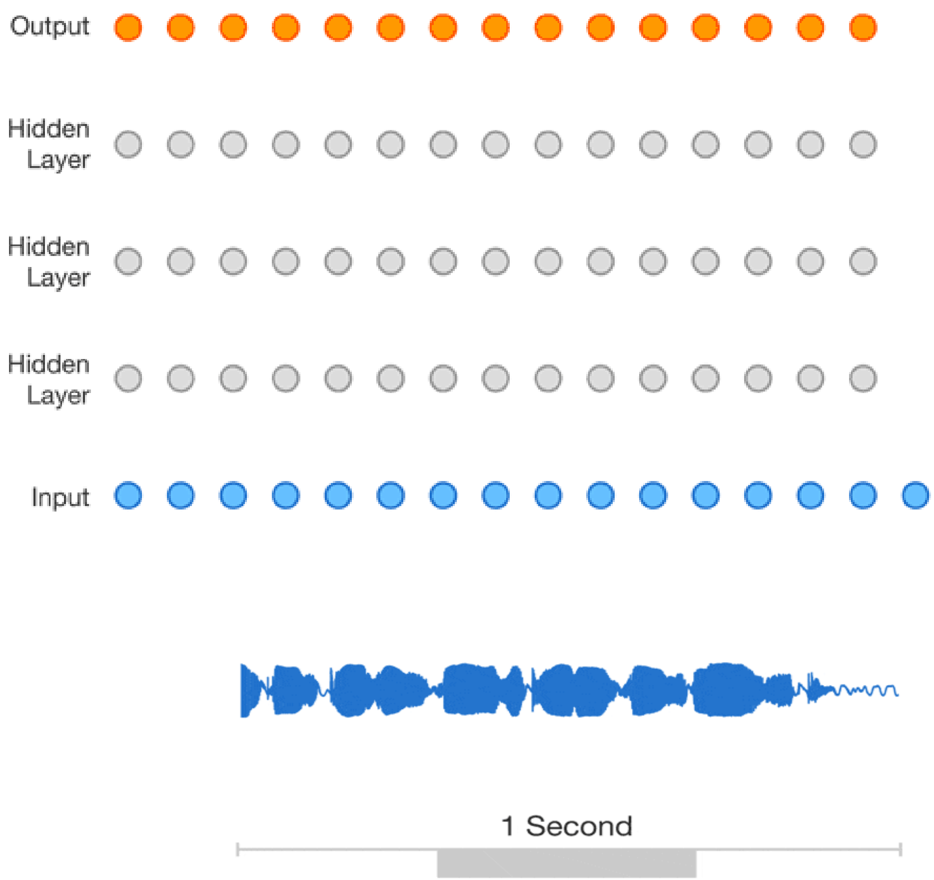
Vocoder —— Autoregressive

WaveNet: autoregressive model with dilated causal convolution



implementation:

- **INPUT:** 4093 (~256 ms) previous sample + contextual tags
- 40 layer
- (1,2,4,8,16,32,64,128,256,512)×4 diletation
- 256 dense layer for skip connections
- **OUTPUT:** 256 μ-law level quantized raw audio

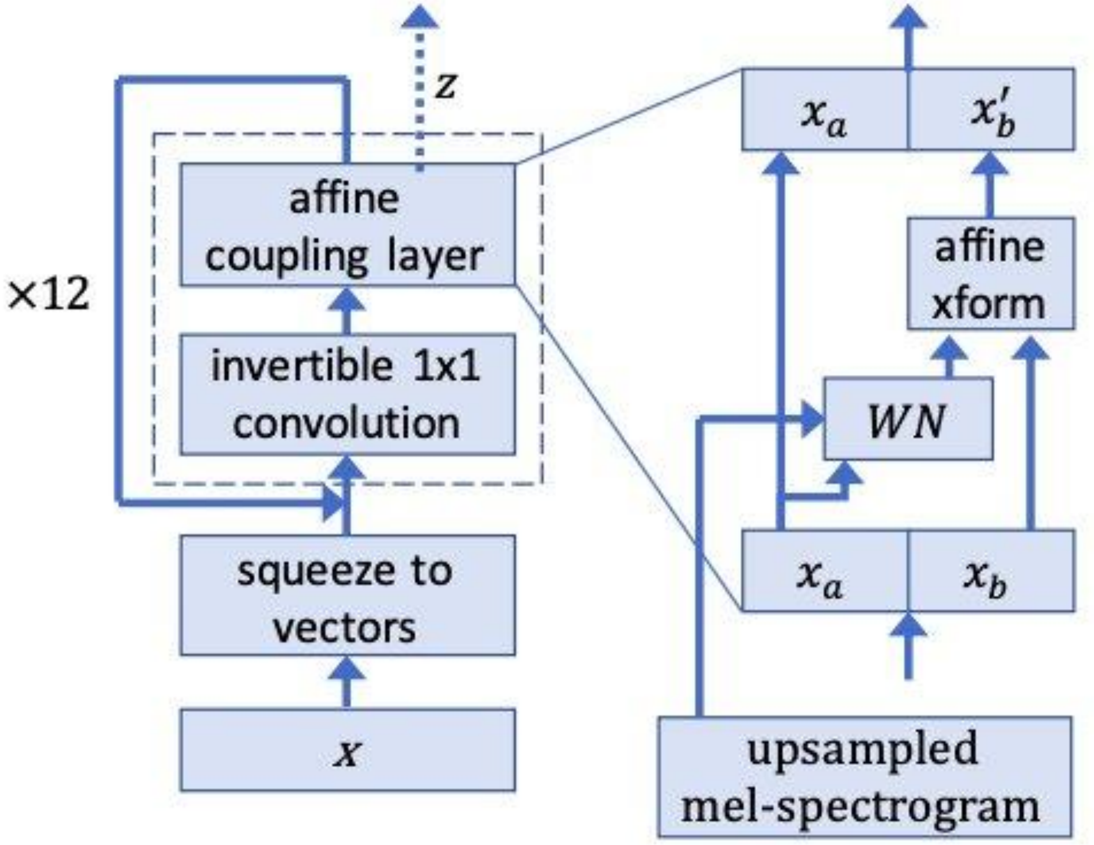


https://arxiv.org/pdf/1609.03499.pdf

Vocoder —— Flow

WaveGlow: A Flow-based Generative Network for Speech Synthesis

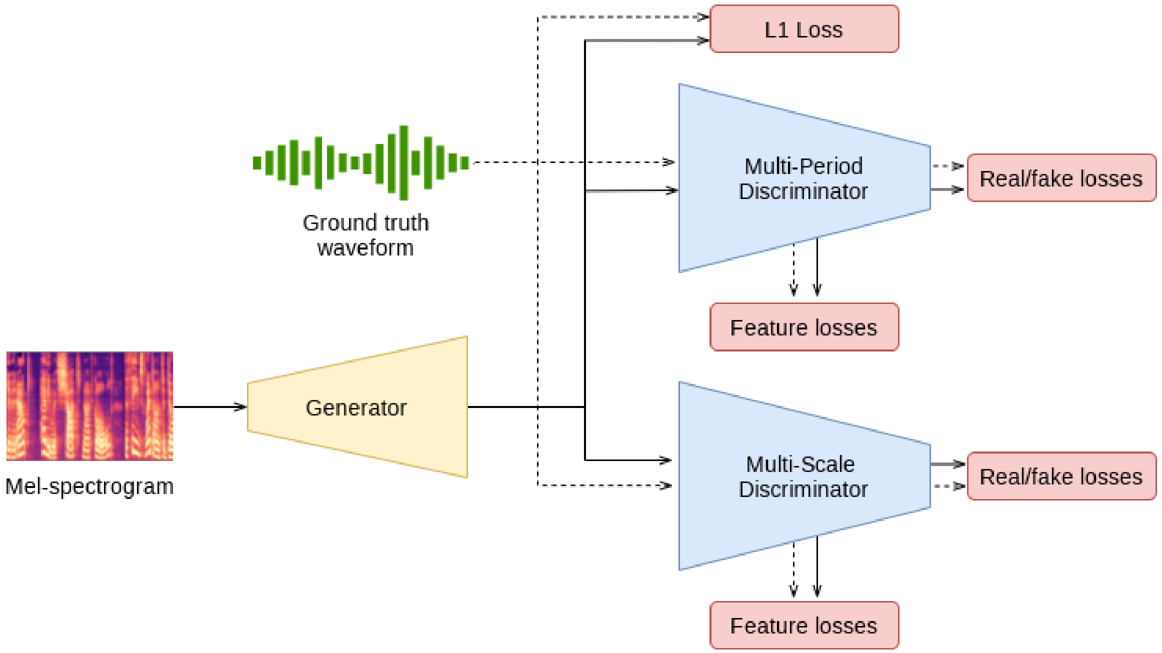
- Flow based transformation
- Affine Coupling Layer



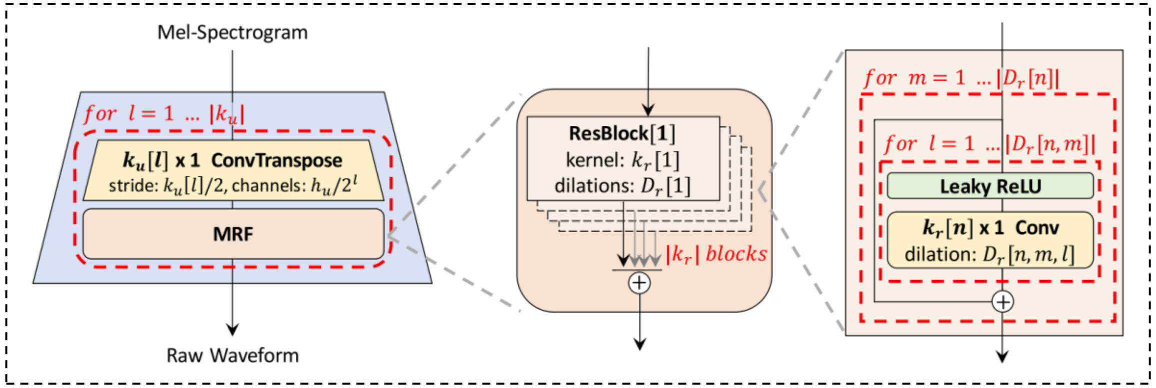
https://arxiv.org/pdf/1811.00002.pdf

Vocoder —— GAN

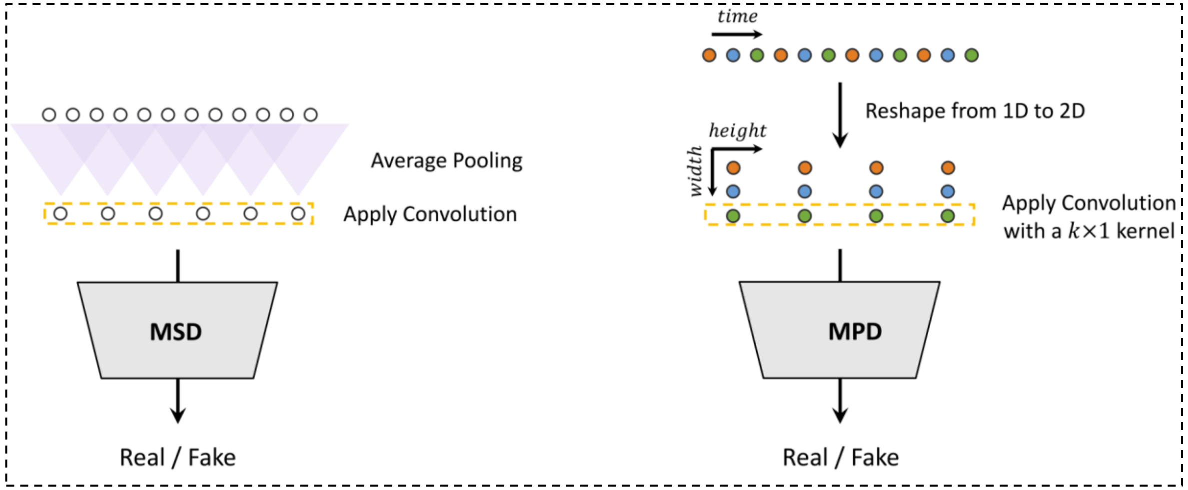
HiFi-GAN: Generative Adversarial Networks for Efficient and High-Fidelity Speech Synthesis



- synthesize speech waveforms from mel-spectrograms.
- follows the generative adversarial network (GAN)
- composed of a generator and a discriminator
- after training, the generator is used for synthesis, and the discriminator is discarded



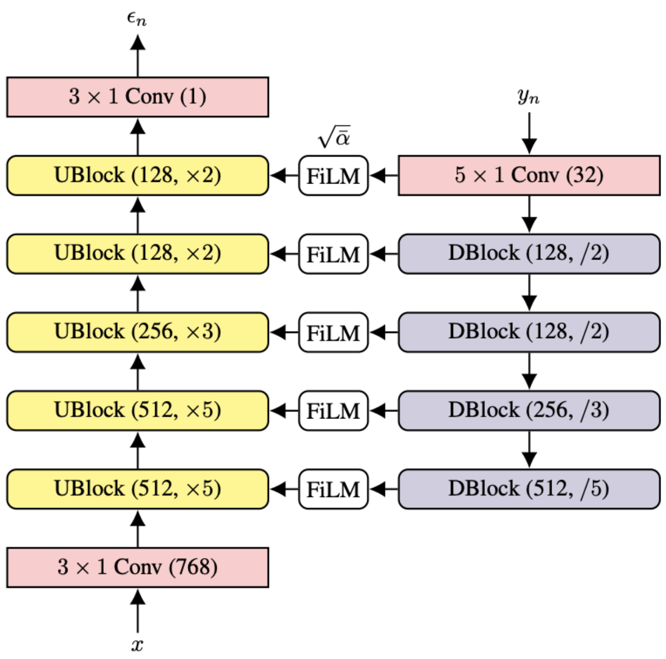
Generator



Discriminator

https://arxiv.org/pdf/2010.05646.pdf

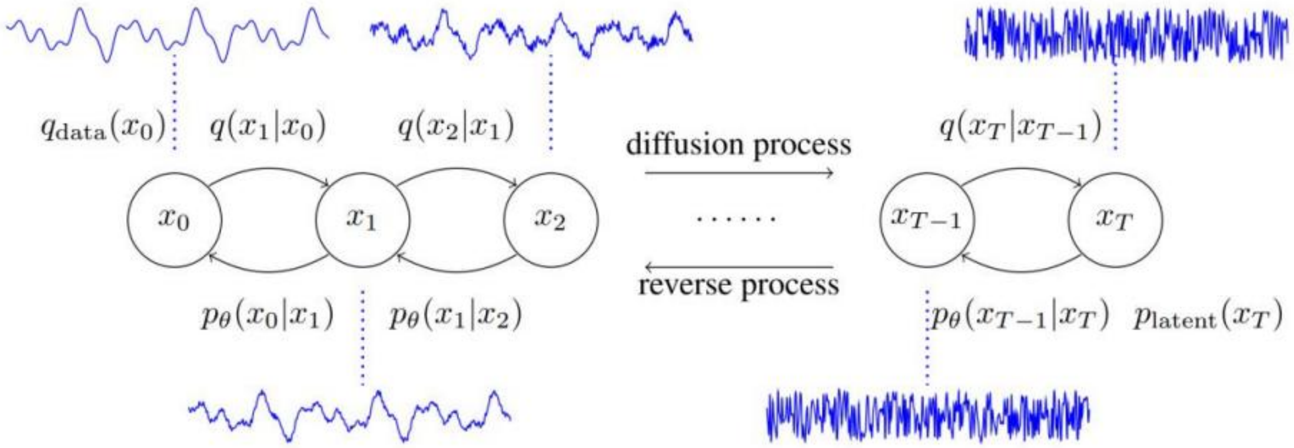
Vocoder —— Diffusion

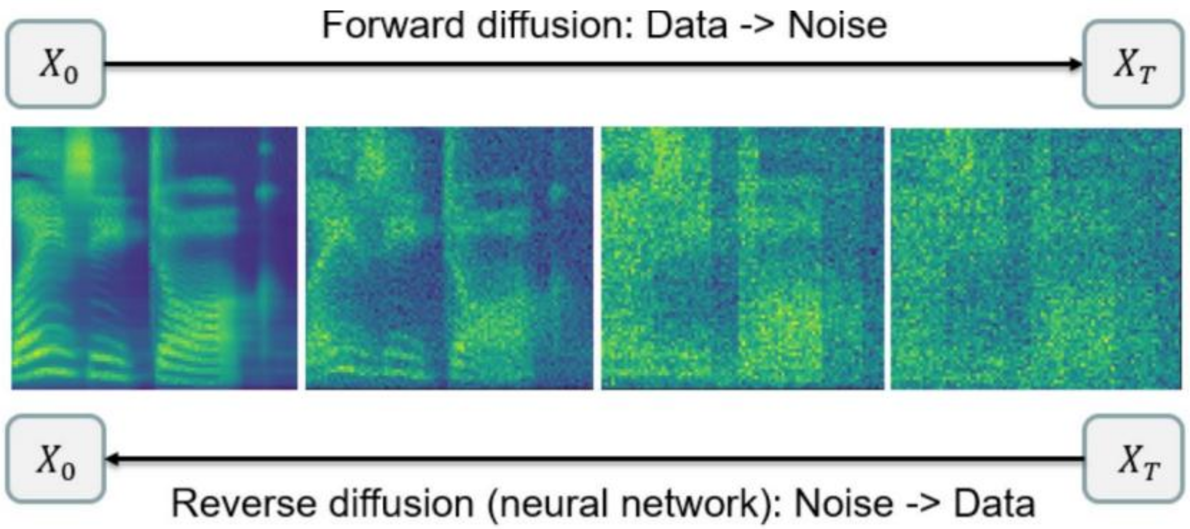


WaveGrad: Estimating Gradients for Waveform Generation

Diffusion probabilistic model

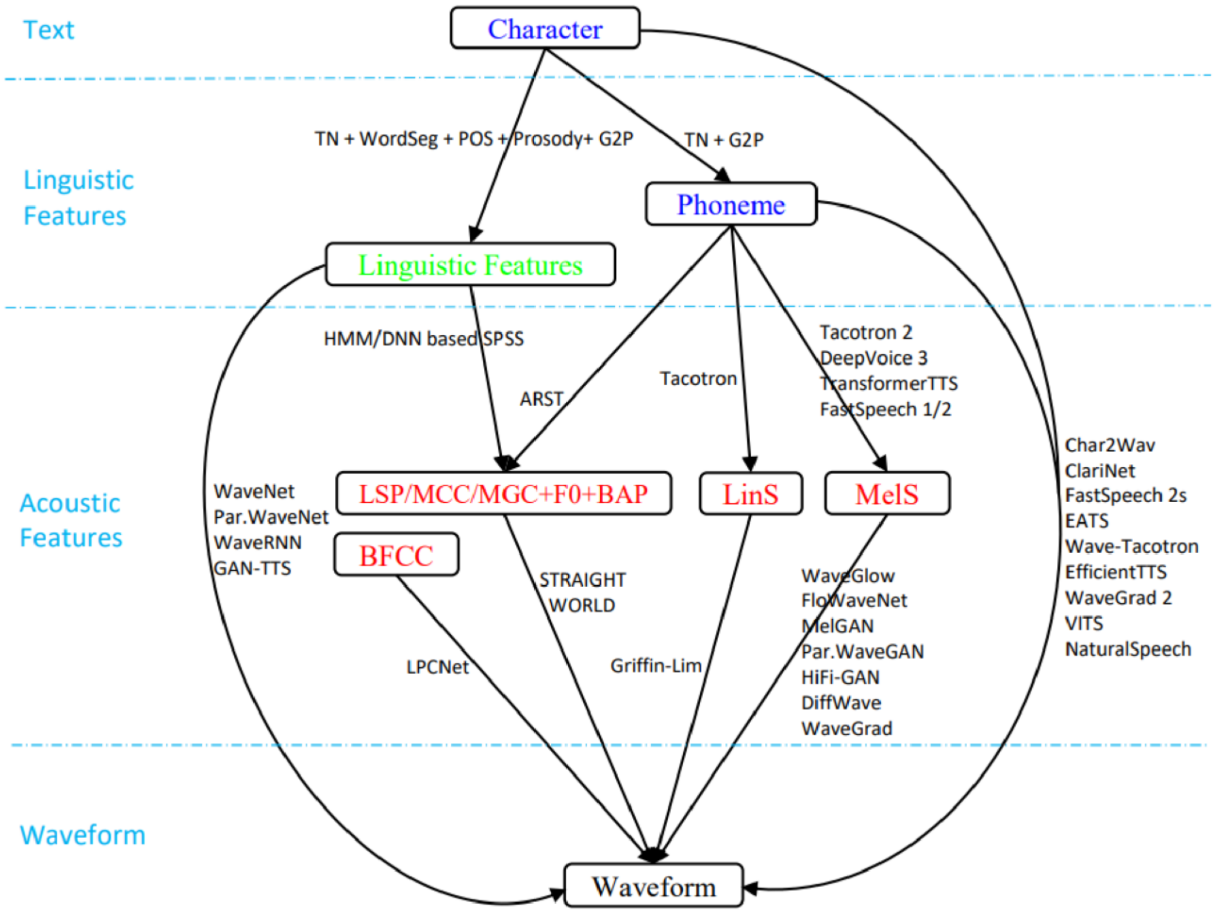
- Forward (diffusion) process
- Reverse (denoising) process

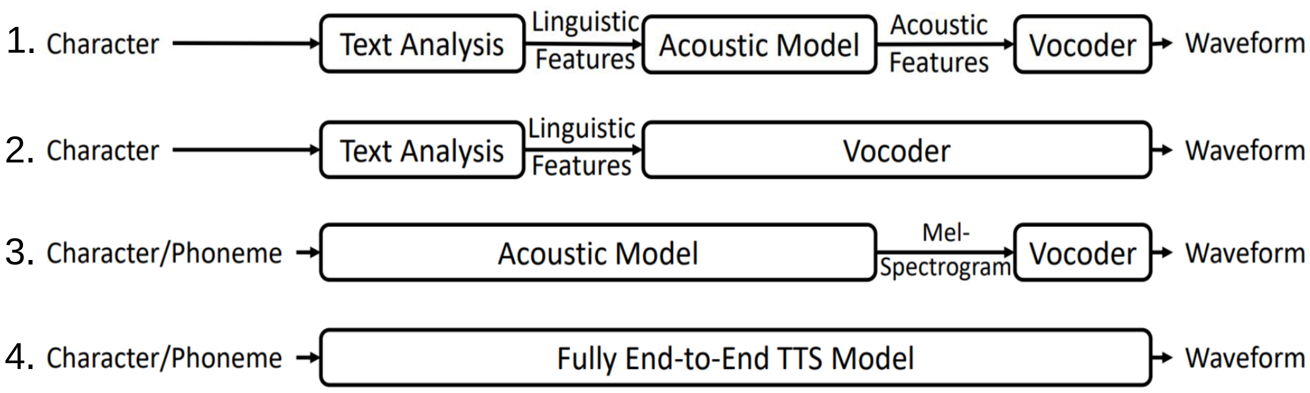




https://arxiv.org/pdf/2009.00713.pdf

Data conversion pipeline

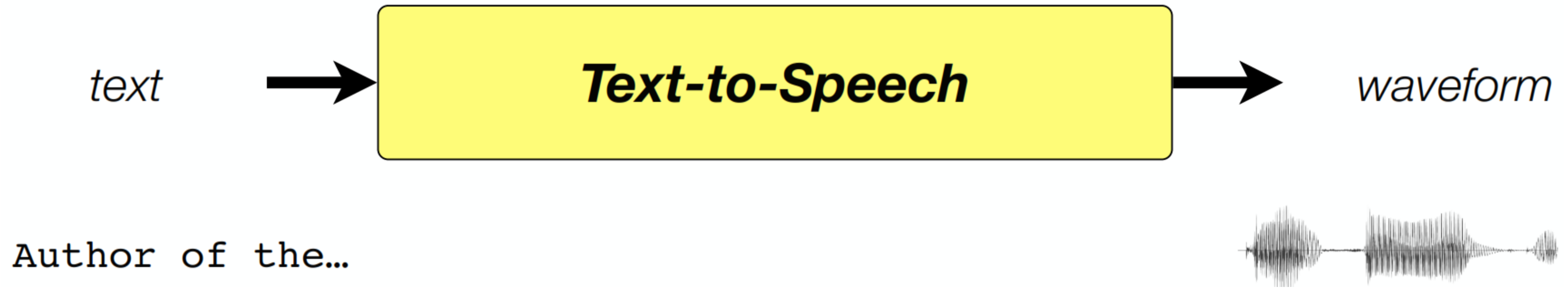




	Model
1	SPSS
2	WaveNet
3	Tacotron2, DeepVoice3
4	CharWav, VITS

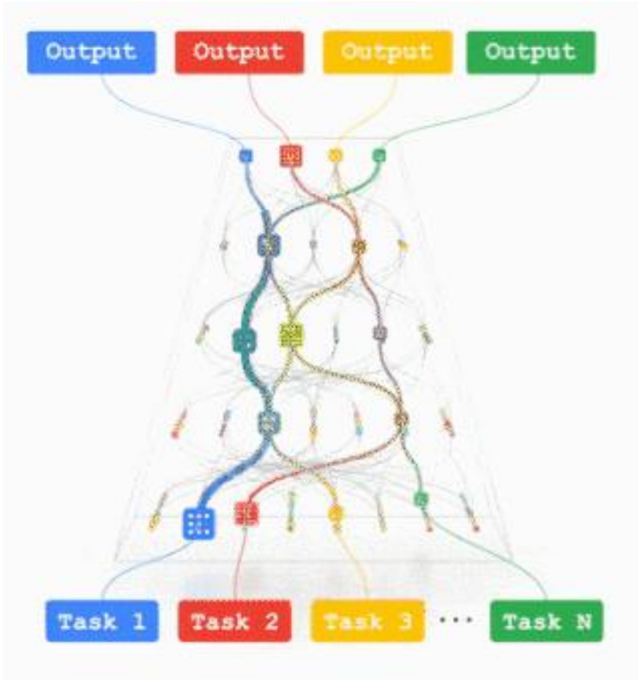
The end-to-end problem we want to solve

- end-to-end systems are systems which learn to directly map from an input sequence X to an output sequence Y , estimating P(Y |X)



Fully End-to-End TTS

Direct text/phoneme to waveform generation

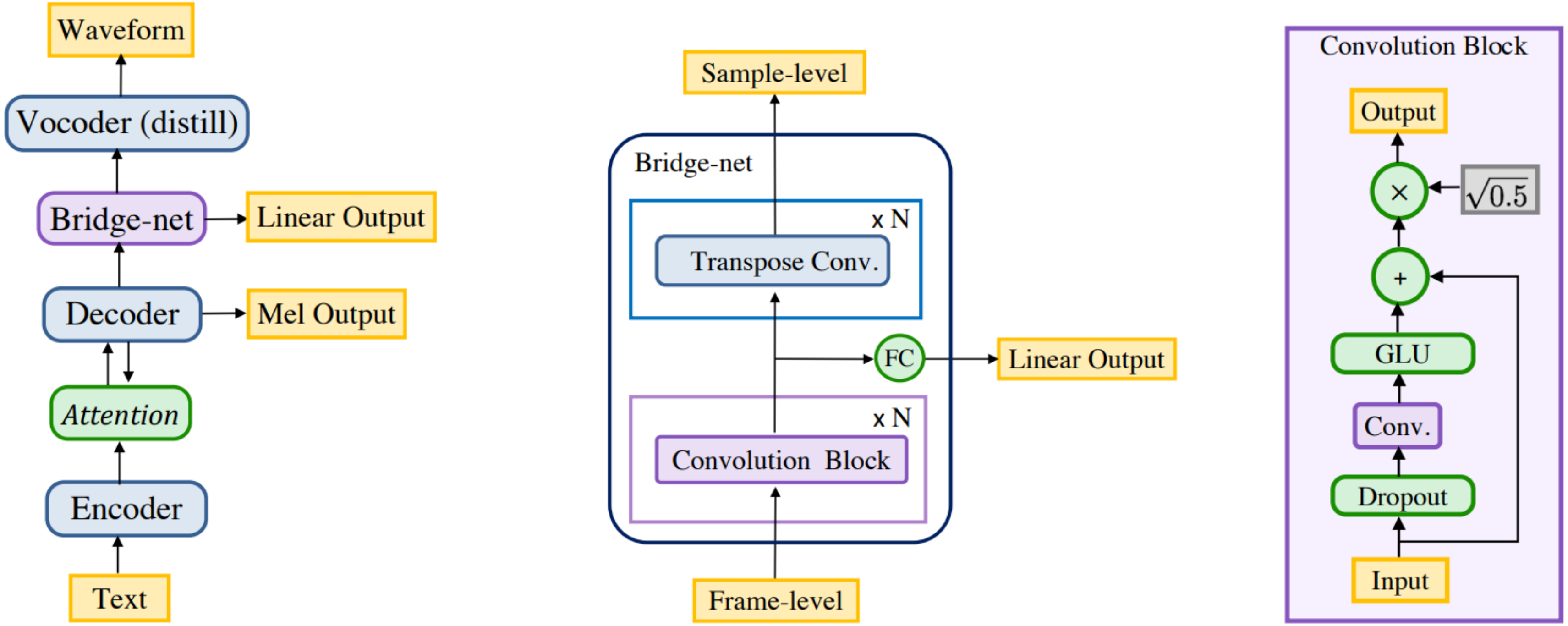


Advantages:

- Fully differentiable optimization (towards the end goal)
- Reduce cascaded errors (training/inference mismatch)
- No mel-spectrogram bias (mel-spectrogram is not an optimal representation)

Fully End-to-End TTS

ClariNet: Parallel Wave Generation In End-to-end Text-to-speech



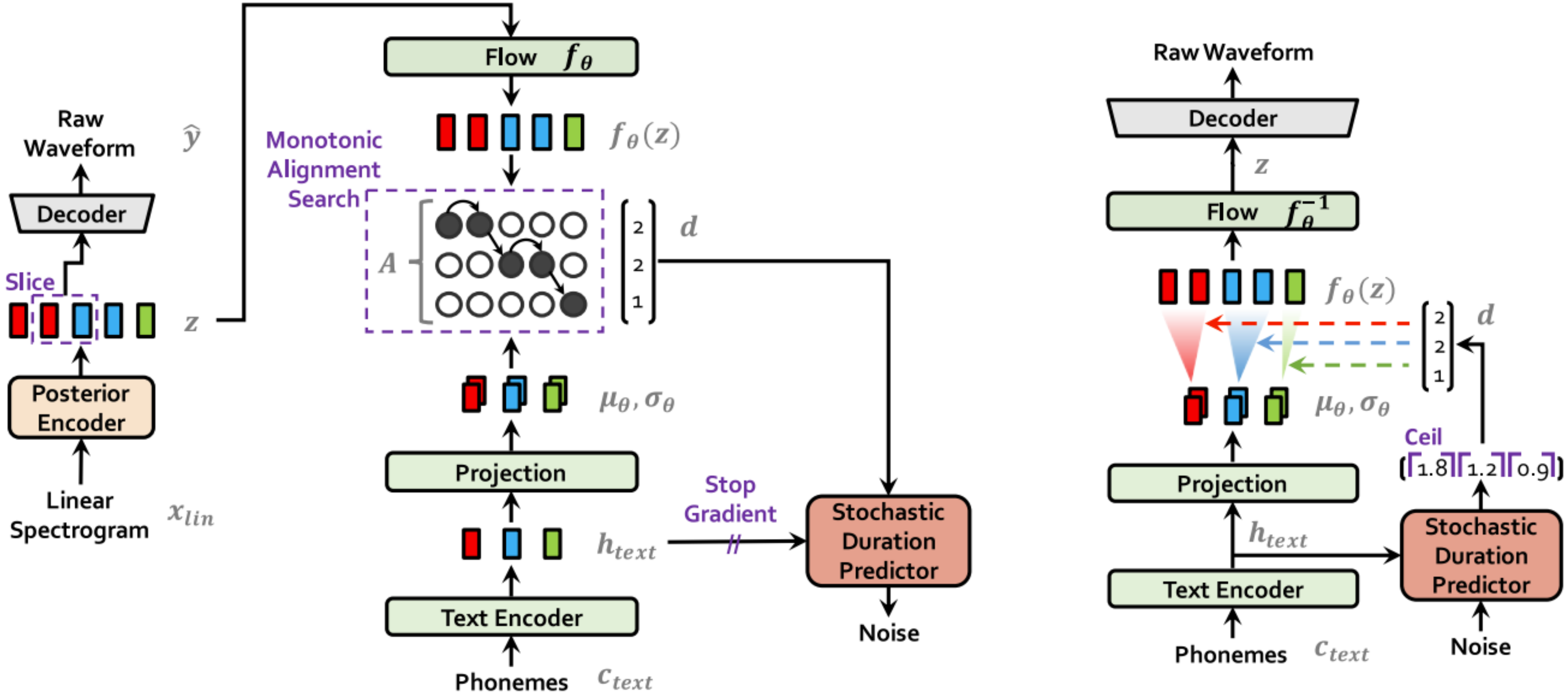
(a) Text-to-wave architecture (b) Bridge-net (c) Convolution block

https://openreview.net/pdf?id=HklY120cYm

Fully End-to-End TTS

VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End TTS

- VAE, Flow, GAN
- VAE: mel→waveform
- Flow for VAE prior
- GAN for waveform generation
- Monotonic alignment search

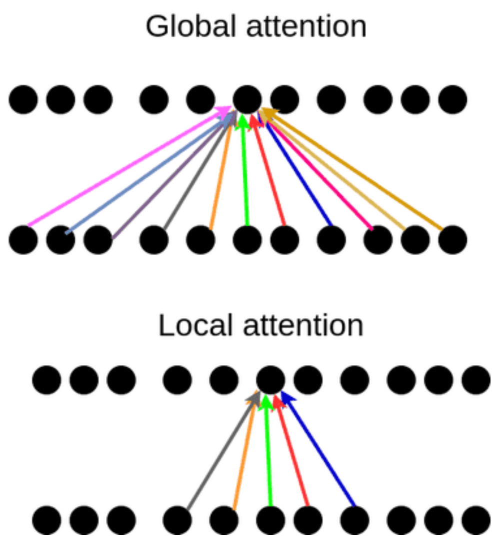


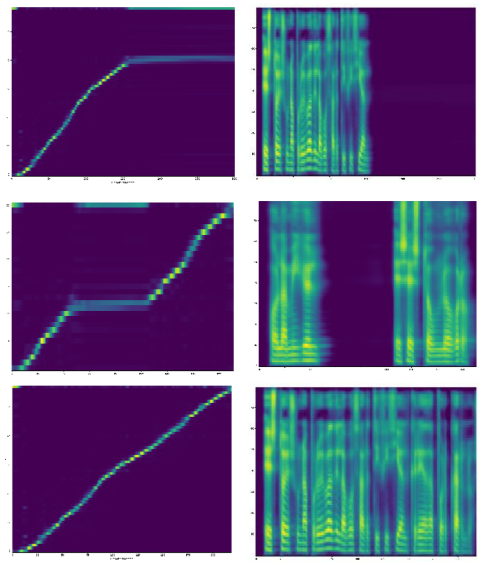
https://arxiv.org/pdf/2106.06103.pdf

Attention and Alignment

- **Attention** is a mechanism in machine learning models that allows the model to focus on specific parts of the input sequence when making predictions.

- **Alignment** refers to the relationship between words in the input and output sequences. It ensures that the model understands which parts of the input correspond to which parts of the output.
  - In translation, alignment ensures that the translated words correspond correctly to the words in the original language.





Why Attention and Alignment Matter?

- Attention helps the model better understand and capture dependencies between words in the input sequence.
- Particularly useful when input and output sequences have different lengths, allowing the model to align information appropriately.

Advanced topics in TTS

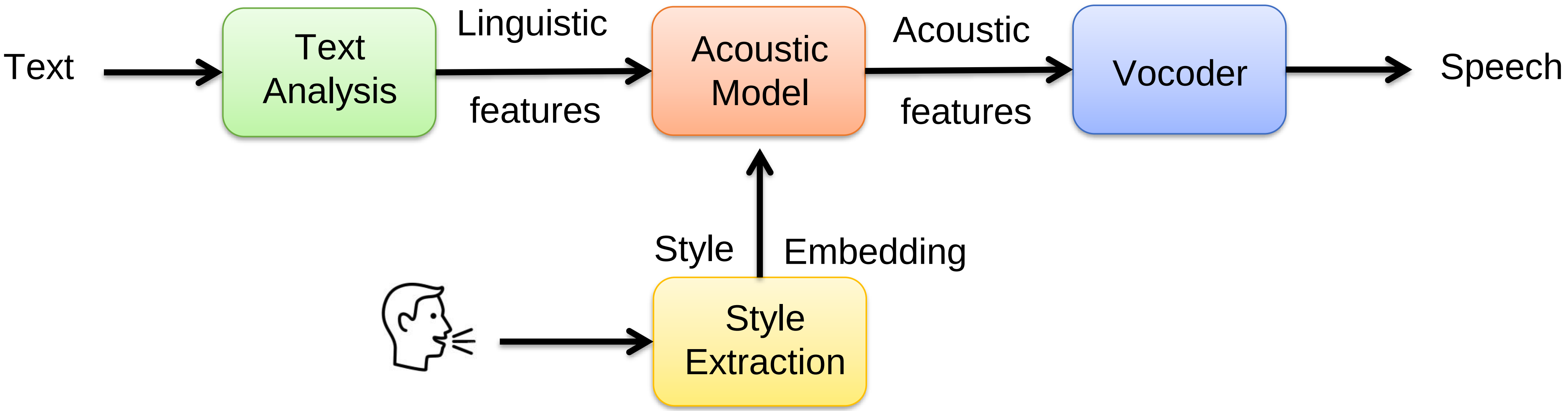
Advanced topics in Neural TTS

- ❑ Expressive TTS
- ❑ Controllable TTS
- ❑ Adaptative TTS

Expressive TTS

Expressiveness:

- what to say → Characterized by content
- who to say → speaker/timbre
- how to say → prosody/emotion/style
- where to say → noisy environment



(duration, pitch, sound volume, speaker, style, emotion, etc)

Expressive TTS

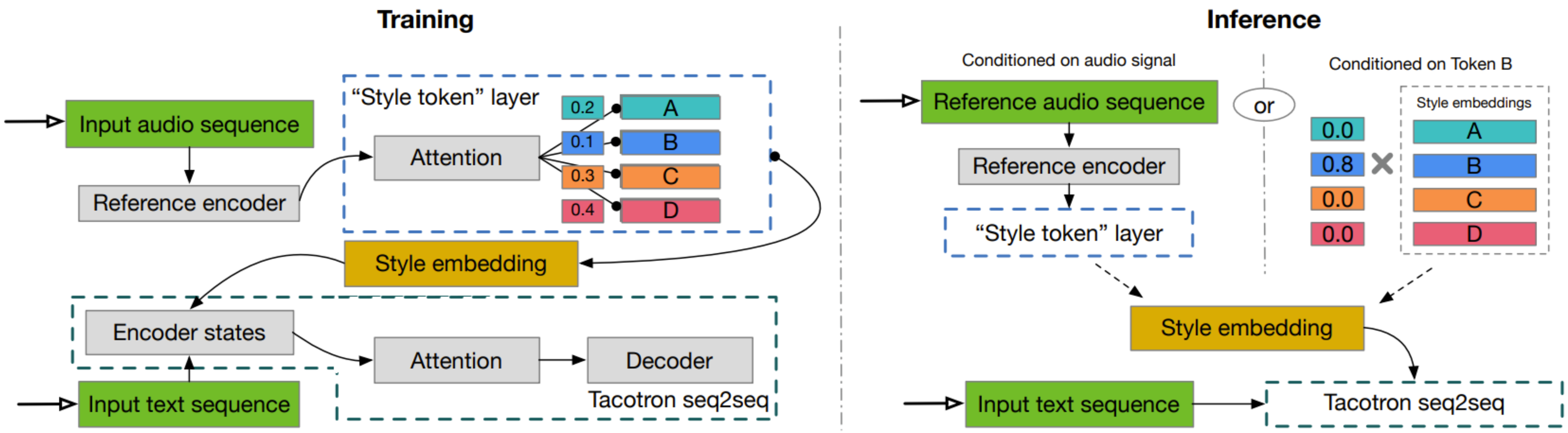
Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis

During training:

- the log-mel spectrogram of the training target is fed to the reference encoder followed by a style token layer.
- The resulting style embedding is used to condition the Tacotron text encoder states.

During inference:

- feed an arbitrary reference signal to synthesize text with its speaking style.



https://proceedings.mlr.press/v80/wang18h/wang18h.pdf

Controllable TTS

Adjustable Parameters

- TTS systems allow control over voice characteristics like pitch, rate, and volume.

Syntax Markup

- Adding annotations or tags in the input text can control aspects like emphasis, pauses, or pronunciation.

Prosody Manipulation

- Direct control over intonation, rhythm, and stress patterns is available in some TTS systems.

Customization and Training

- Advanced systems permit customization and training for specific voices, accents, or speech styles, offering more nuanced control over the output.

Voice-Controlled

Controllable TTS

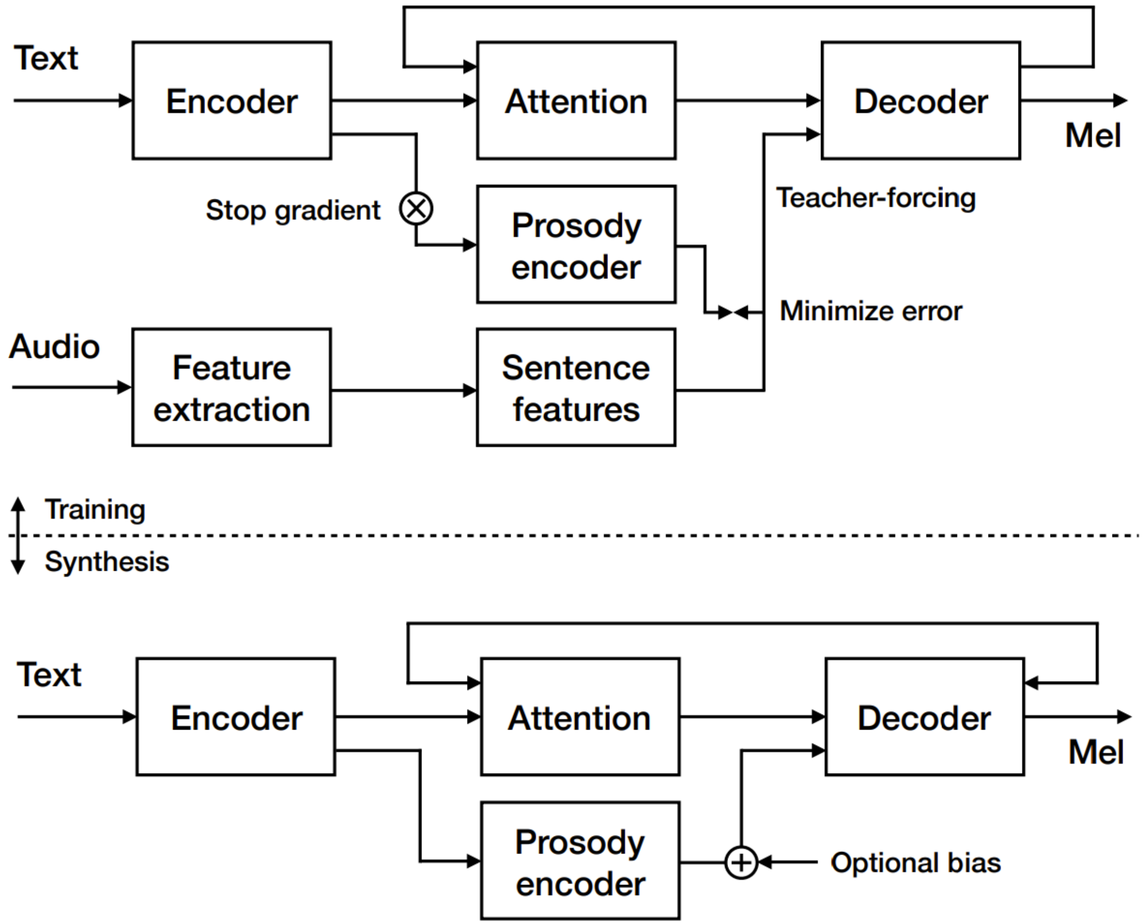
Controllable neural text-to-speech synthesis using intuitive prosodic features

training phase

- the prosody encoder learns to predict the sentence-wise prosodic features
- the decoder is conditioned on the ground-truth features (teacher-forcing). T

inference phase

- prosody encoder predicts prosodic features to condition the decoder, with an additional bias option for prosody control.



https://arxiv.org/pdf/2009.06775.pdf

Adaptive TTS

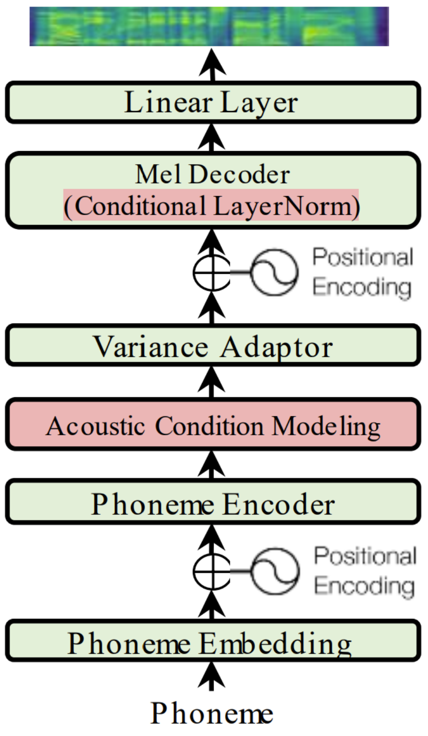
Empower TTS for everyone

- Pre-training on multi-speaker TTS model
- Fine-tuning on speech data from target speaker
- Inference speech for target speaker

Challenges

- To support diverse customers, the source model needs to be generalizable enough
- The target speech may be diverse (different acoustics/styles/languages)

Adaptive TTS



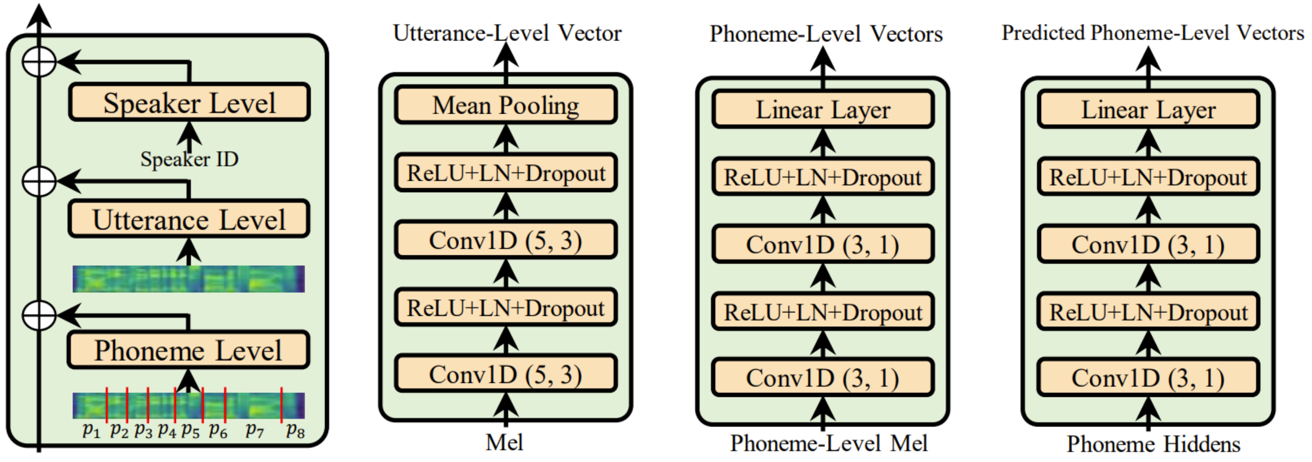
AdaSpeech: Adaptive Text to Speech for Custom Voice

Acoustic condition modeling

- Model diverse acoustic conditions at speaker/utterance/phoneme level
- Support diverse conditions in target speaker

Conditional layer normalization

- To fine-tune as small parameters as possible while ensuring the adaptation quality



(a) Overall. (b) Utterance level. (c) Phoneme level. (d) Phoneme level.

https://arxiv.org/pdf/2103.00993.pdf

TTS Model Evaluation

Objective Evaluation	Subjective Evaluation
Mel Cepstral Distortion (MCD)	MUSHRA (Multiple Stimuli with Hidden Reference and Anchor)
root mean square error (RMSE)	Mean Opinion Score (MOS)
Short-Time Objective Intelligibility (STOI)	
Perceptual Evaluation of Speech Quality (PESQ)	
Segmental Signal-to-Noise Ratio (SNRseg)	
etc.	

TTS Demos

Festival

http://www.cstr.ed.ac.uk/projects/festival/morevoices.html

Cereproc

https://www.cereproc.com/en/products/voices

TTS & STT for all languages

There are 7,000+ languages in the world, but popular commercialized speech services only support hundreds of languages

Google Cloud

Amazon Polly

aws

IBM Watson™

Microsoft Azure

Please, don’t forget to send feedback:

https://bit.ly/bme-dl

Thank you
for your attention

Dr. Mohammed Salah Al-Radhi

malradhi@tmit.bme.hu

12 November 2024