ELEC-C5220
Lecture 9:
Speech synthesis

Machine learning in information technology

Aalto University
School of Electrical
Engineering

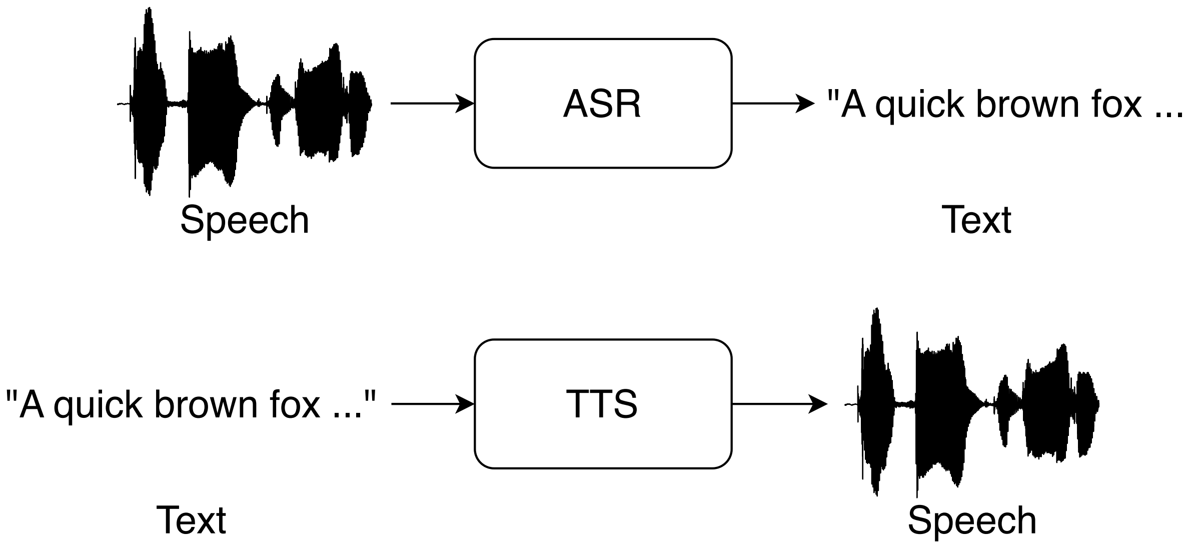
Lauri Juvela

14.3.2024

Lecture 9 content

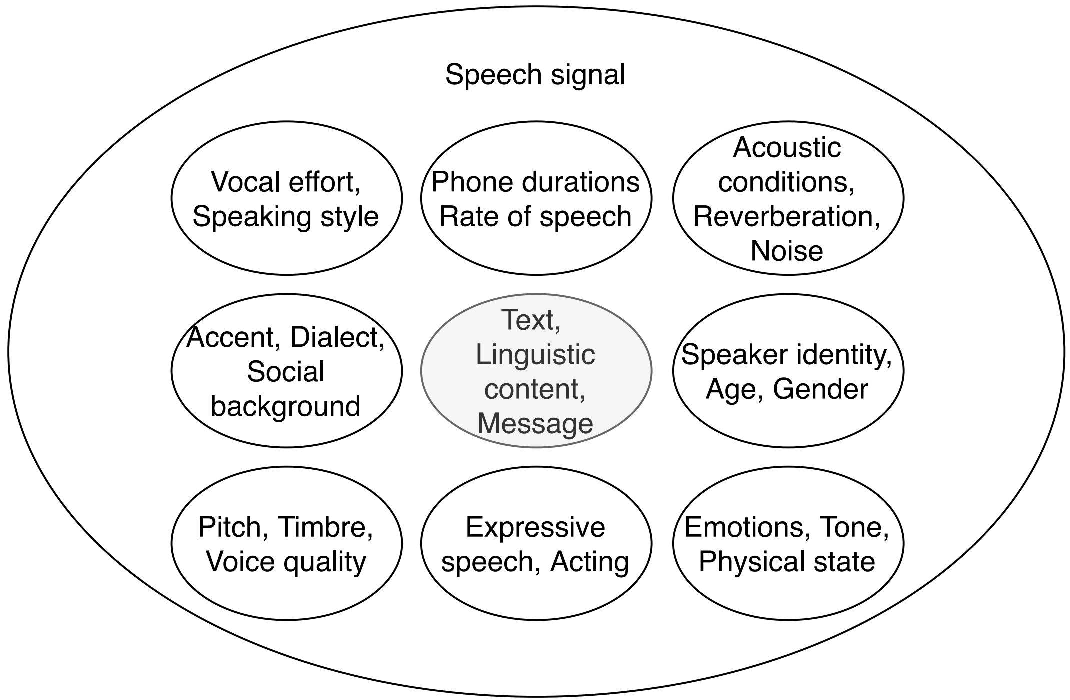
- **Speech synthesis**
- **Acoustic model**
  - Text to mel-spectrogram
  - Tacotron 2
- **Waveform model**
  - Mel-spectrogram to speech
  - Neural vocoders
  - HiFi-GAN
- **Voice cloning**

Speech synthesis and recognition



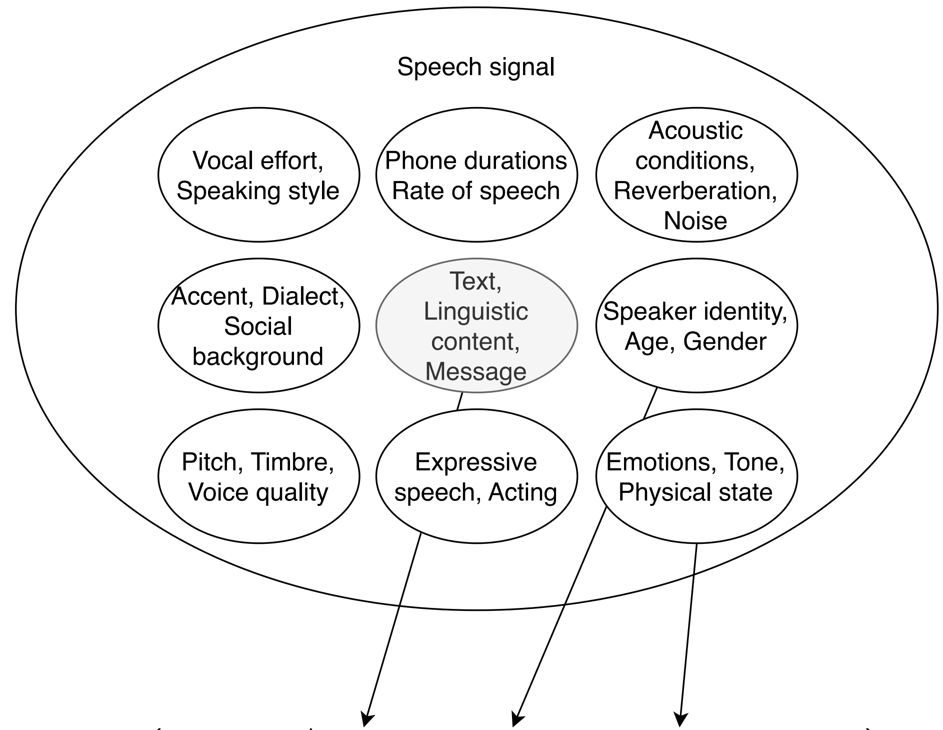
- **Automatic Speech Recognition (ASR) – many-to-one mapping**
- **Text-to-speech synthesis (TTS) – one-to-many mapping**

Attributes in speech signal



One-to-many mapping problems

- The same text can be read in many equally acceptable ways even by the same speaker
- Synthesis by averaging over all possible conditions gives unrealistic results
- Conditioning helps if you have labels
- Generative models help



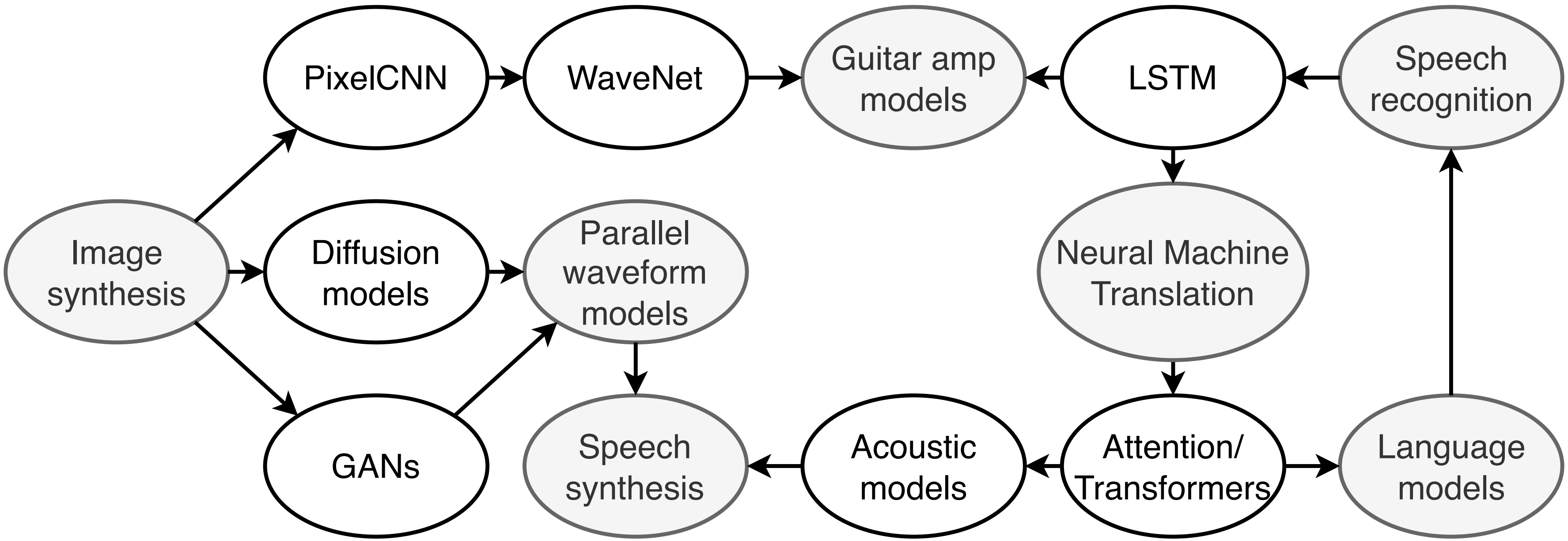
$$p(\text{speech}|\text{text}, \text{speaker}, \text{emotion}, \ldots)$$

$$\text{speech} = f(\text{text}, \text{speaker}, \text{emotion}, \ldots; \theta)$$

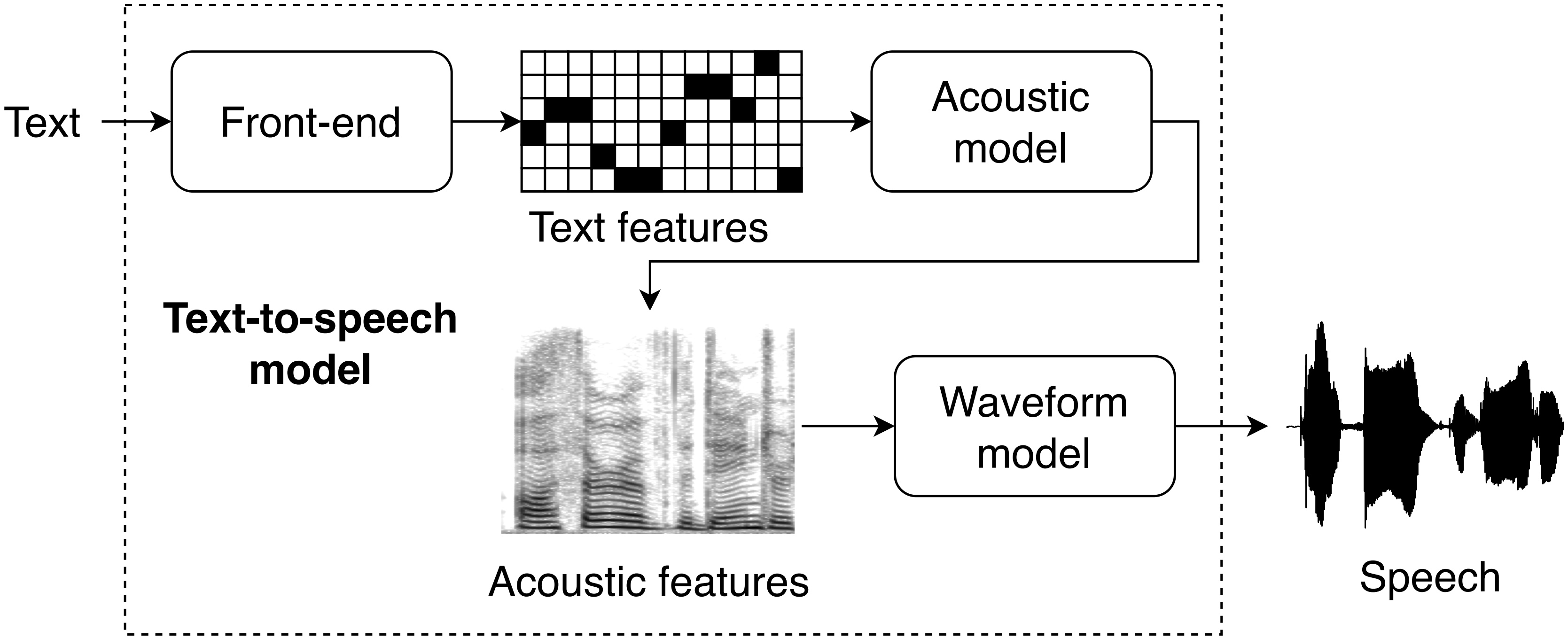
Speech synthesis applications

- Screen readers and assistive devices
- Voice prostheses
- Speech interfaces
- Voice chatbots for customer service
- Conversational AI
- Voice cloning for entertainment, virtual avatars, DeepFakes
- …

Speech synthesis - related fields and technology transfer



Text-to-speech systems



TTS front-end

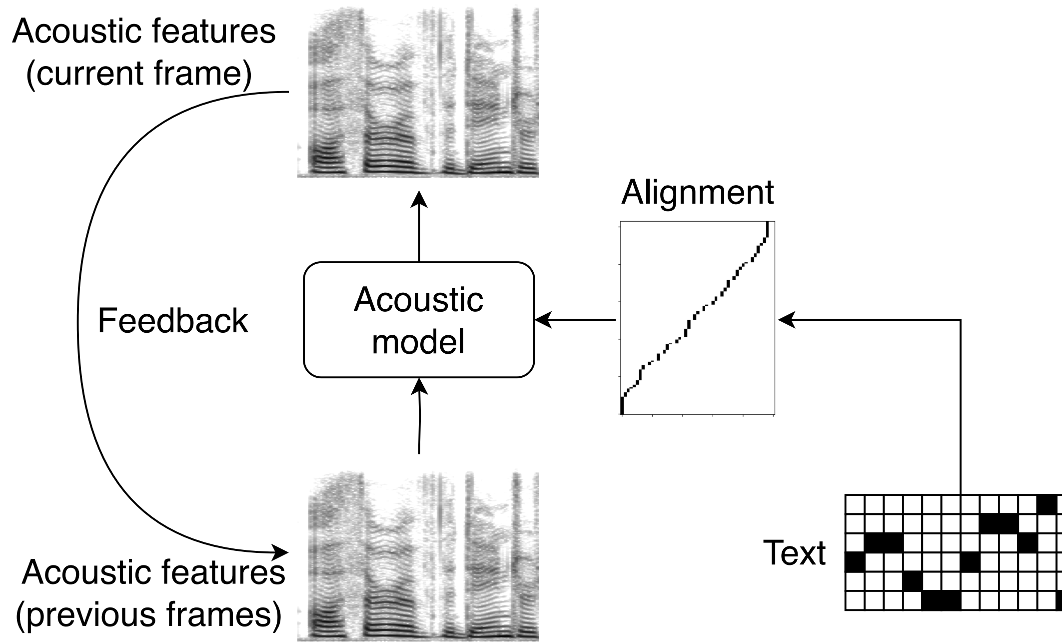
- **Pronounciation dictionary with letter-to-sound rules**
  - Very necessary in English, not so necessary in Finnish
  - G2P: grapheme-to-phoneme
- **Text normalisation, spell out numbers and abbreviations etc.**
  - Mr. -> Mister
  - Etc. -> et cetera
  - Today is March 14th -> Today is March fourteenth

TTS acoustic model

- **Map text sequences to acoustic feature sequences**
- **Acoustic model has to somehow solve the *sequence alignment problem***
- **Mel-spectograms are commonly used the acoustic features nowadays**
- **Previously common: pitch and vocal tract envelope features for parametric synthesis**
- **Often used in tandem with a duration model (how long should each text token last in seconds / acoustic frames)**

Autoregressive acoustic model

- Use cross-attention to align text with spectrogram
- Predict current spectrogram frame frame from previous frames
- Similar concept to Whisper ASR (L8) and LSTM Language model (L7)



TTS acoustic modeling with Tacotron 2

NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

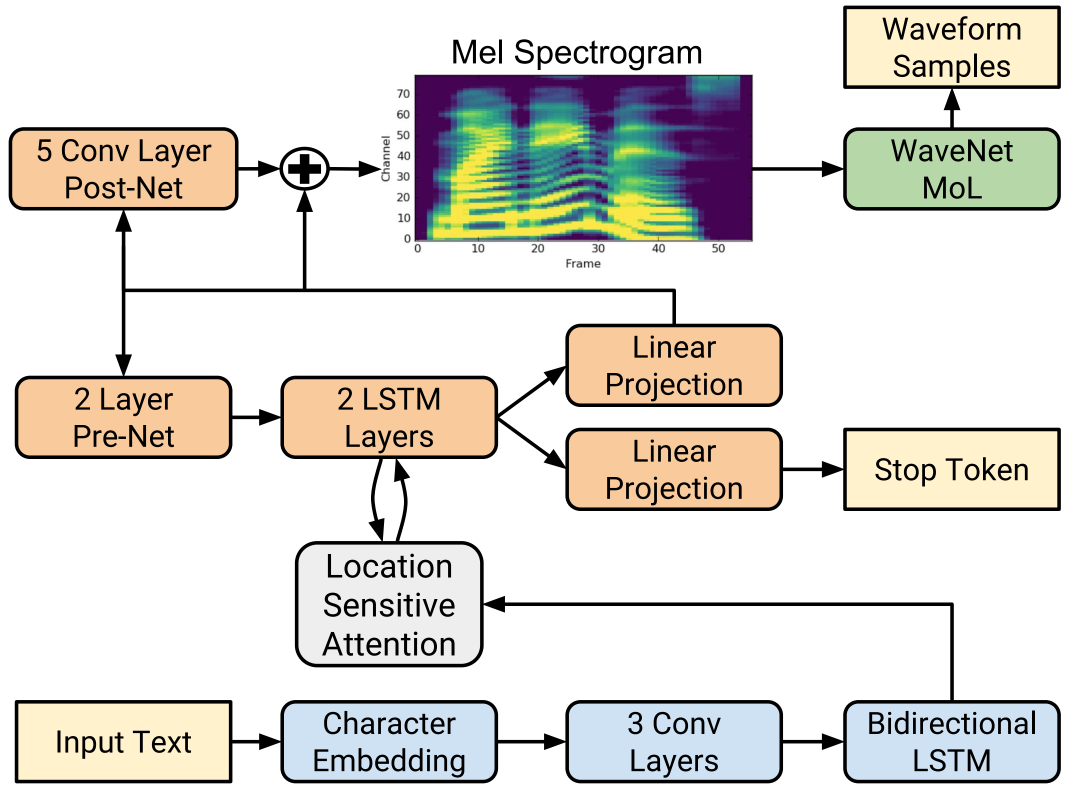
Jonathan Shen[1], *Ruoming Pang*[1], *Ron J. Weiss*[1], *Mike Schuster*[1], *Navdeep Jaitly*[1], *Zongheng Yang*[*2], *Zhifeng Chen*[1], *Yu Zhang*[1], *Yuxuan Wang*[1], *RJ Skerry-Ryan*[1], *Rif A. Saurous*[1], *Yannis Agiomyrgiannakis*[1], *and Yonghui Wu*[1]

[1]Google, Inc., [2]University of California, Berkeley,
{jonathanasdf,rpang,yonghui}@google.com

https://google.github.io/tacotron/publications/tacotron2/

Tacotron 2 architecture

- **LSTM text encoder on characters or phonemes**
- **Cross-attention to align text and spectrogram (one head, one layer)**
- **Autoregressive LSTM spectrogram decoder**

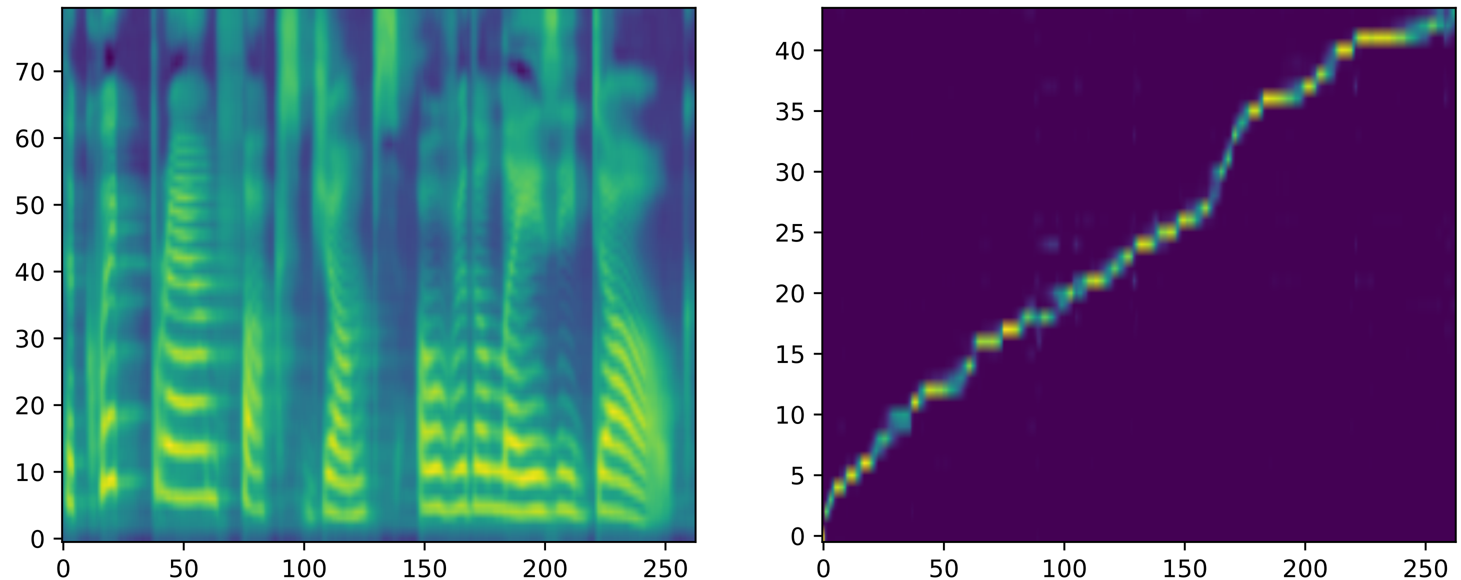


Tacotron 2 evaluation

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

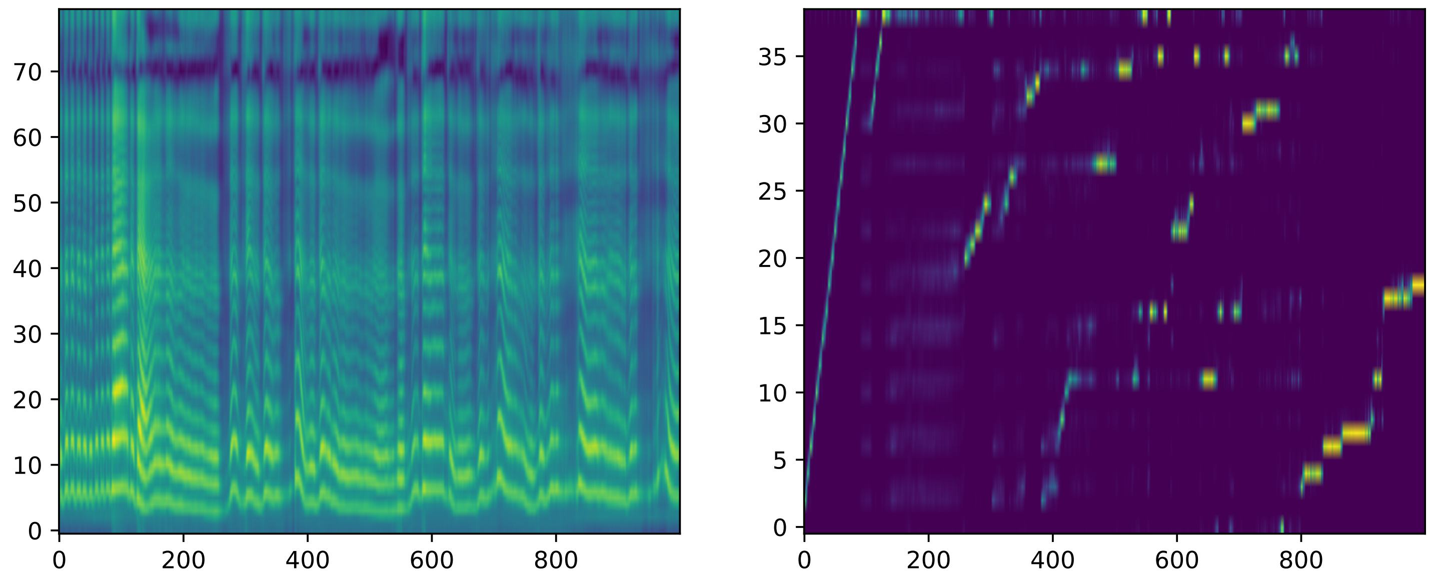
Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

Tacotron 2: generated mel-spectrogram and attention plot



"The quick brown fox jumps over the lazy dog."

Tacotron 2: example of attention failure



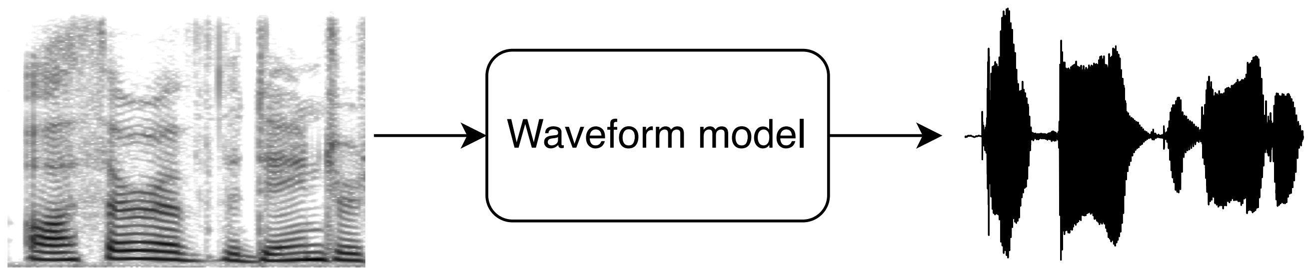
"The the the the the the the the the the"

TTS waveform model

- We can’t listen to mel spectrograms directly, but need some way to generate waveforms
- Magnitude spectrograms are missing phase information and phase is difficult to invent from scratch
- Waveform synthesis models are also known as Vocoders (from voice coders)

Mel-spectrum to waveform

(Batch, Channels=Freq-bins, Frames) (Batch, Channels=1, Samples)



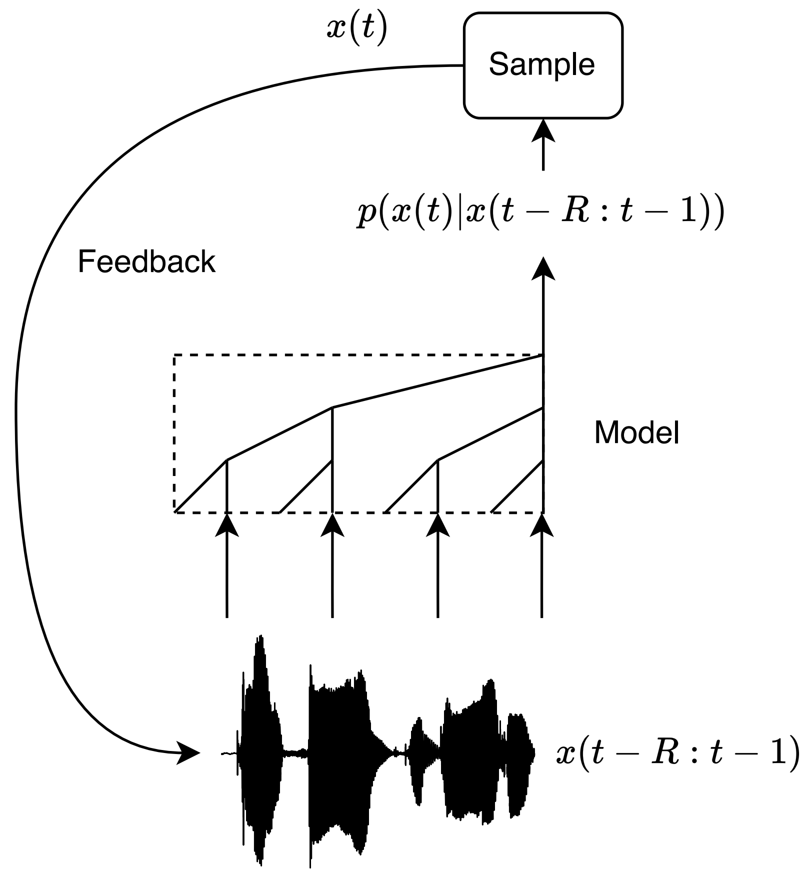
Mel-spectrogram Speech

Samples = Hop-size (stride) $\times$ Frames

- **Model needs to upsample from frame rate (around 100 Hz) to sample rate (around 20 000 Hz), total factor of 200!**
- **Usually done in multiple stages (progressive upsampling)**

Autoregressive waveform models

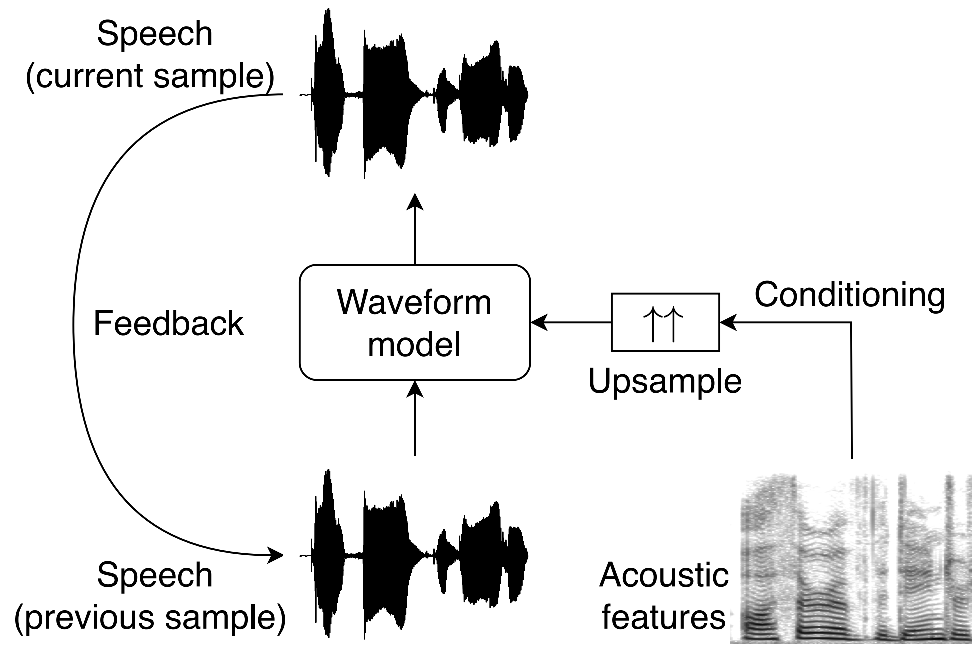
- **Predict the distribution of next sample amplitude, given previous amplitude values**
- **Similar to the LSTM language model in Lecture 6!**
- **Use dilated convolutions to deal with long sequences (1s is 16 000 samples at 16kHz rate)**



Aalto University
School of Electrical
Engineering

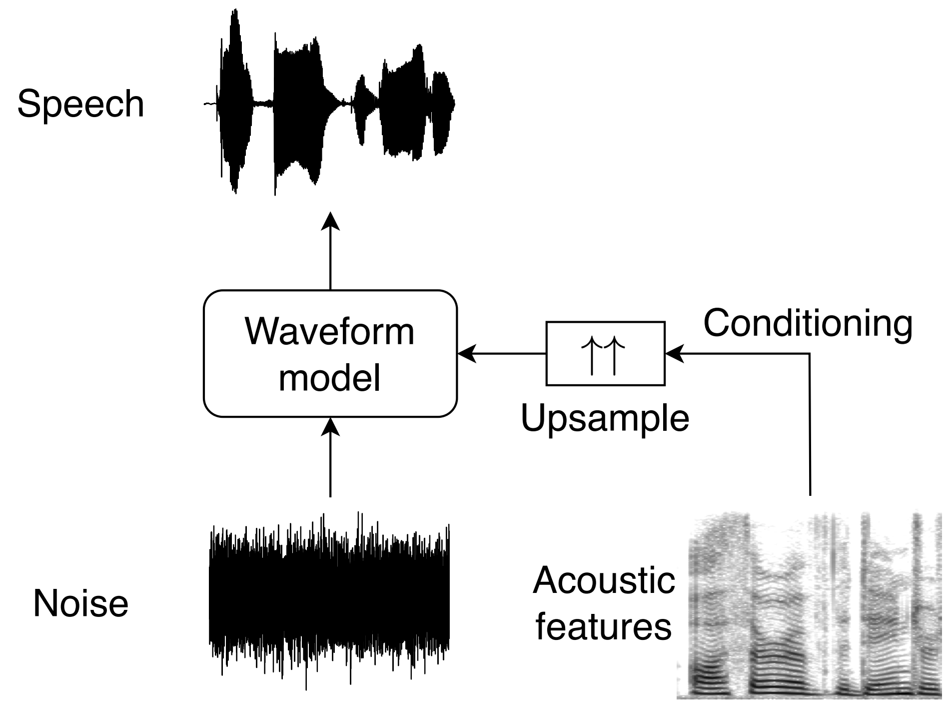
Autoregressive waveform models

- **Condition on mel spectrograms to make the model useful in TTS**
- **WaveNet and WaveRNN were were the first big success stories**



Parallel waveform synthesis

- **It's more convenient to generate the full waveform sequence in a single forward pass**
- **Modern systems can do parallel syntehesis using diffusion models, GANs or neural flows**
- **We use HiFi-GAN in the exercise**



TTS waveform synthesis with HiFi-GAN

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

Jungil Kong
Kakao Enterprise
henry.k@kakaoenterprise.com

Jaehyeon Kim
Kakao Enterprise
jay.xyz@kakaoenterprise.com

Jaekyoung Bae
Kakao Enterprise
storm.b@kakaoenterprise.com

HiFi-GAN generator architecture

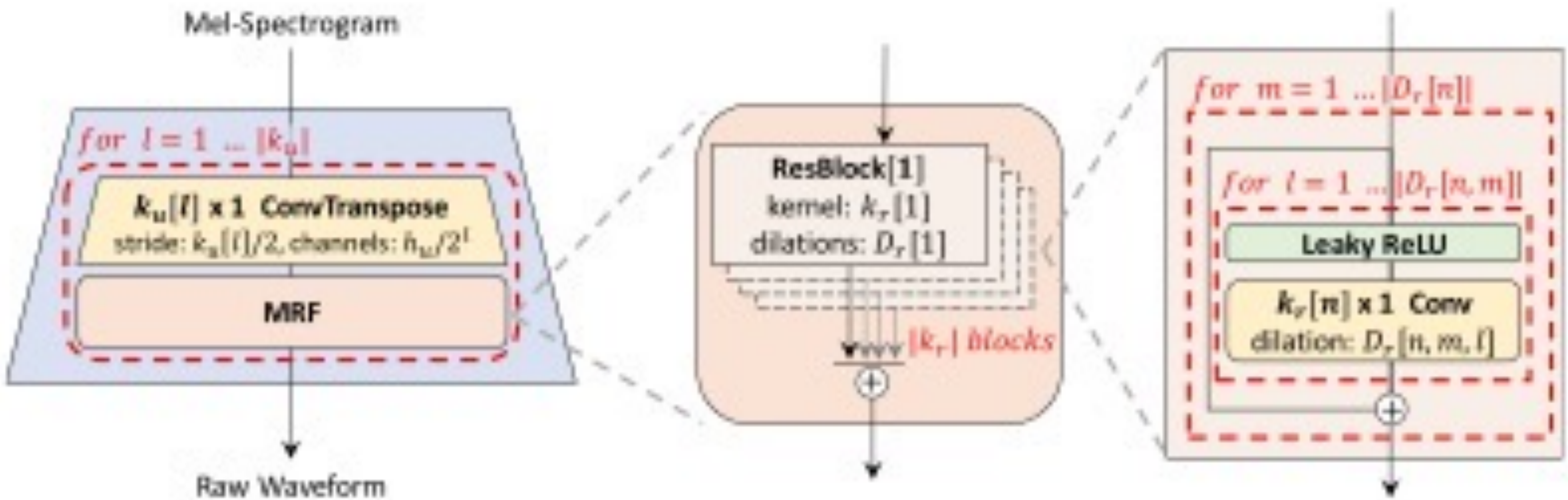


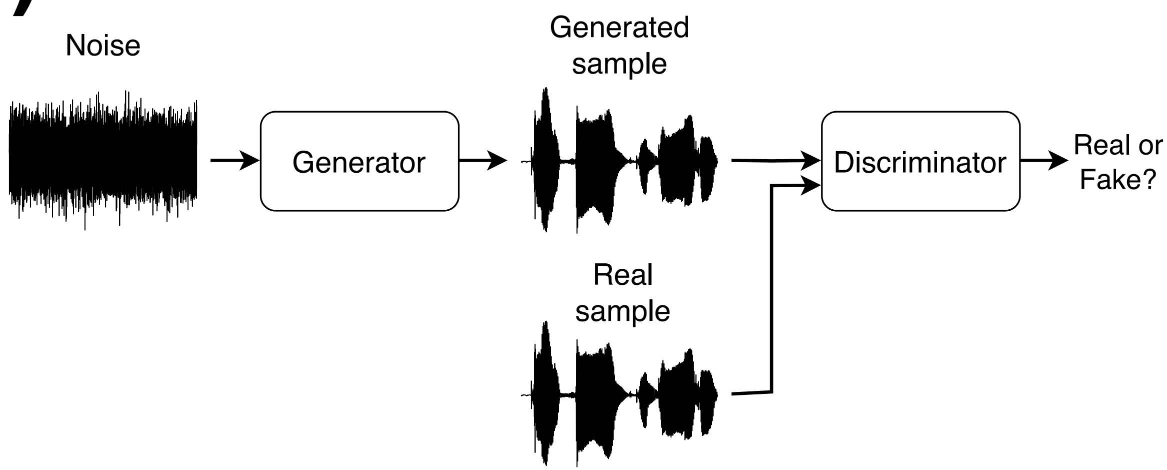
Figure 1: The generator upsamples mel-spectrograms up to $|k_u|$ times to match the temporal resolution of raw waveforms. A MRF module adds features from $|k_r|$ residual blocks of different kernel sizes and dilation rates. Lastly, the n-th residual block with kernel size $k_r[n]$ and dilation rates $D_r[n]$ in a MRF module is depicted.

HiFi-GAN evaluation

Table 1: Comparison of the MOS and the synthesis speed. Speed of n kHz means that the model can generate $n \times 1000$ raw audio samples per second. The numbers in () mean the speed compared to real-time.

Model	MOS (CI)	Speed on CPU (kHz)		Speed on GPU (kHz)		# Param (M)
Ground Truth	4.45 ($\pm$0.06)	–		–		–
WaveNet (MoL)	4.02 ($\pm$0.08)	–		0.07	($\times$0.003)	24.73
WaveGlow	3.81 ($\pm$0.08)	4.72	($\times$0.21)	501	($\times$22.75)	87.73
MelGAN	3.79 ($\pm$0.09)	145.52	($\times$6.59)	14,238	($\times$645.73)	4.26
HiFi-GAN $V1$	**4.36** ($\pm$0.07)	31.74	($\times$1.43)	3,701	($\times$167.86)	13.92
HiFi-GAN $V2$	4.23 ($\pm$0.07)	214.97	($\times$9.74)	16,863	($\times$764.80)	**0.92**
HiFi-GAN $V3$	4.05 ($\pm$0.08)	**296.38**	**($\times$13.44)**	**26,169**	**($\times$1,186.80)**	1.46

Generative adversarial networks (GANs)



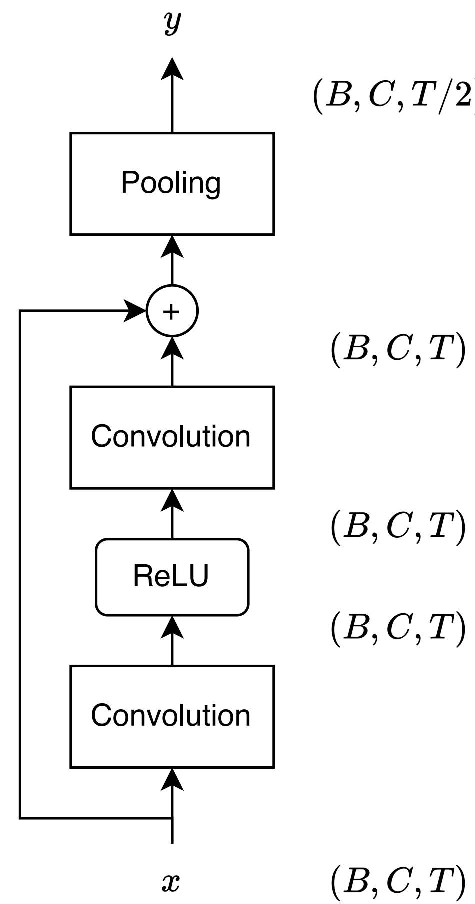
- **Generator transforms noise to look like it came from the real data distribution**
- **Discriminator attempts to classify between real and generated samples**

Discriminator design

- Classifier model architectures are generally suitable for Discriminator use
- Generator applies progressive upsampling
- Discriminator applies progressive downsampling
- Recap: Spoken Digit Classification from Lecture 3

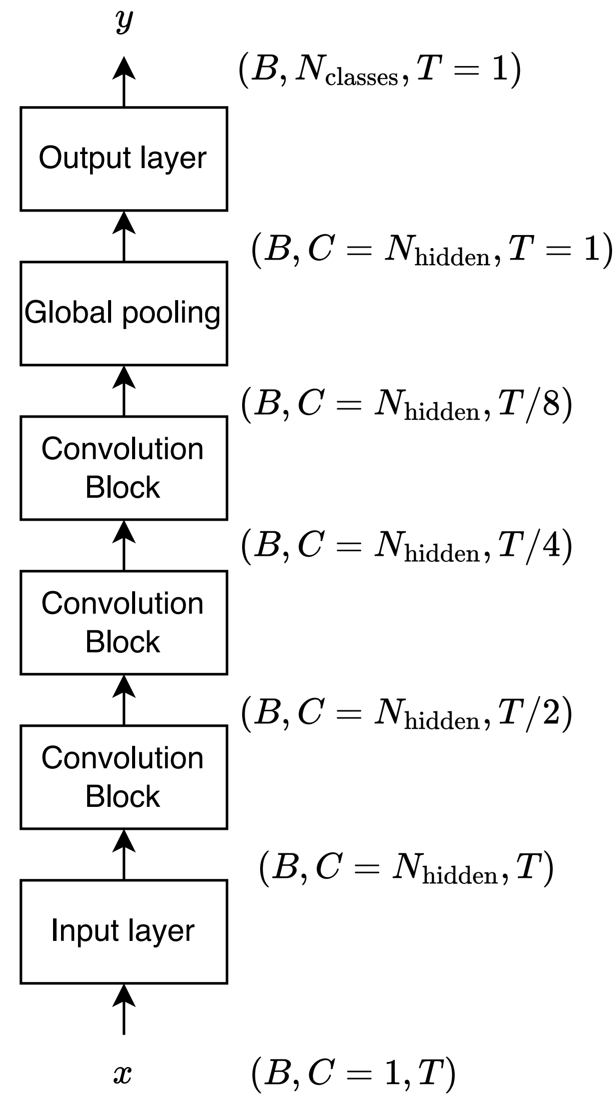
Convolution Block

- **Typical convolution layers (aka Blocks) contain**
  - Convolutions
  - Activations (ReLU)
  - Residual connections
  - Pooling (Max or Avg.)



CNN classifier model

- **Input layer embeds the data to hidden dimension**
- **Convolution layers learn representations and gradually downsample the input**
- **Global pooling deals with whatever sequence length remains**
- **Output layer projects to number of classes**



More courses on generative models

- **Training GANs is out-of-scope for this course, let's use a pre-trained generator model and skip the training**
- **CS-E4890 Deep Learning D**
    - Currently includes generative topics, will focus on DL basics from 2025
- **CS-E4891 Deep Generative Models D**
    - New course starting in spring 2025
    - Covers generative topics, including GANs, Autoregressive models, Diffusion, Variational autoencoders, etc.

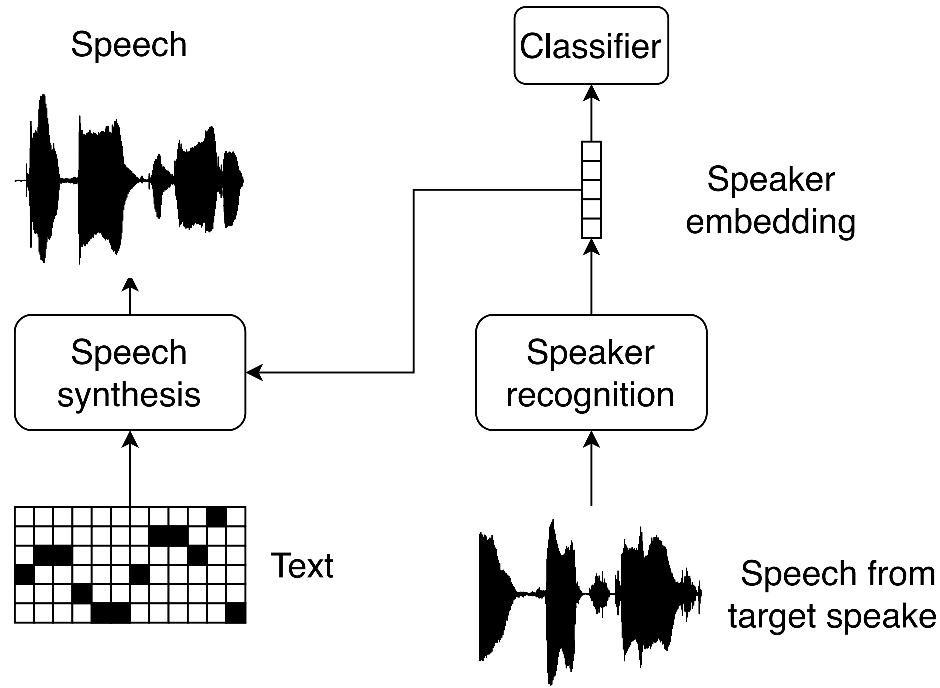
Weekly exercise

- Dissect a pre-trained TTS system made from a Tacotron 2 acoustic model and a HiFi-GAN vocoder
- Visualise how cross attention between speech and text works
- Track how the signal flows in the system

Voice cloning with TTS

Voice cloning system

- **Extract speaker embeddings from a speaker recognition system**
- **TTS system is trained on multiple speakers and conditioned on speaker embeddings**
- **Embeddings are content-agnostic – easy to enroll new speakers**



Reading list: Tacotron 2

Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

Paper: https://arxiv.org/abs/1712.05884

Demo: https://google.github.io/tacotron/publications/tacotron2/

Code: https://github.com/NVIDIA/tacotron2

Reading list: HiFi-GAN

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

Paper: https://arxiv.org/abs/2010.05646

Demo: https://jik876.github.io/hifi-gan-demo/

Code: https://github.com/jik876/hifi-gan

Lecture 9 summary

- **Speech synthesis**
- **Acoustic model**
  - Text to mel-spectrogram
  - Tacotron 2
- **Waveform model**
  - Mel-spectrogram to speech
  - Neural vocoders
  - HiFi-GAN
- **Voice cloning**