Introduction to Speech Technology

You (Neil) Zhang

you.zhang@rochester.edu

(Some slides are adapted from
http://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20 and
http://tts.speech.cs.cmu.edu/courses/11492/)

Outline

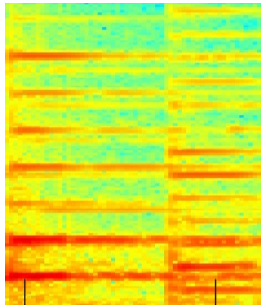
Introduction

Research Topics

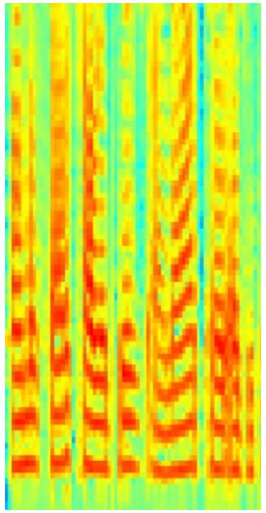
Future horizons

Q & A

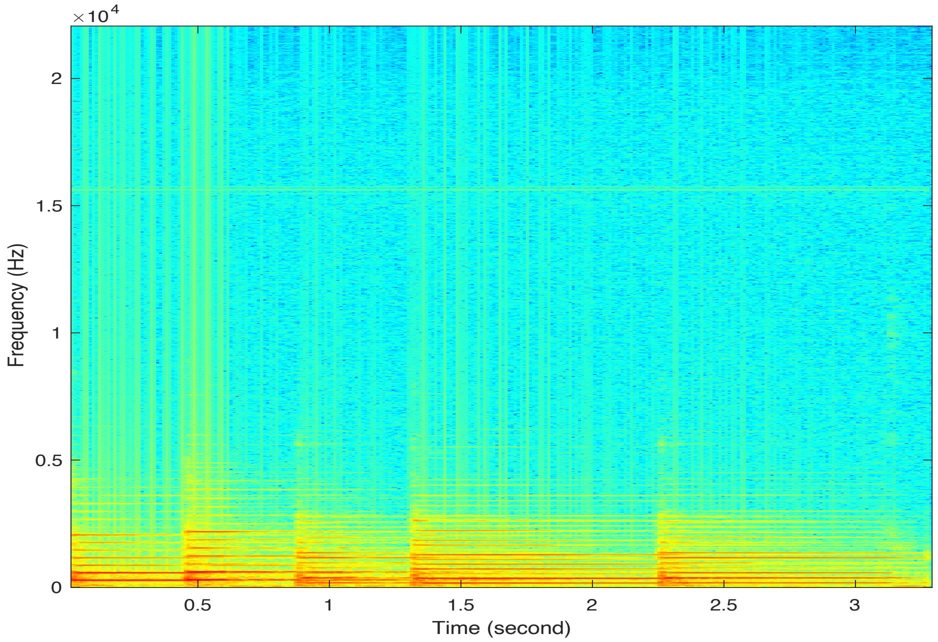
Audio Signals

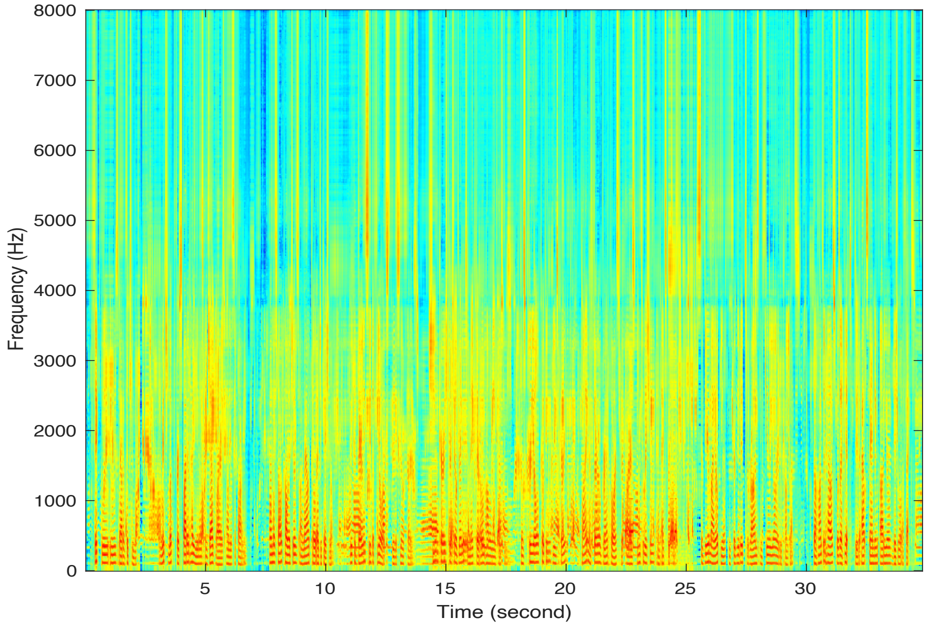


Music

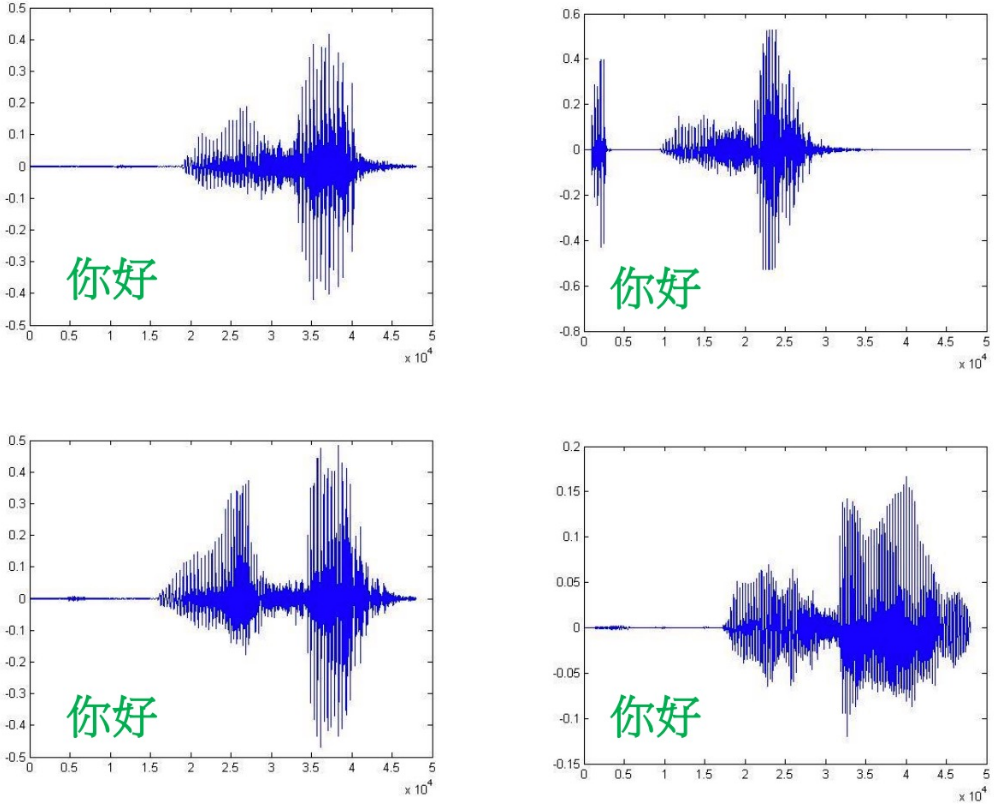


Speech





Why speech?

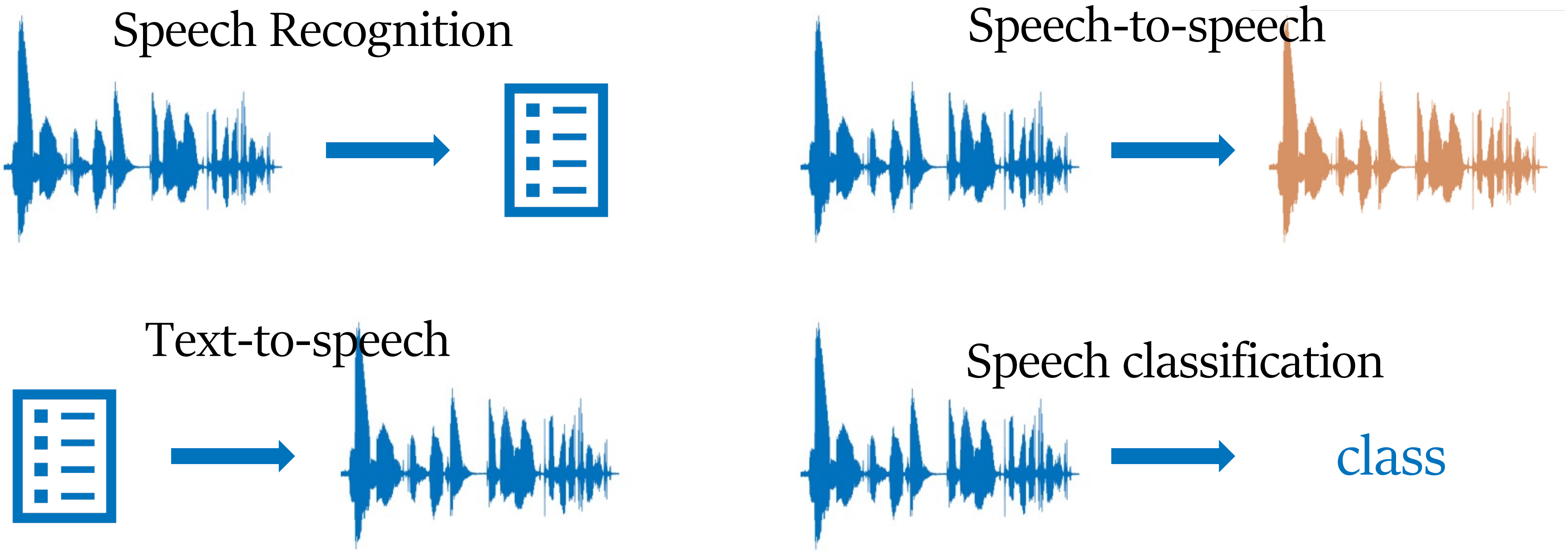


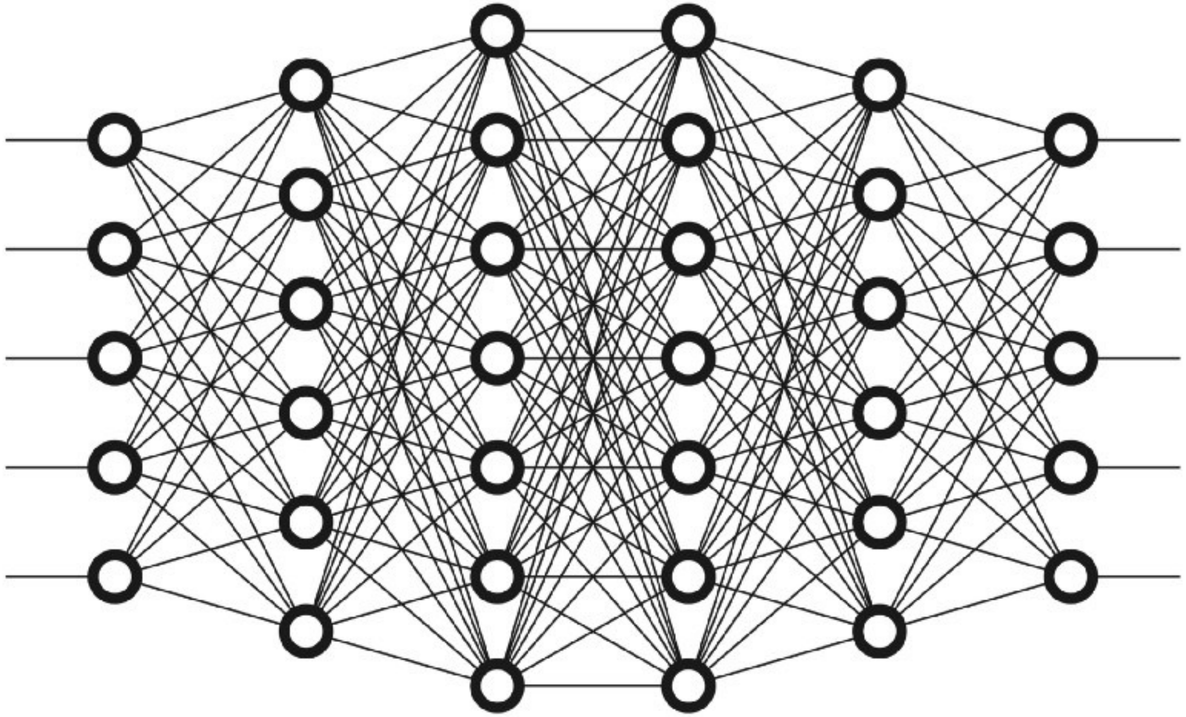
- Most natural way for human communication
- Hard to represent (You cannot speak the same twice)
- Hard to search

Speech Applications

- Google Maps
- Apple's Siri, Google Home, Amazon Echo/Alexa
- Screen readers
- Voice biometrics
- …

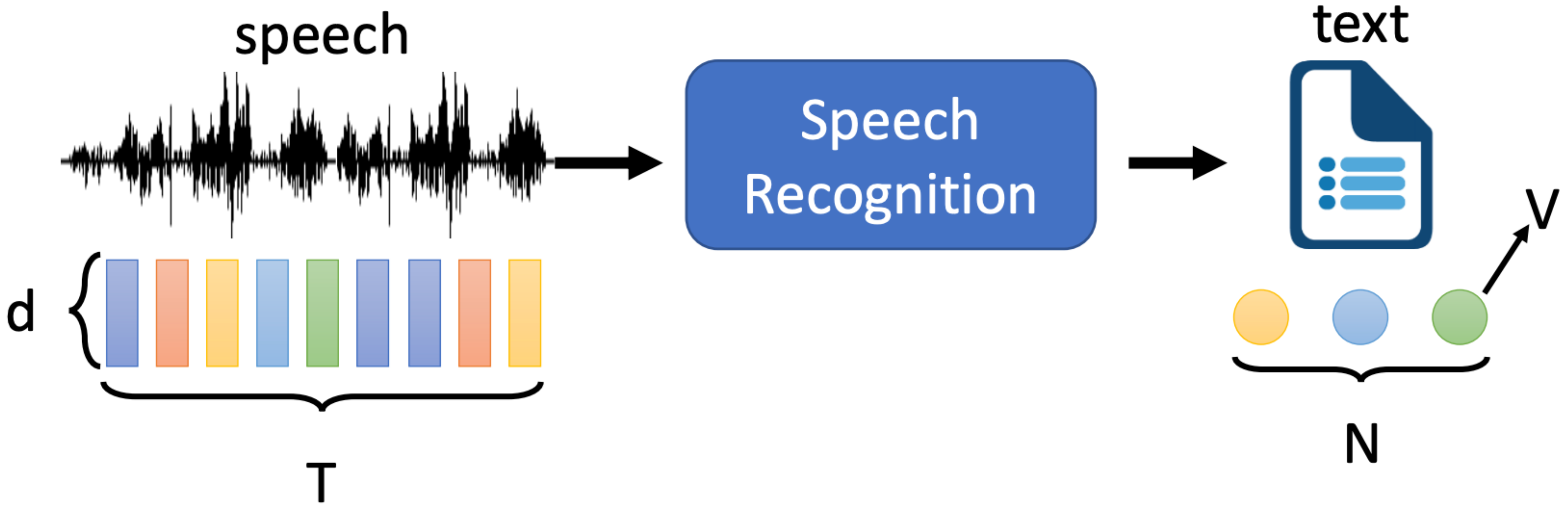
Overview of speech topics





- Besides training Deep Neural Networks, what does each topic care about?

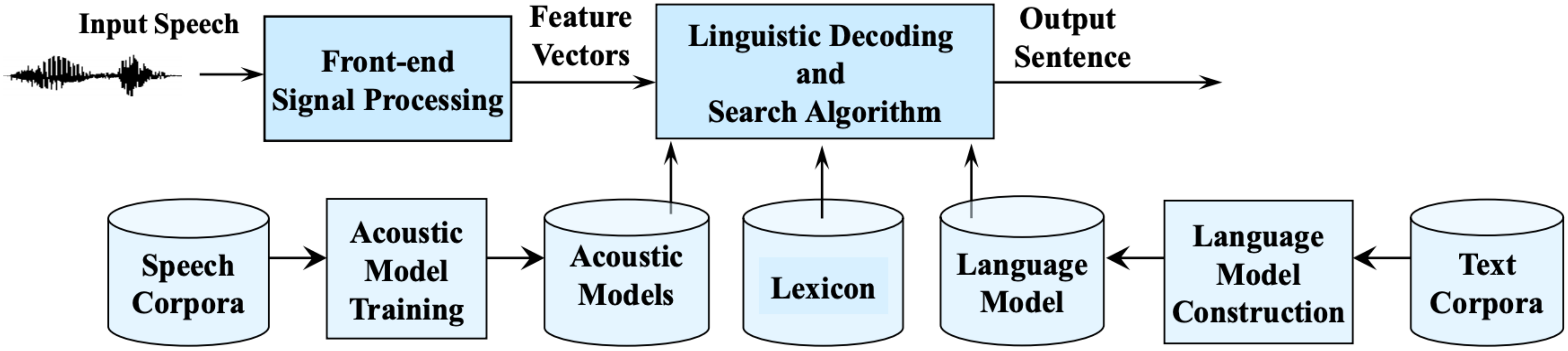
Speech recognition



- Speech: a sequence of vector (length T, dimension d)
- Text: a sequence of token (length N, V different tokens)

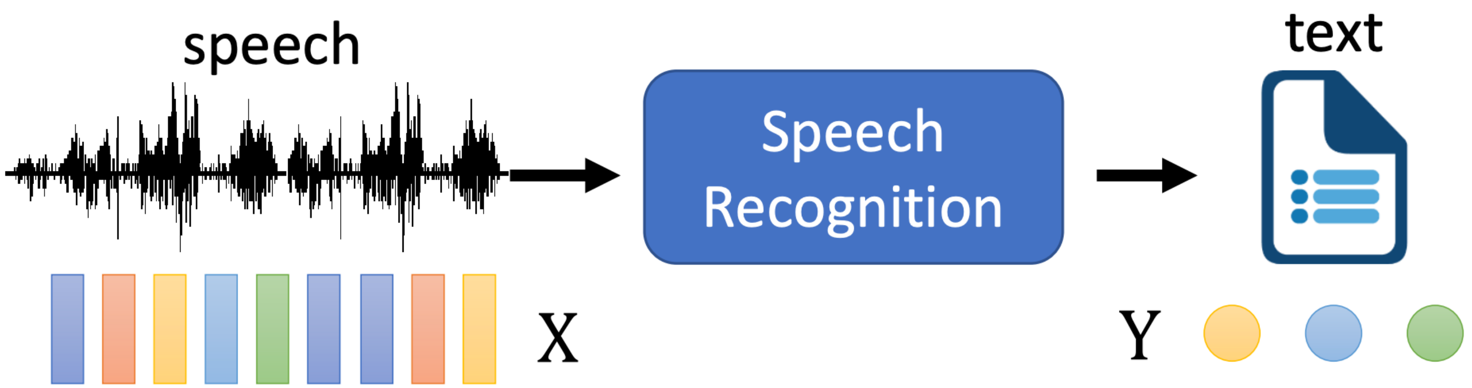
Speech recognition

Traditional Speech Recognition



Speech recognition

- HMM



$$Y^* = arg\max_{Y} P(Y|X)$$

Decode

$$= arg\max_{Y} \frac{P(X|Y)P(Y)}{P(X)}$$

$$= arg\max_{Y} P(X|Y)P(Y)$$

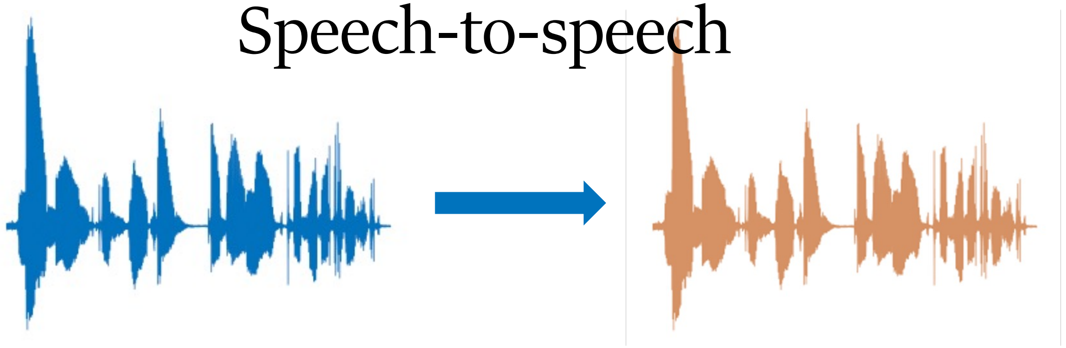
$P(X|Y)$: HMM

Acoustic Model

$P(Y)$:

Language Model

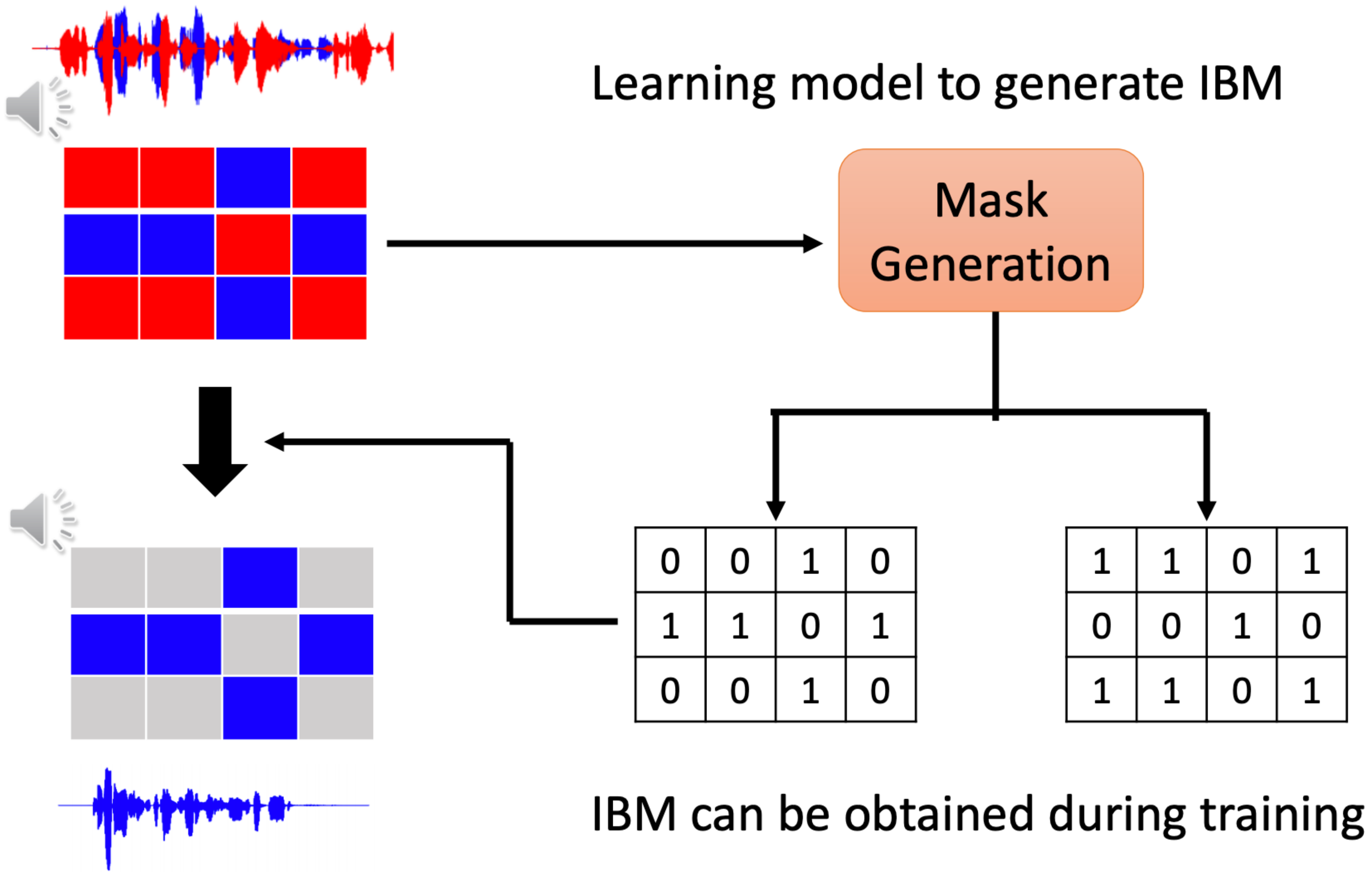
Speech-to-speech



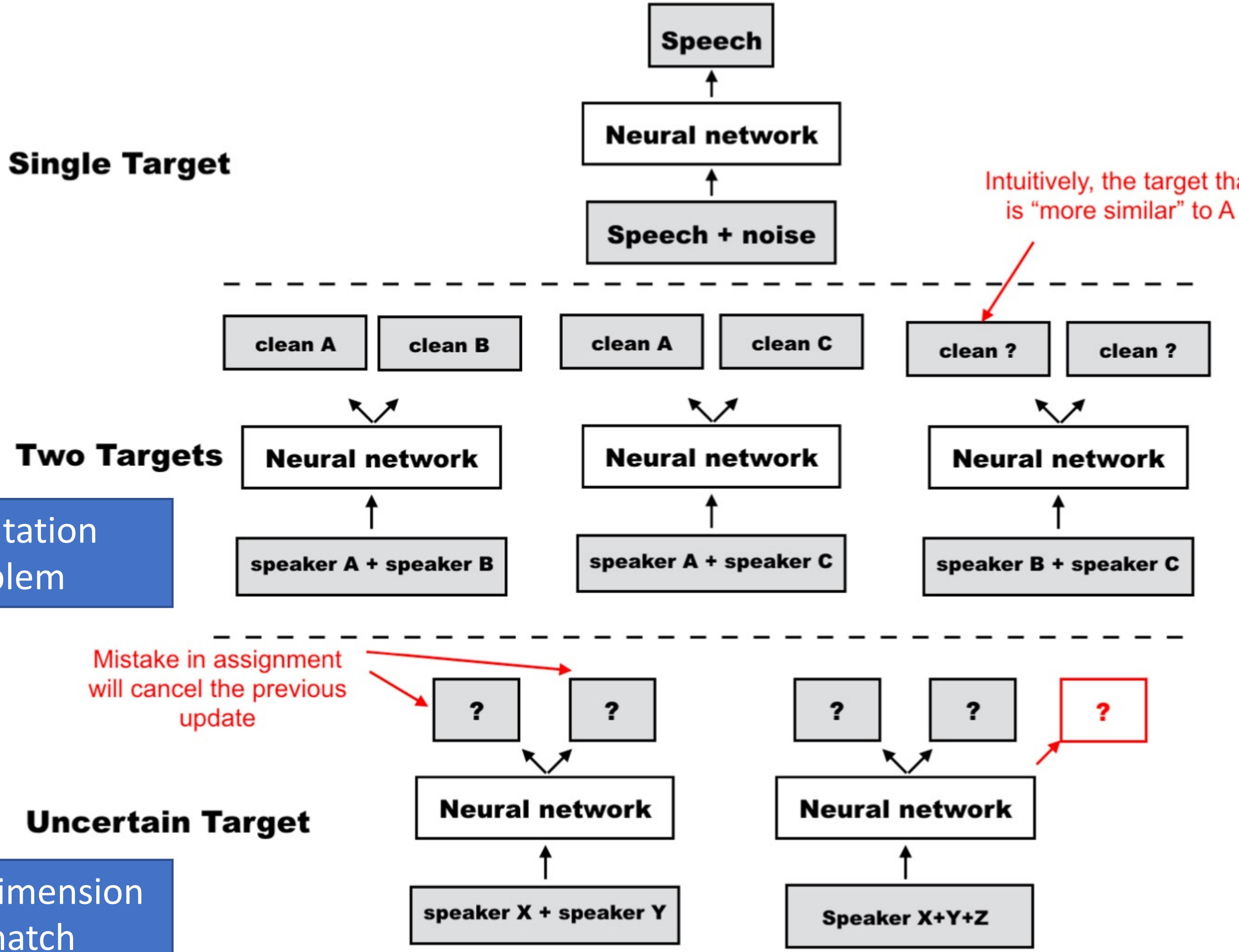
Speech separation

https://researcher.watson.ibm.com/researcher/view_group.php?id=2819

- Ideal binary mask



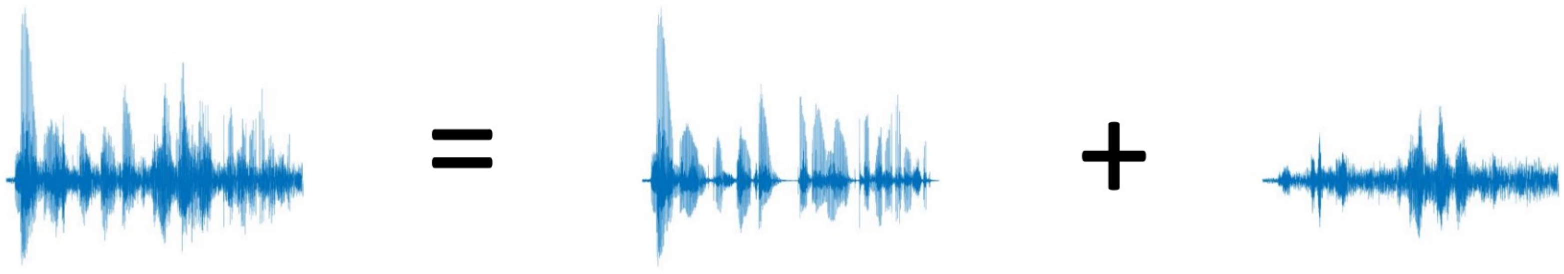
Two Problems in the speech separation task



Permutation problem

Output dimension mismatch

Speech Enhancement

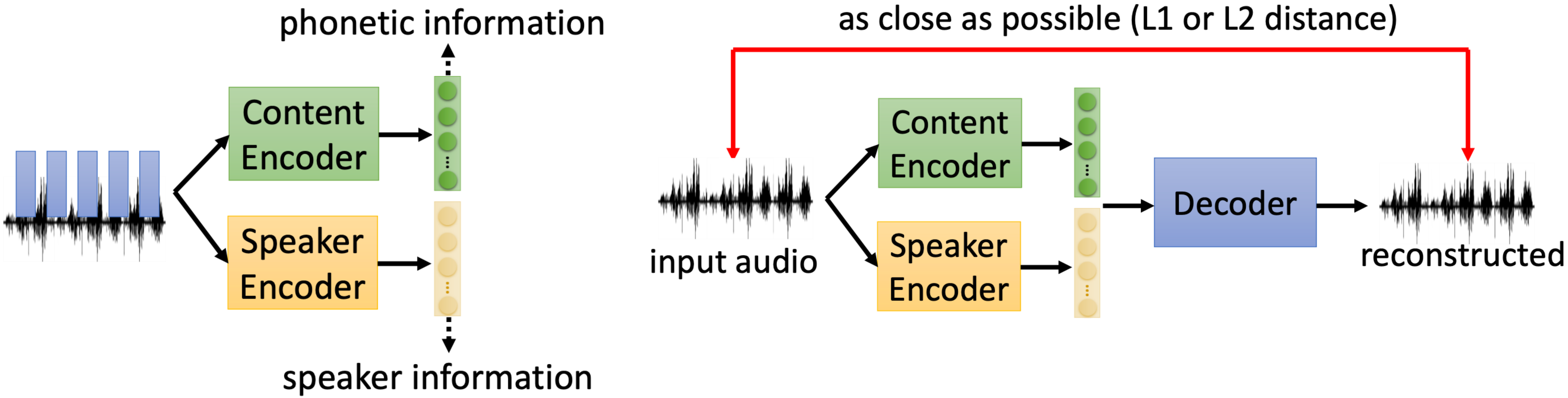


Noisy speech = clean speech + noise

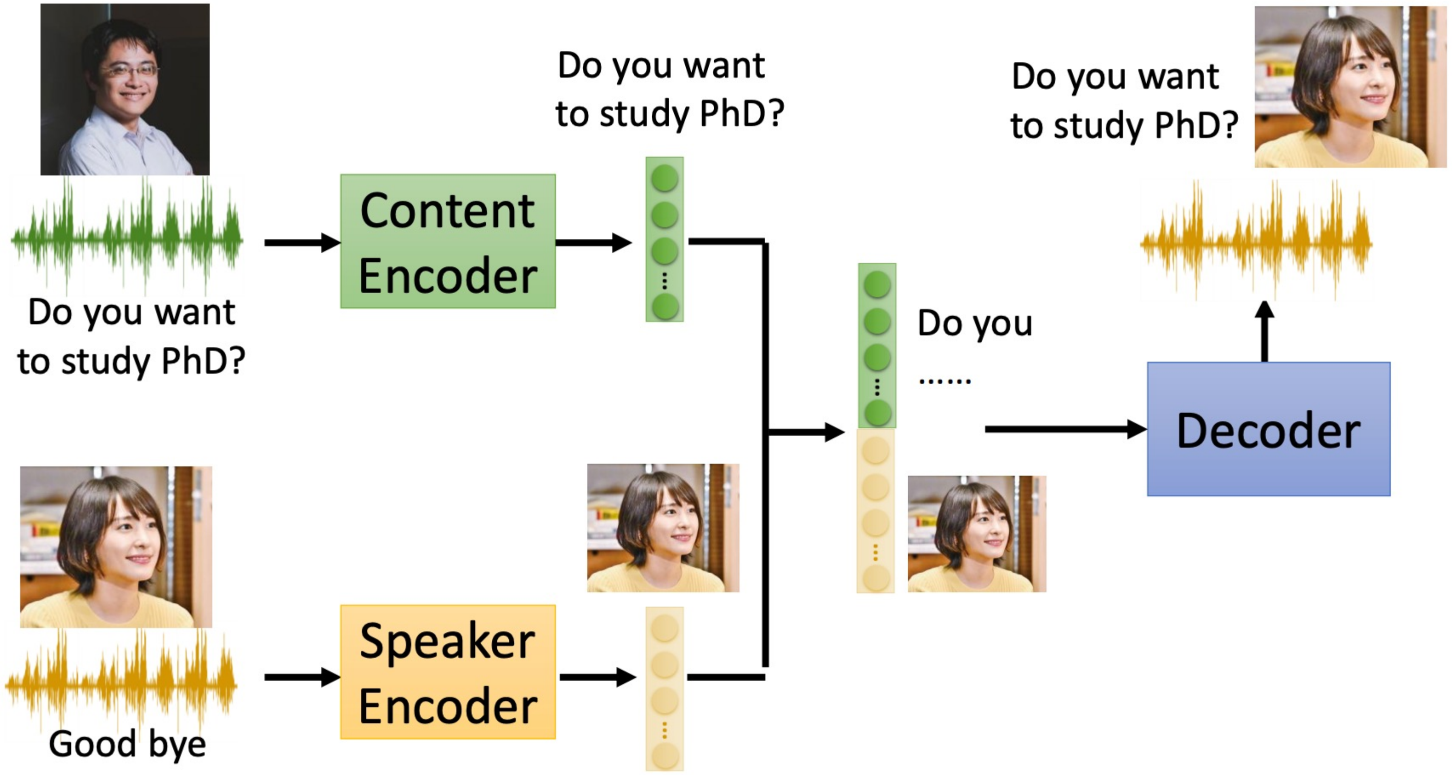
- Well-solved
- Perceptual clean speech

Voice Conversion

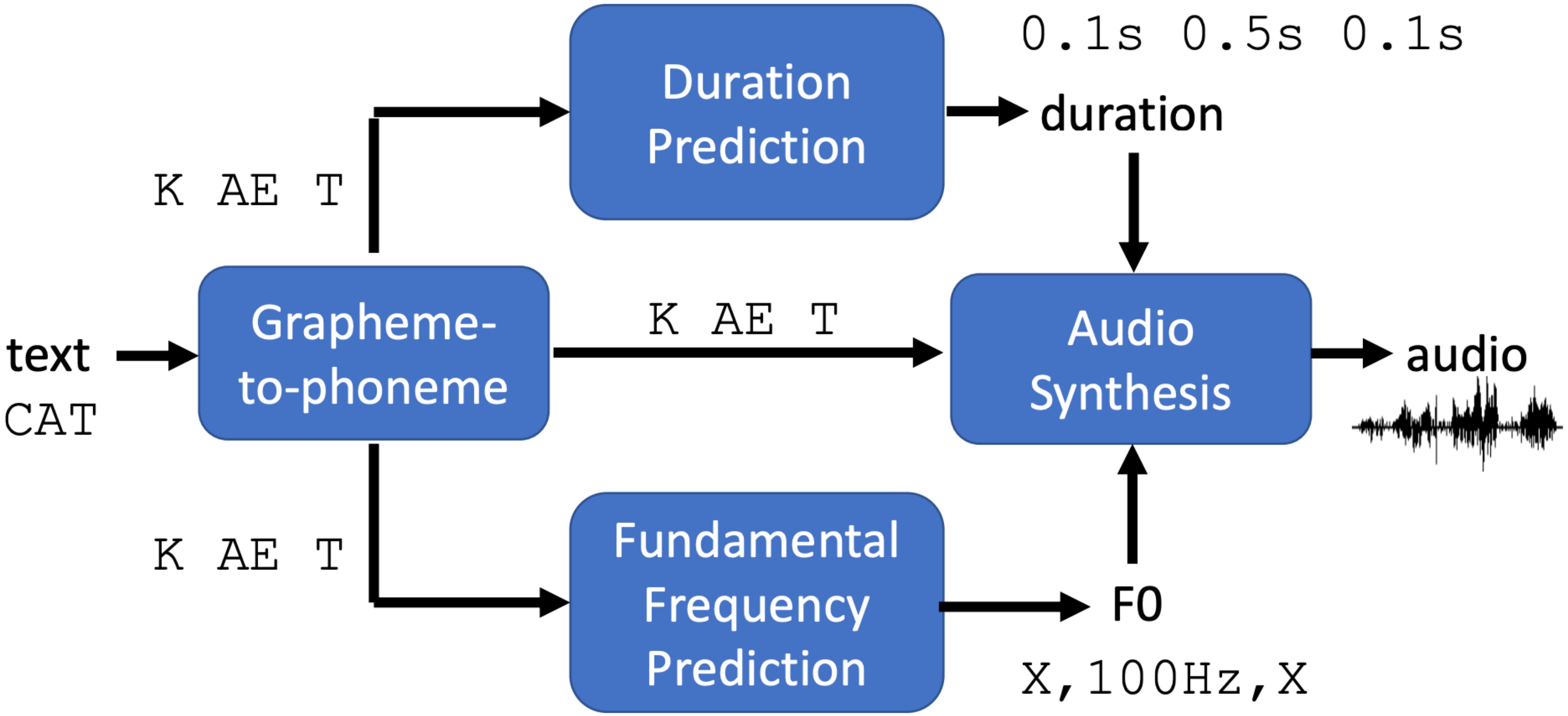
- Feature Disentangle



- Inference pipeline

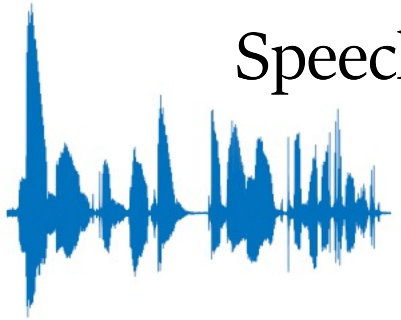


Text-to-speech



- Natural
- Lack of evaluation metric

Speech classification



class

Gender classification

- **Energy Entropy** - Male low and distributed
  Female high and stays for short period of time

$$P(k) = \frac{|X(k)|^2}{\sum_{k=0}^{K} |X(k)|^2}, \qquad H = \sum_{k=0}^{K/2} P(k)\log(P(k)); \qquad M = (E - C_E)(H - C_H), \quad EE = \sqrt{(1 + |M|)}$$

- **Short time energy** – Male low , Female High

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n} - m])^2 = \sum_{m=-\infty}^{\infty} x^2[m]w^2[\hat{n} - m].$$

- **Zero –crossing rate** – Female ZCR higher than male

$$\text{ZCR}, Z = \frac{1}{N}\sum_{i=1}^{N-1} \frac{\operatorname{sgn}\{x(i)\} - \operatorname{sgn}\{x(i-1)\}}{2} \qquad \operatorname{sgn}\{x(i)\} = \begin{cases} 1; x(i) > 0 \\ 0; x(i) = 0 \\ -1; x(i) < 0 \end{cases}$$

- **Spectral Centroid**

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

- **Frame based teager energy**

$$f_i = w_i^2 X(w_i). \qquad T_i = (\sum_{k=1}^{K} f_k)^{1/2}.$$

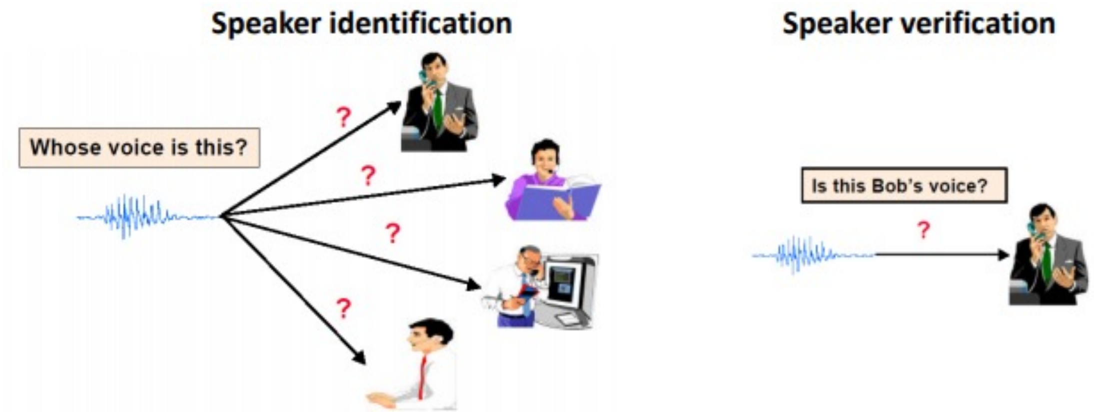
- **Position of Maximum FFT coefficient**

  Position of Maximum FFT coefficient divided by sampling frequency

Emotion recognition

- Categorized emotion, sometimes confusing
- Dataset, actor performance
- Continuous change of emotion

Speaker recognition



- Speaker Verification:

  Supervised binary classification: Given a speech sequence and a claimed identity, accept or reject the identity.

- Speaker Identification:

  Supervised multi-class classification: Determine which speaker (from a predetermined set of speakers) has uttered the sequence.

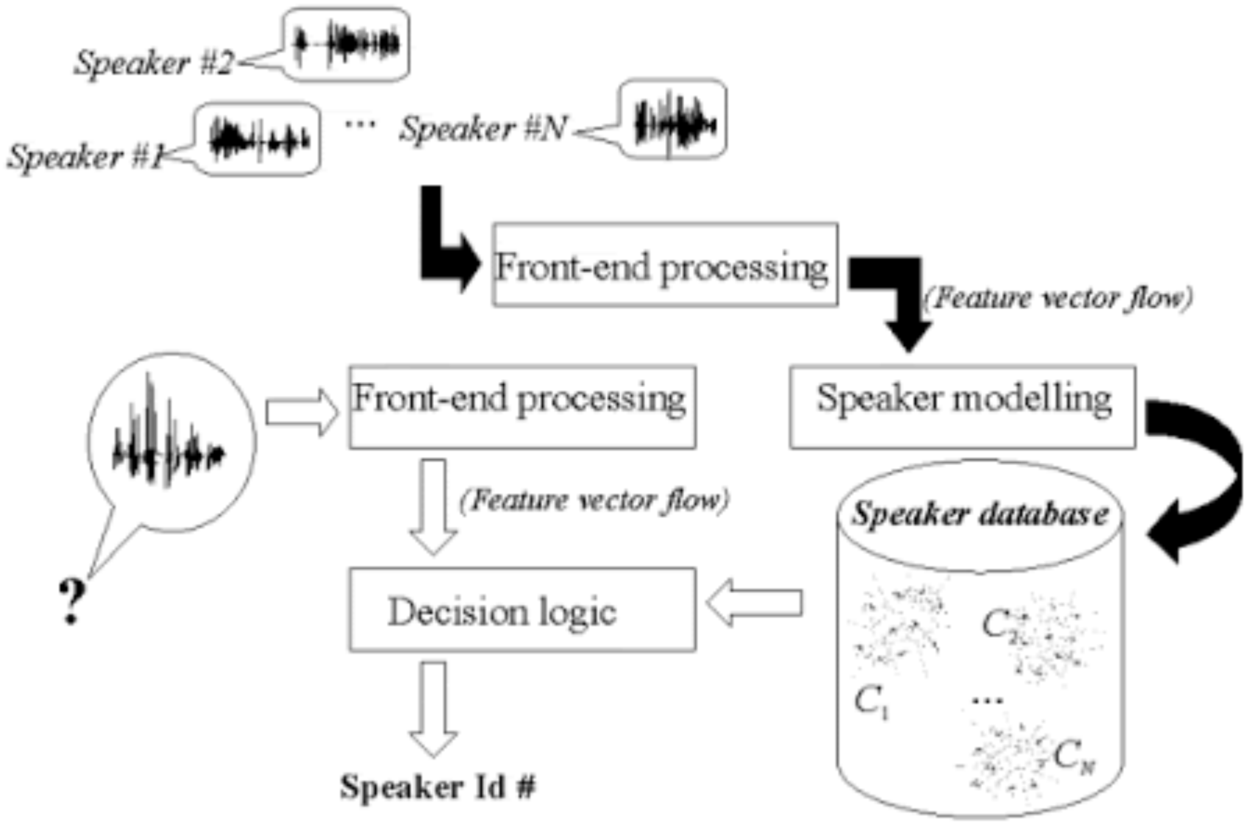
- Speaker Diarization:

  Clustering and segmentation: Partition an input audio stream into homogeneous segments. according to the

Speaker recognition

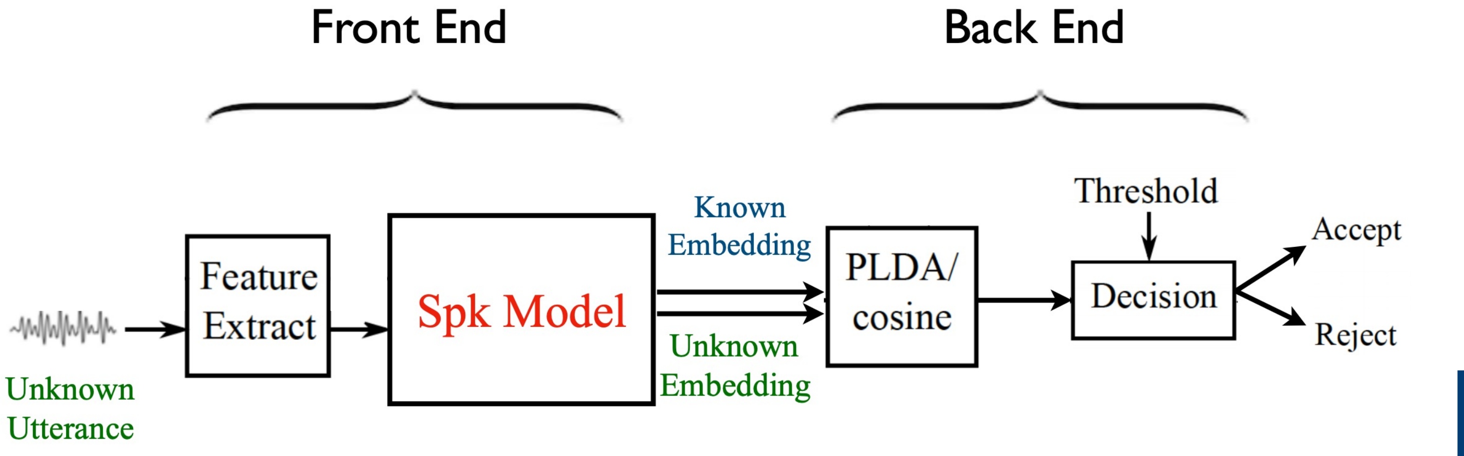
Speaker Embedding:

- Represent speaker info
- Measure the similarity

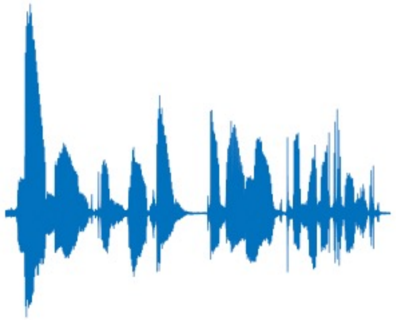


Speaker verification

- Verify the identity of a speaker



Voice Anti-spoofing



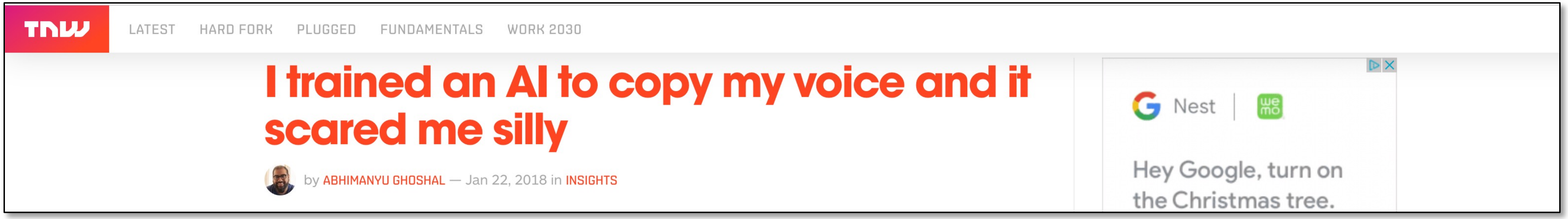
Decision
Genuine or Spoofing attacks


Synced

AI TECHNOLOGY

Clone a Voice in Five Seconds With This AI Toolbox

A new Github project introduces a remarkable Real-Time Voice Cloning Toolbox that enables anyone to clone a voice from as little as five seconds of sample audio.


TNW LATEST HARD FORK PLUGGED FUNDAMENTALS WORK 2030

I trained an AI to copy my voice and it scared me silly

by ABHIMANYU GHOSHAL — Jan 22, 2018 in INSIGHTS


THE WALL STREET JOURNAL.

Subscribe | Sign In

Special Offer

English Edition | Print Edition | Video | Podcasts | Latest Headlines

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine Search

SHARE

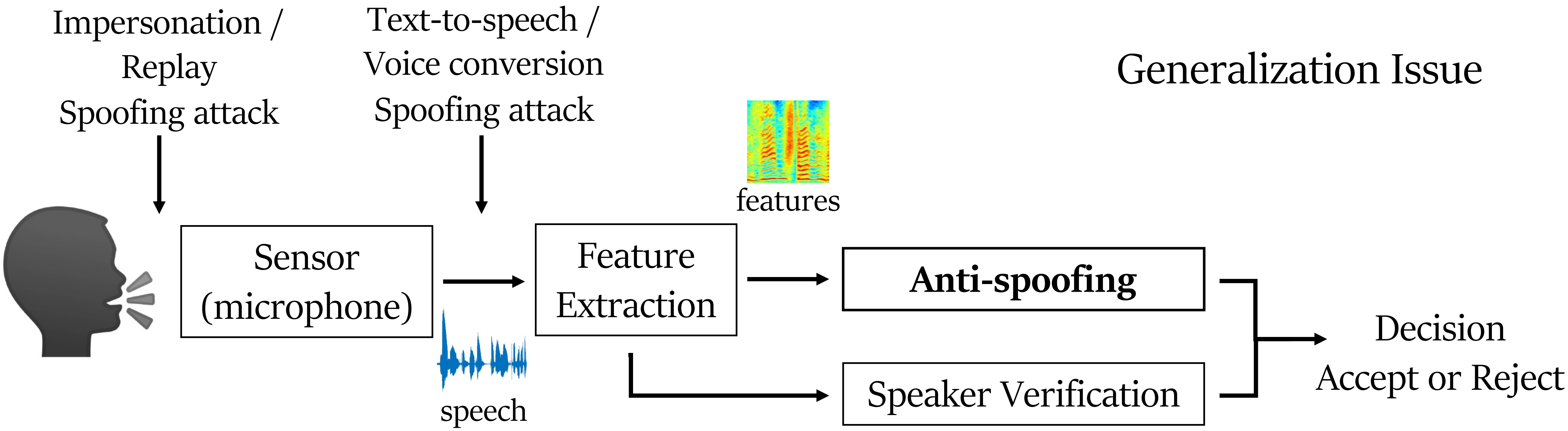
PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

Voice Anti-spoofing

- Detect spoofing attacks (fake speech)



https://vimeo.com/345075279

Anti-Spoofing Demo from ID R&D

Keyword spotting

Overview

Keyword selection
- *Not a common word*
- *Easy to pronounce*

↓

Record samples
- *Usually >200 talkers, 2000 samples*

↓

Data preparation
- *Split into training/test set*
- *Label the utterance onset/offset*

↓

Model training
- *GMM-HMM*

↓

Test & Parameter tuning
- *Trade-off between detection accuracy and false alarm (precision and recall)*

↓

Release
- *Embedded into device*

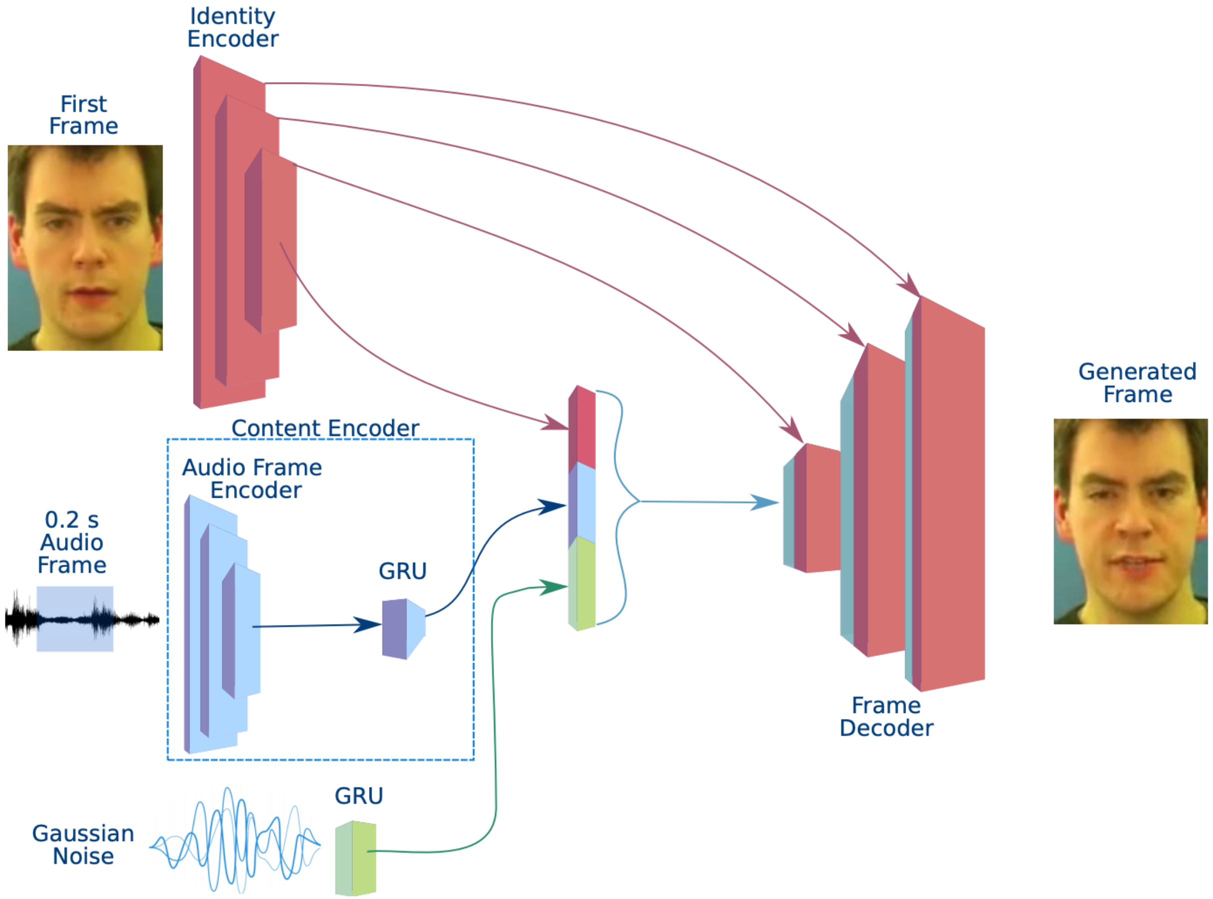
"Alexa"

"Ok Google"

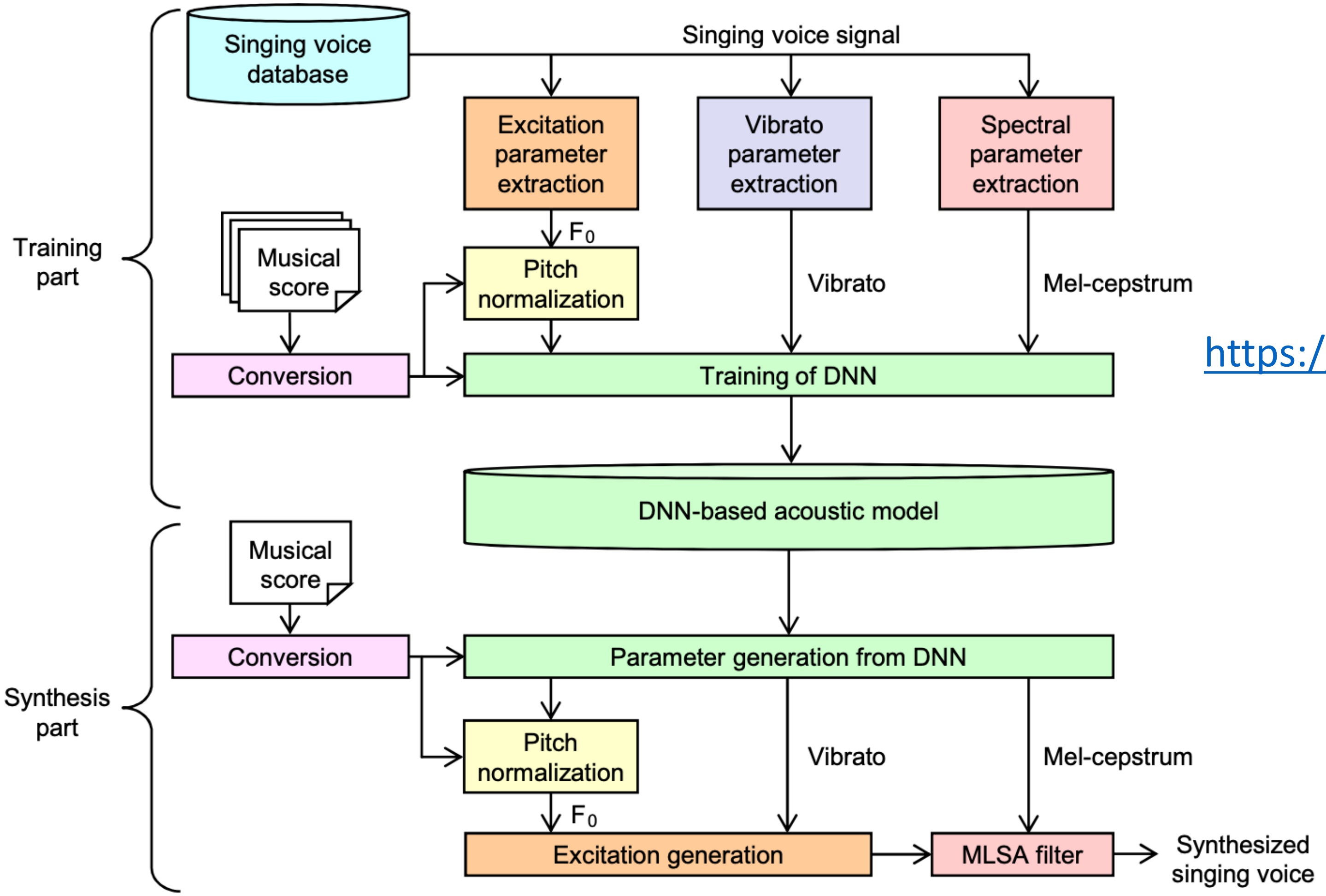
"Hey Siri"

Other topics

Talking face generation



Singing voice synthesis



https://bytesings.github.io/paper1.html

Future horizons

- General speech understanding
- Disentangled speech representation
- Human-Computer Interaction with speech

Thank you !

Q & A

Speech Features

- Resonance peak
- Features (What aspects does each one models?):
- PLP
- MFCC
- PNCC

Audio Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC)

Steps

1. Audio frame → FFT → Spectrum
2. Spectrum → Mel-Filters → Log-Mel Spectrum
3. Perform cepstral analysis
4. Take the first multiple cepstral coefficients as MFCCs