Speech Technology

Neil Zhang

ECE 277/477 - Computer Audition, Fall 2023

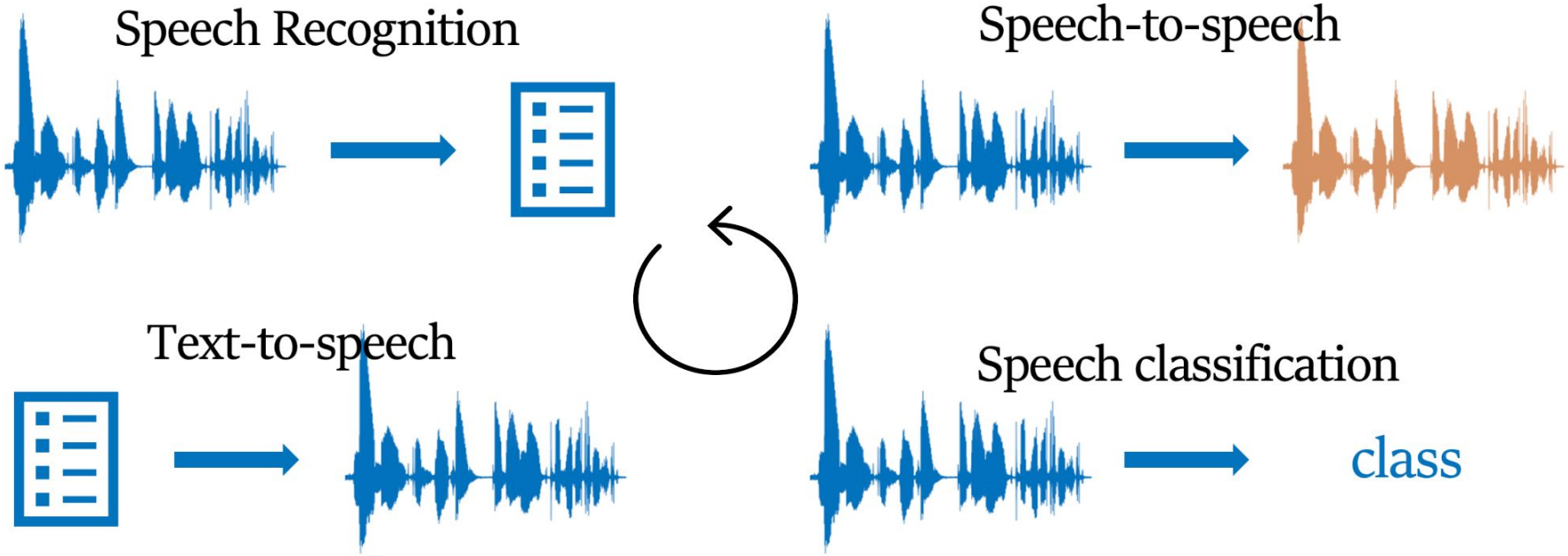
Outline

Overview of research topics in speech technology

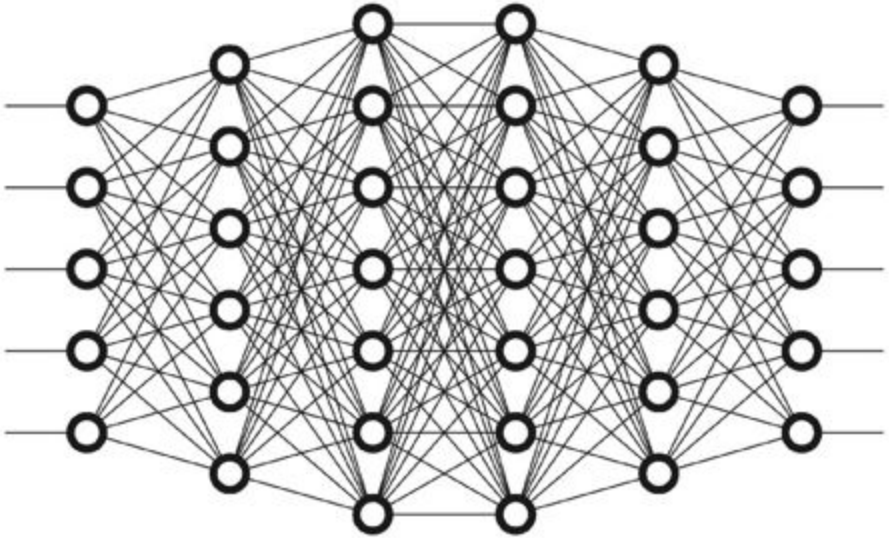
Common front-end for various tasks of speech processing

Speaker verification and speaker diarization for HW6

Overview of Speech Topics

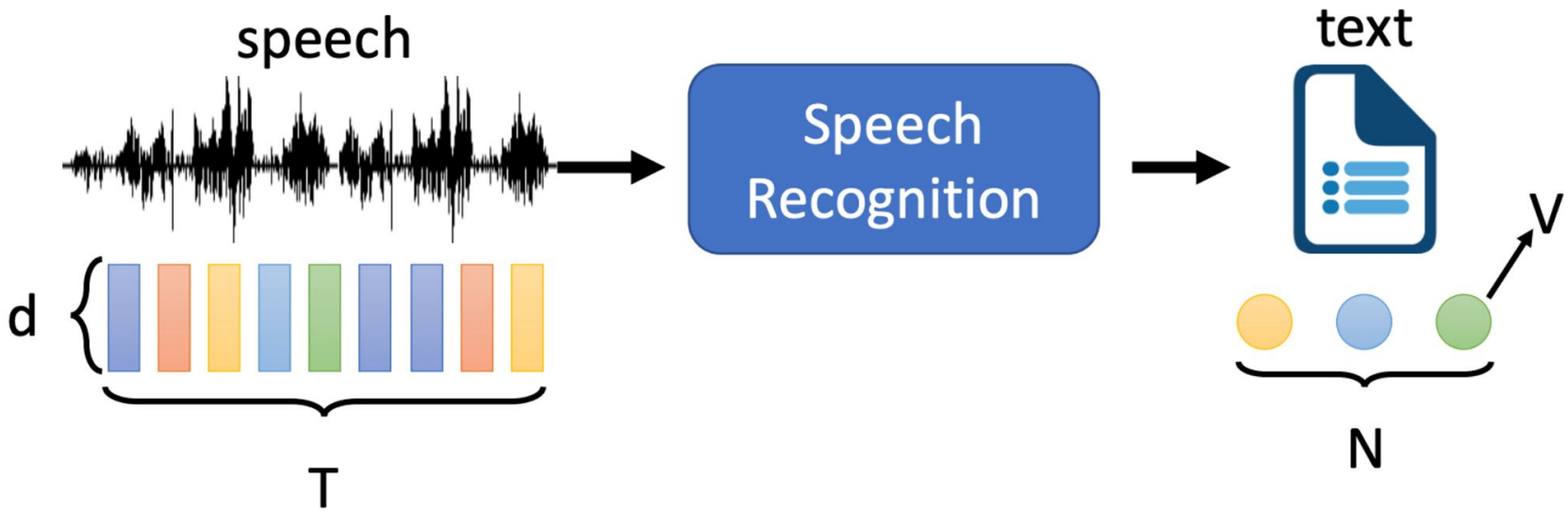


Beyond Training DNNs



What are the additional concerns of each research topic beyond the training of Deep Neural Networks?

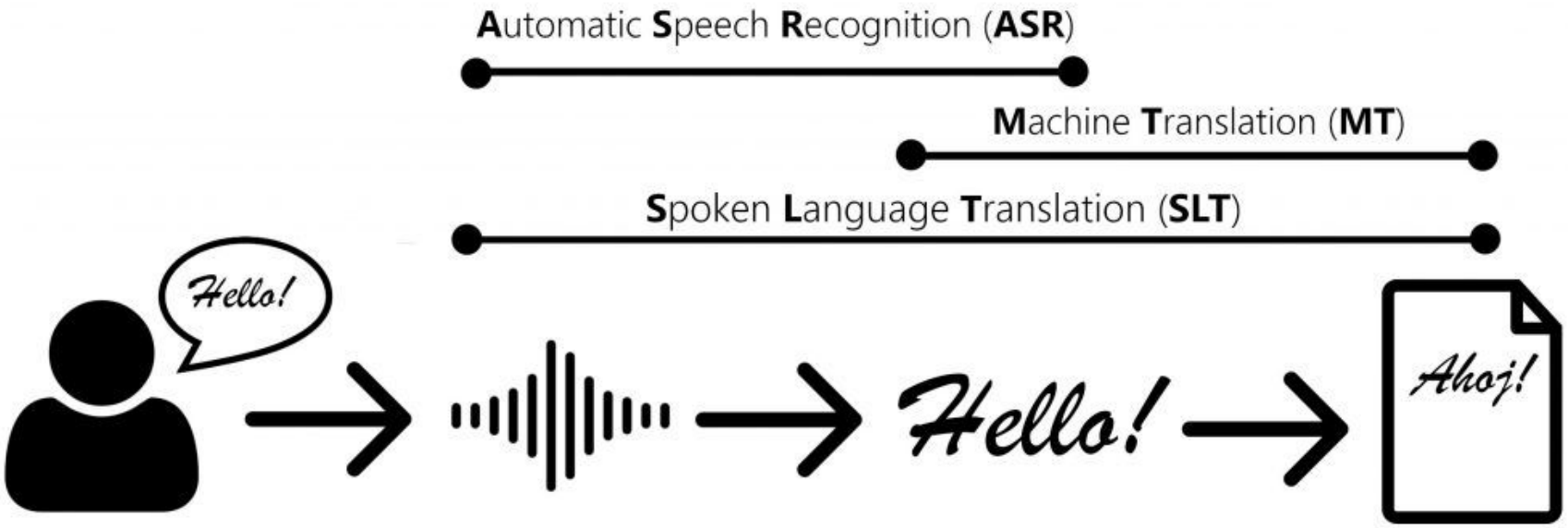
Speech Recognition



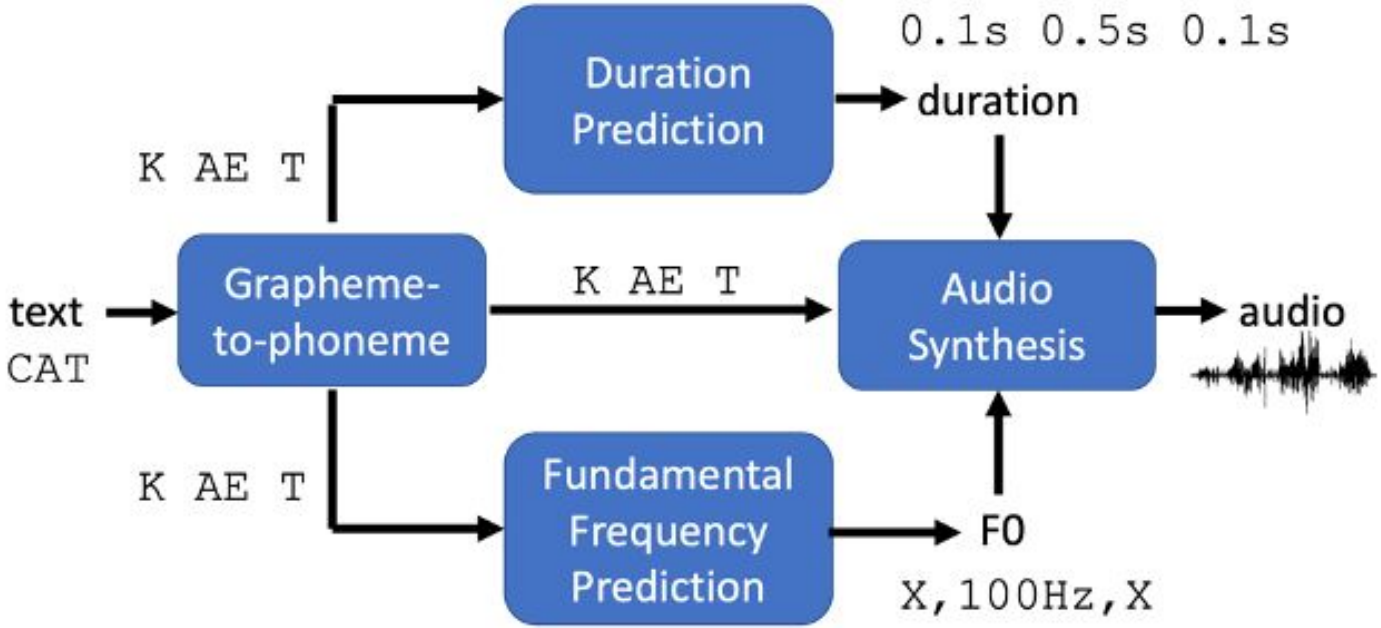
Speech: a sequence of vector (length T, dimension d)

Text: a sequence of token (length N, V different tokens)

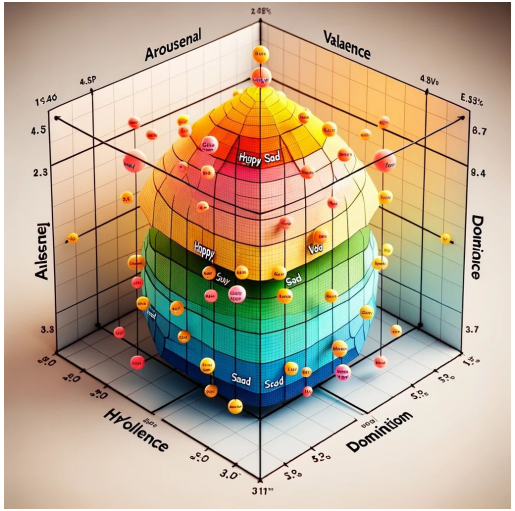
Speech Translation



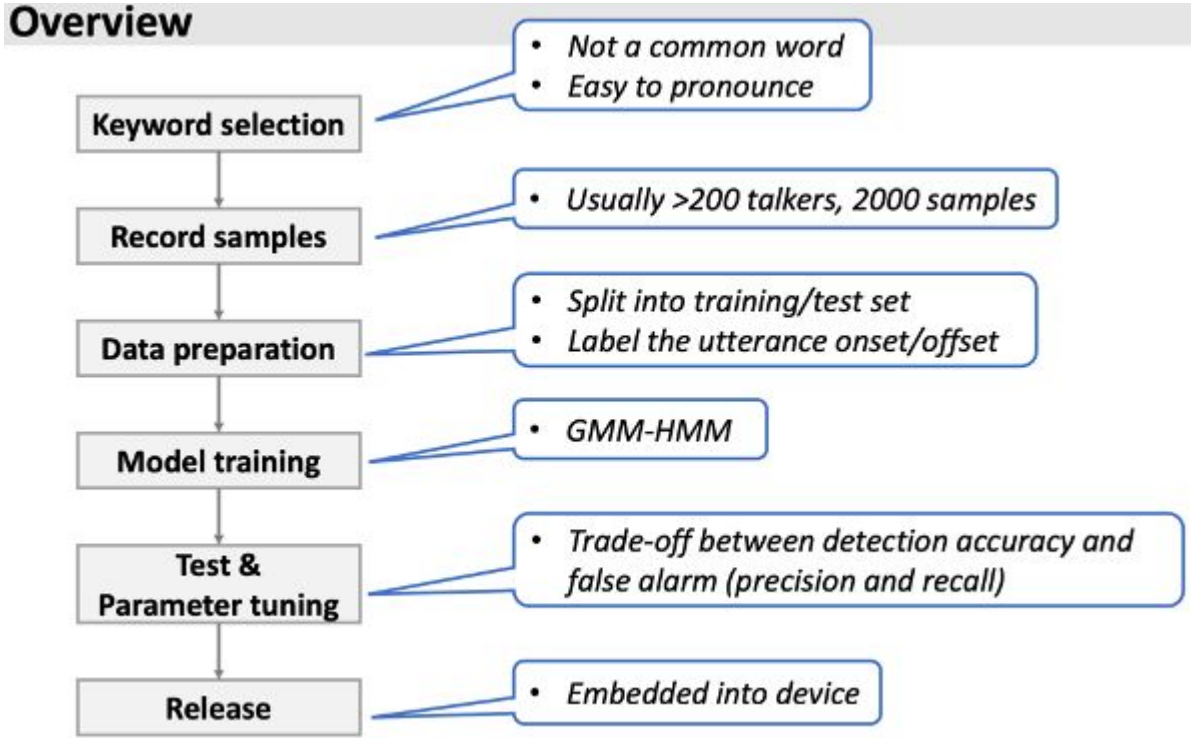
Text-to-speech



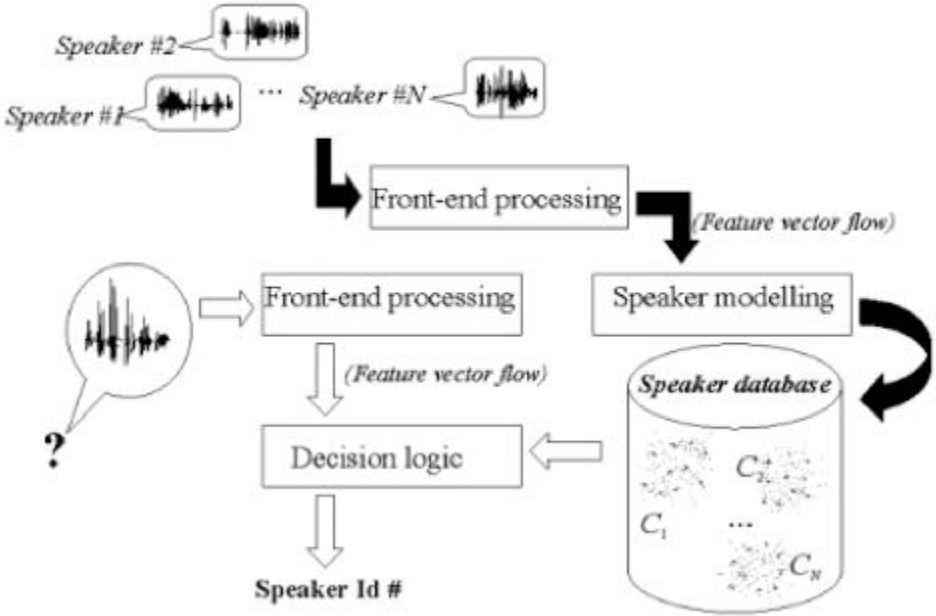
Speech Emotion Recognition



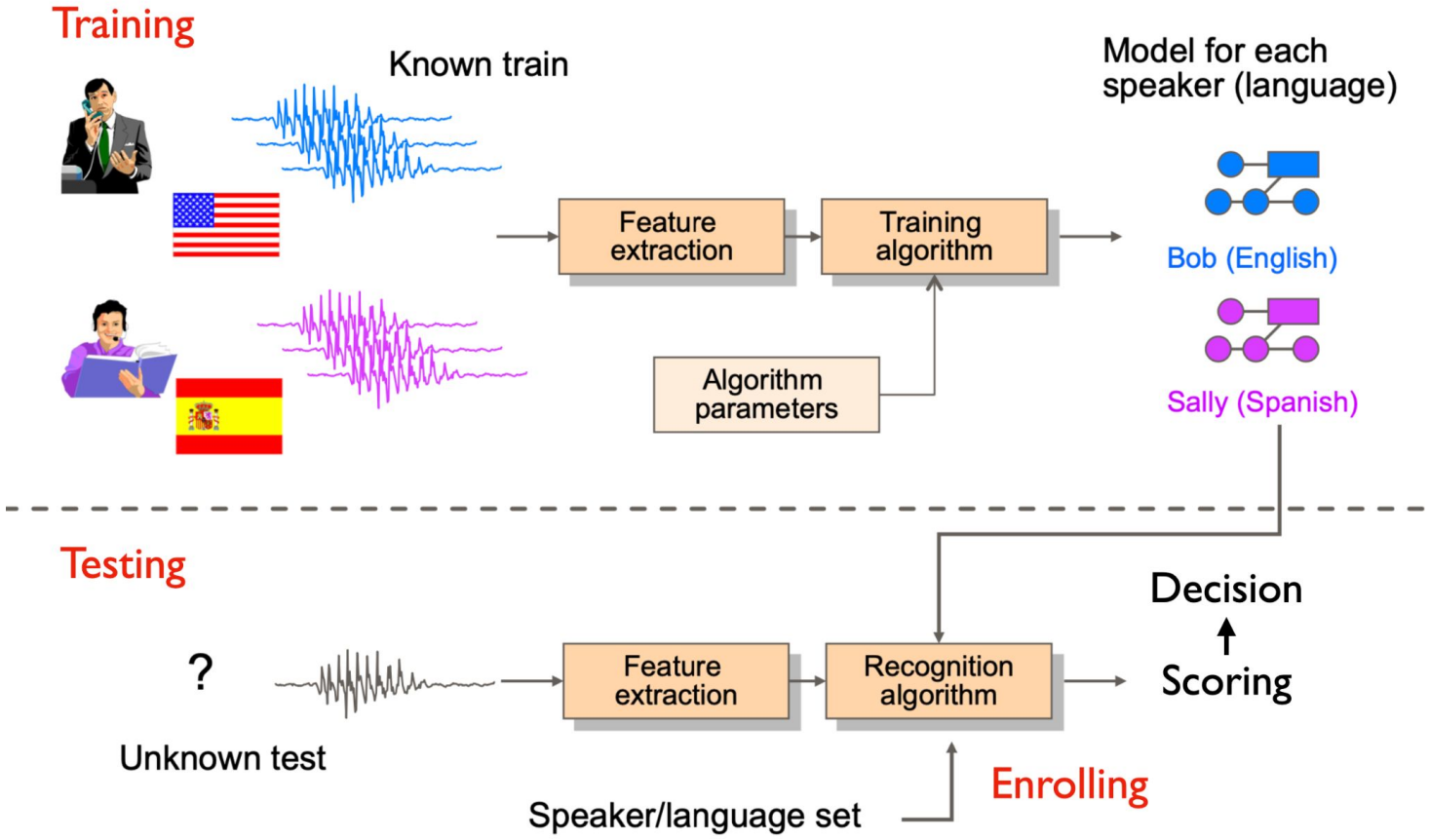
Keyword Spotting



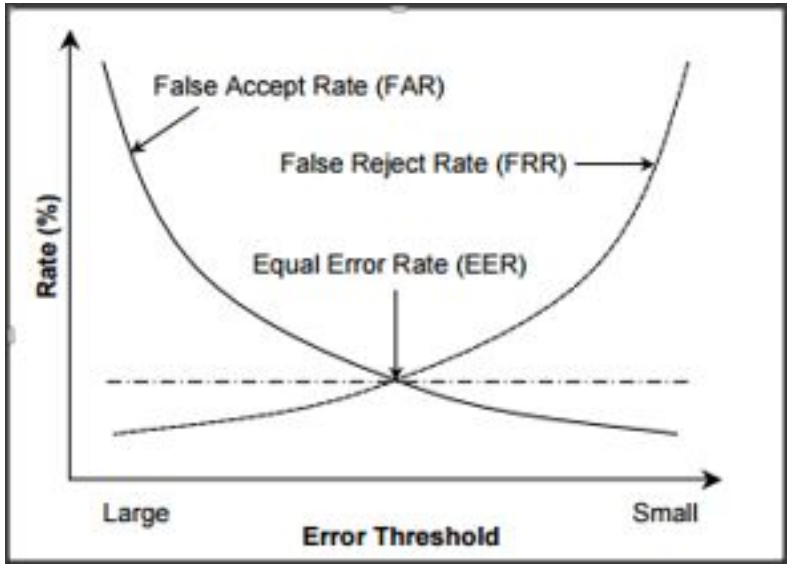
Speaker Recognition



Speaker Verification

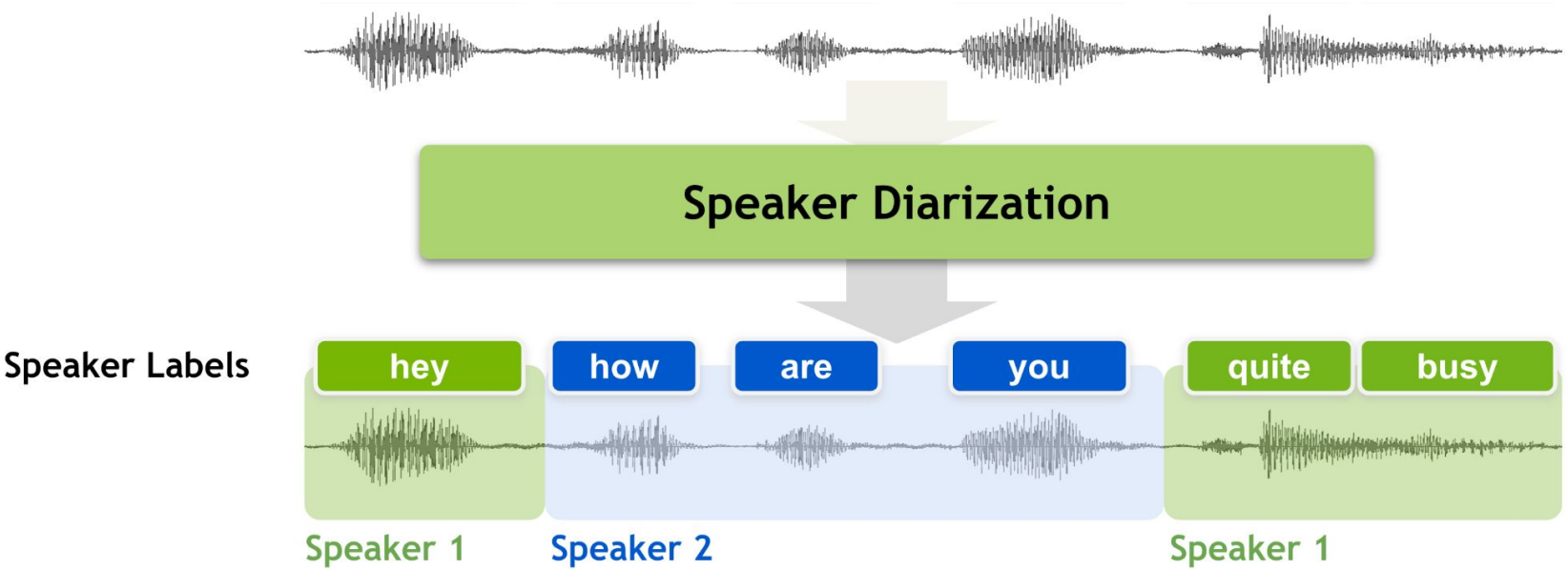


Equal Error Rate (EER)

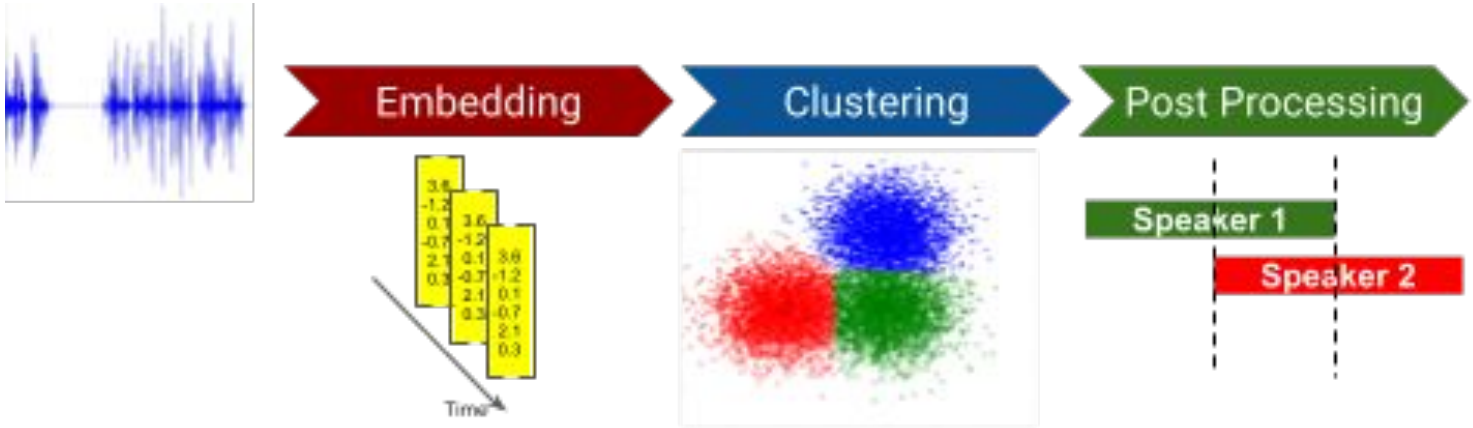


Speaker Diarization

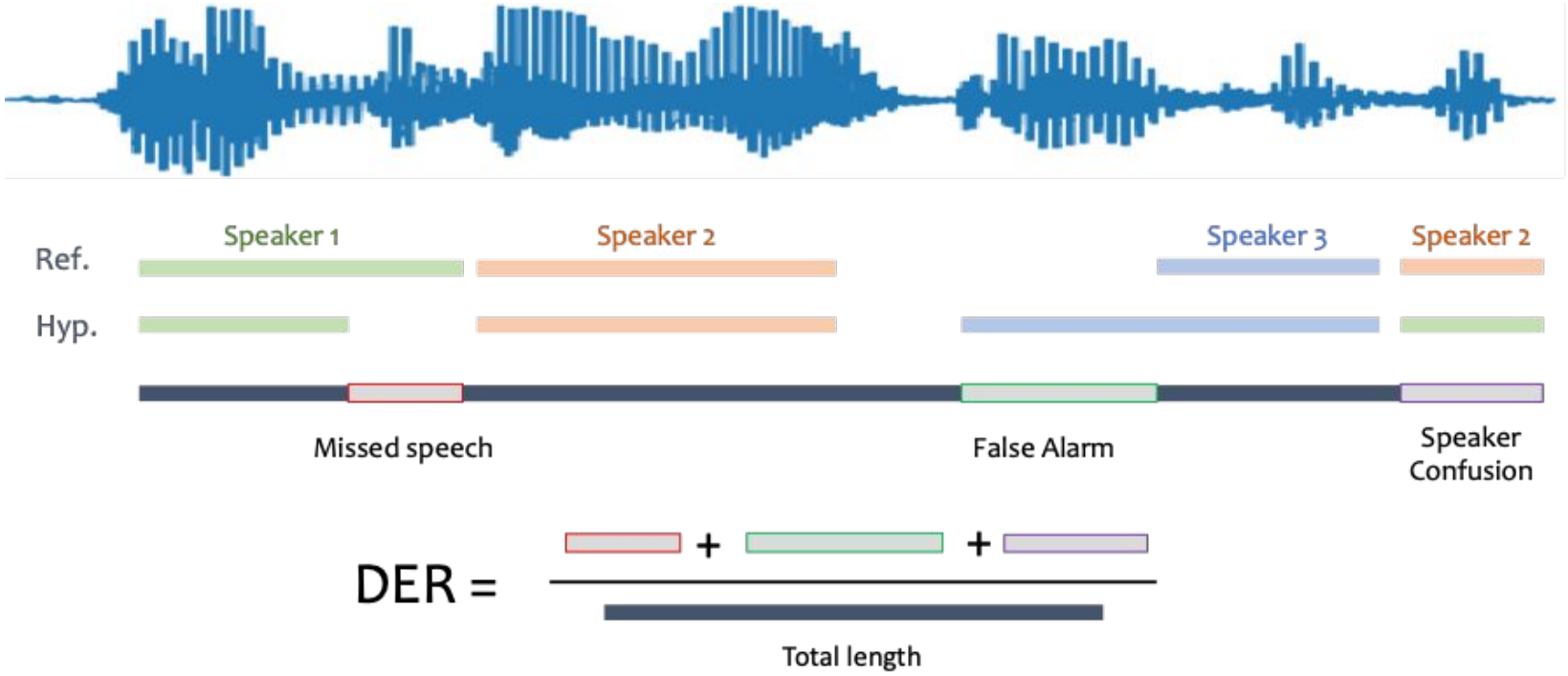
Who spoke when



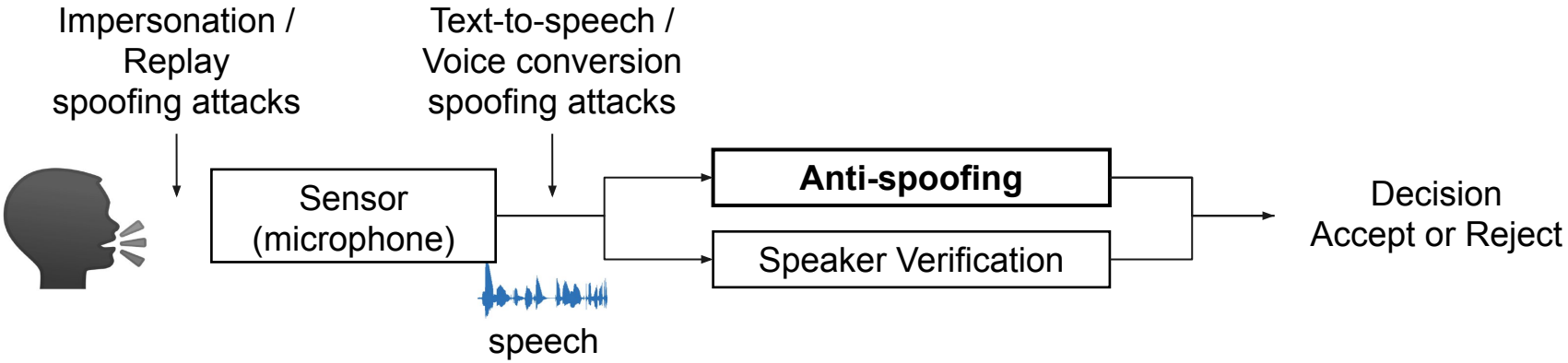
Speaker Diarization



Diarization Error Rate (DER)



Speech Anti-Spoofing



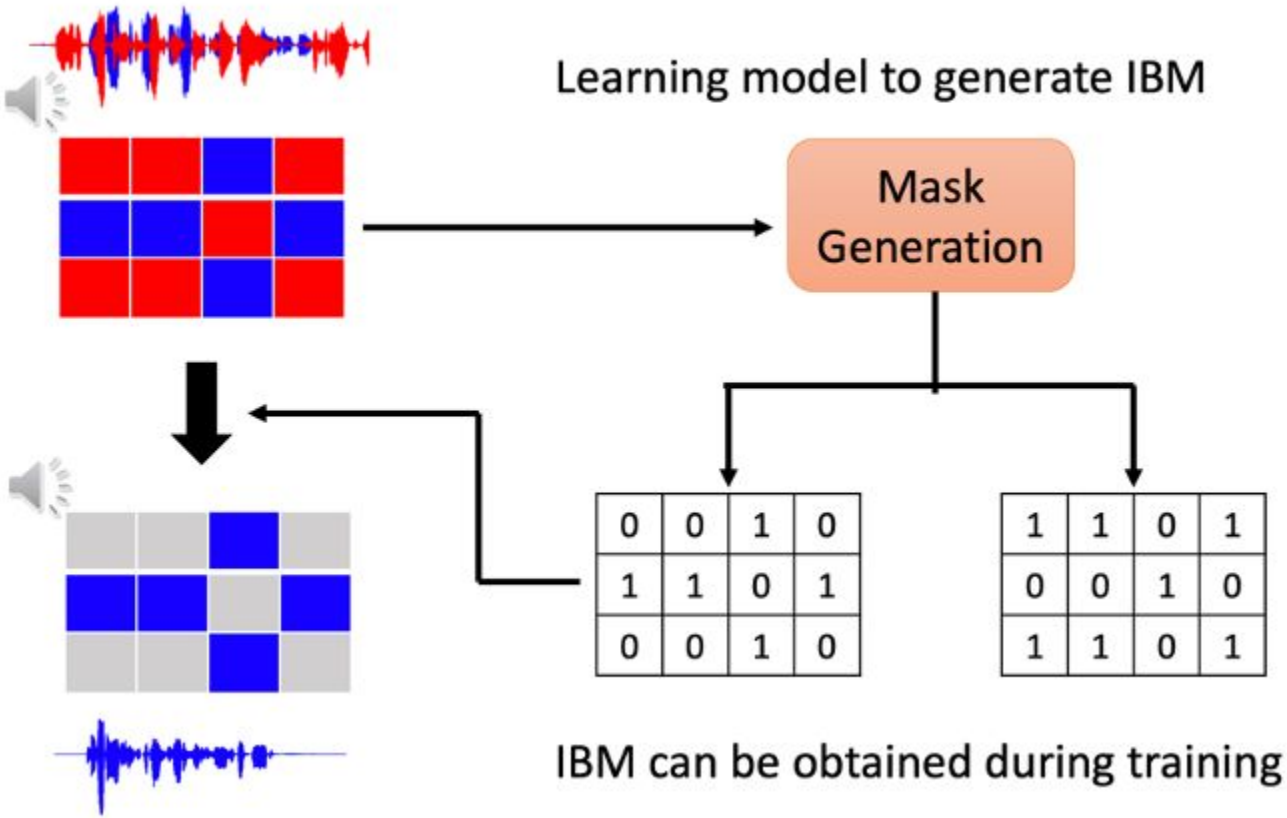
Speech Enhancement

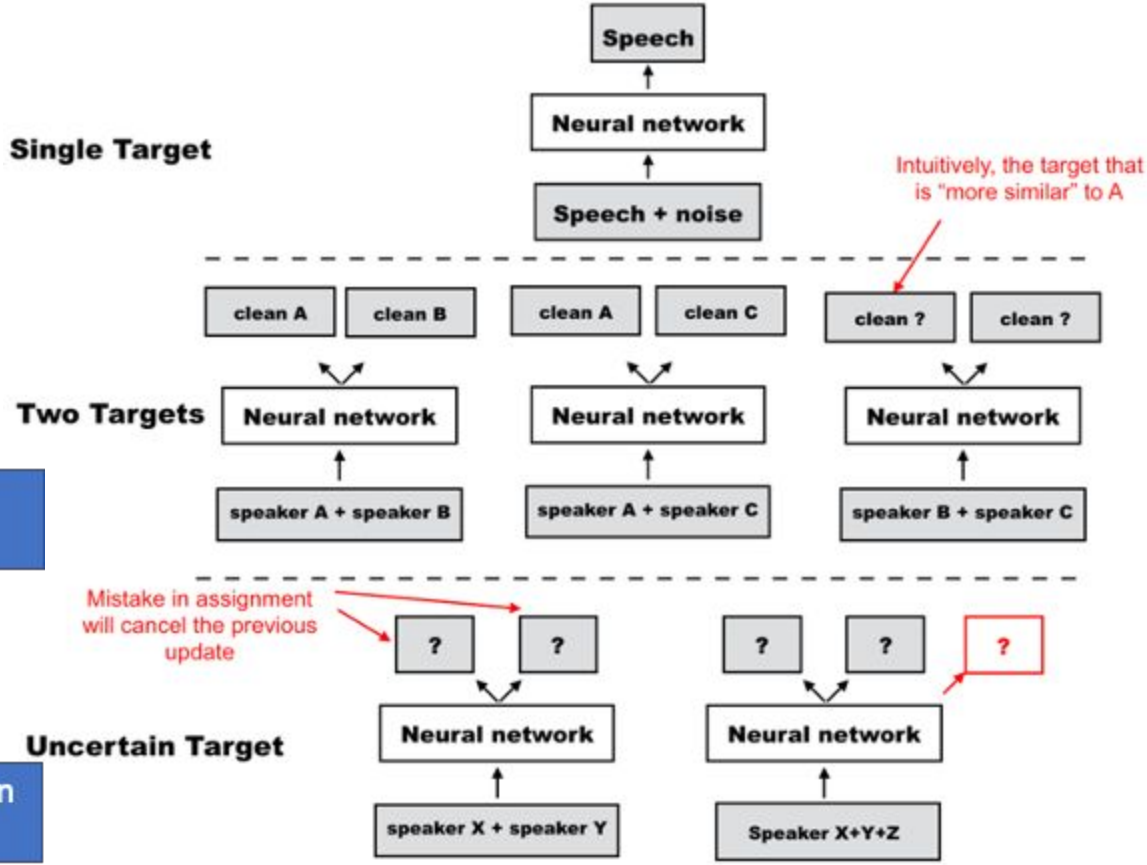
Noisy Speech Clean Speech Noise

Speech Separation

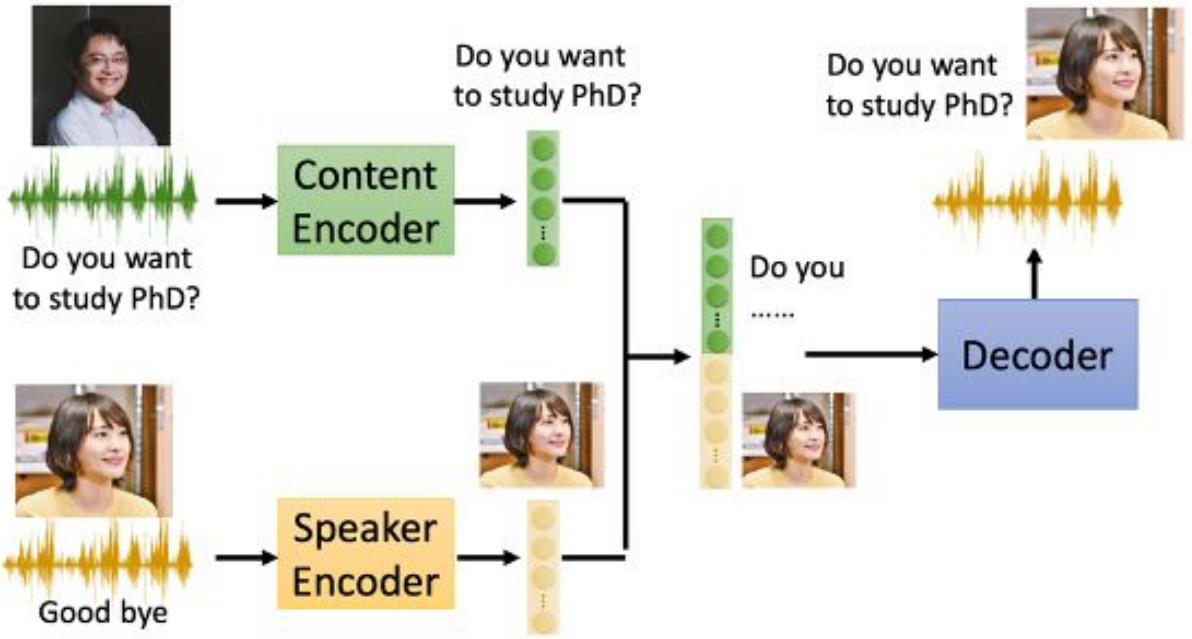
https://researcher.watson.ibm.com/researcher/view_group.php?id=2819

Ideal Binary Mask

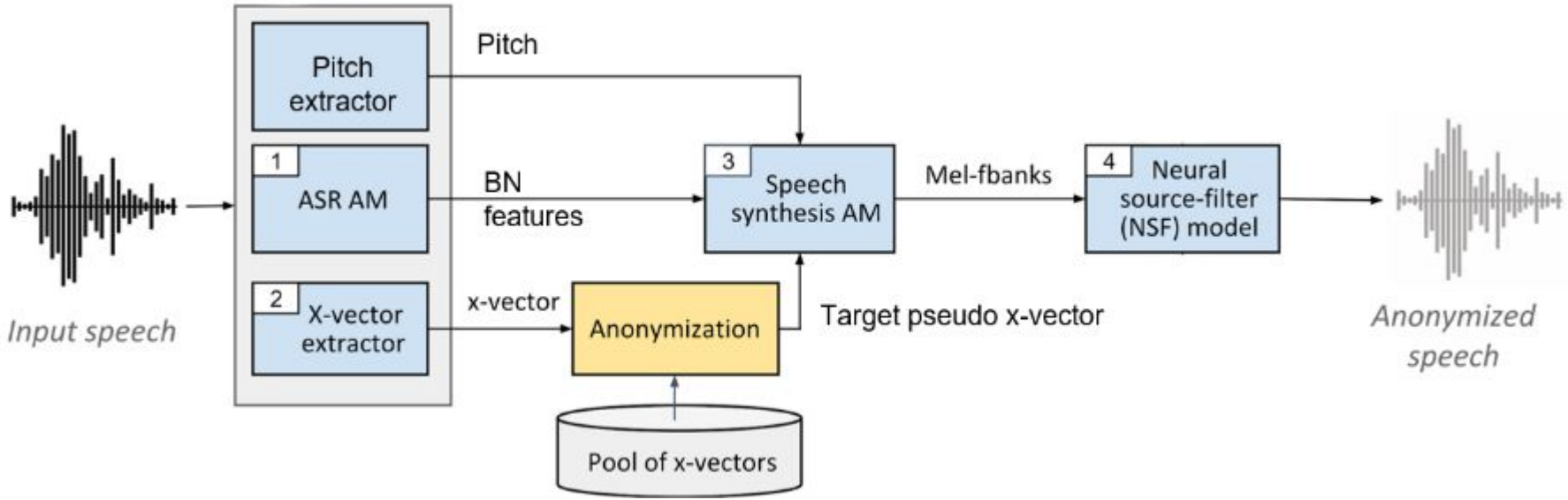



Speech
Neural network
Speech + noise
Single Target
Intuitively, the target that is "more similar" to A
clean A
clean B
clean A
clean C
clean ?
clean ?
Two Targets
Neural network
Neural network
Neural network
speaker A + speaker B
speaker A + speaker C
speaker B + speaker C
Permutation problem
Mistake in assignment will cancel the previous update
?
?
?
?
?
Uncertain Target
Neural network
Neural network
speaker X + speaker Y
Speaker X+Y+Z
Output dimension mismatch

Voice Conversion

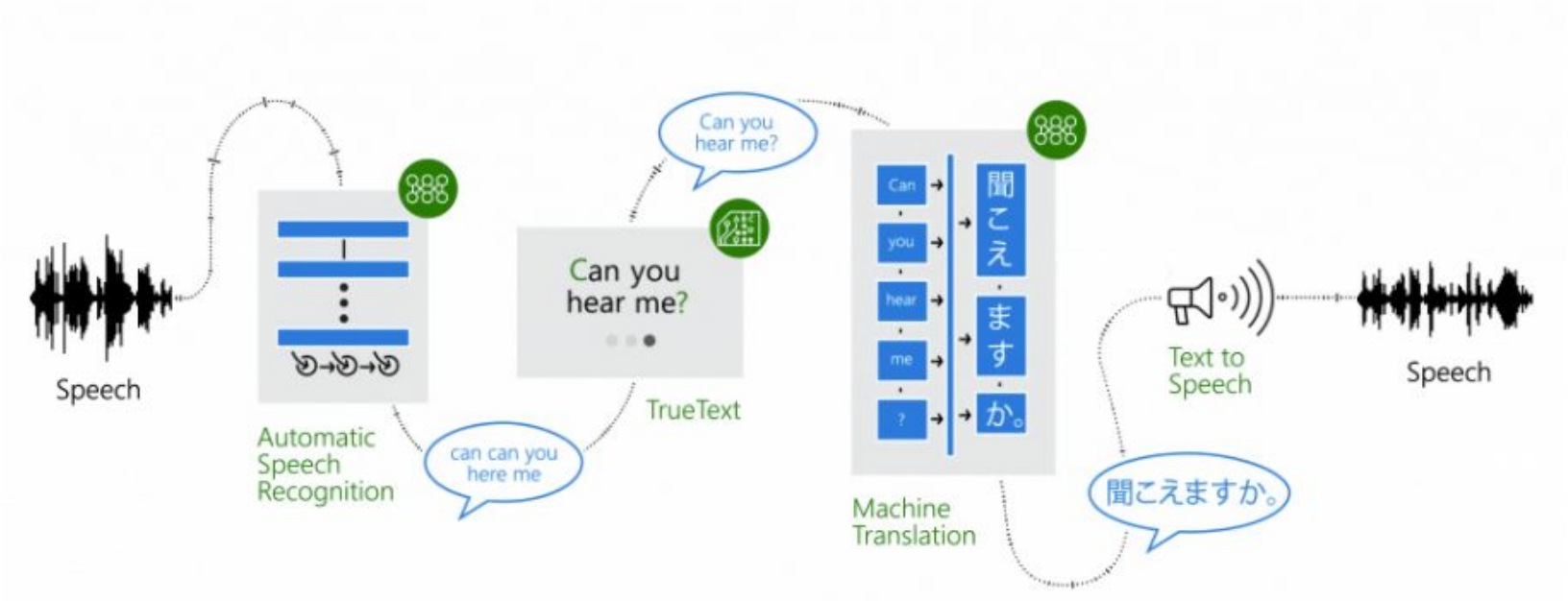


Speech Anonymization

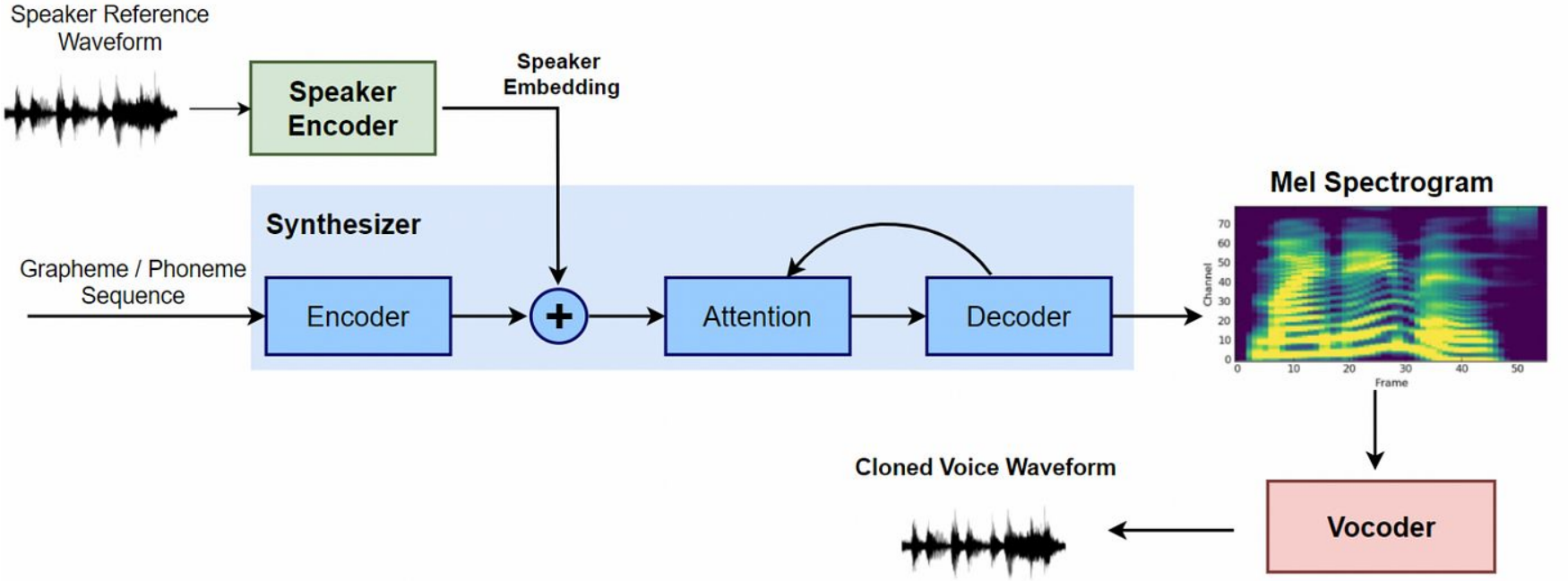


Speech-to-Speech Translation

https://about.fb.com/news/2022/10/hokkien-ai-speech-translation/



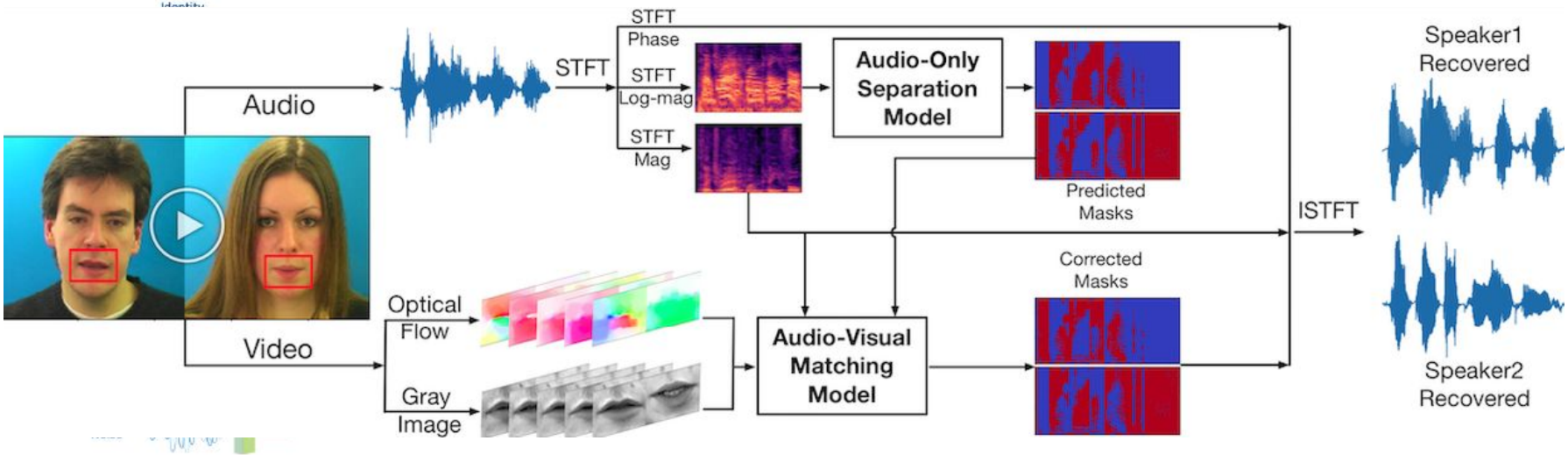
Voice Cloning



Other topics

Beyond speech: extend to singing voice https://bytesings.github.io/paper1.html

Cross modality: audio-visual



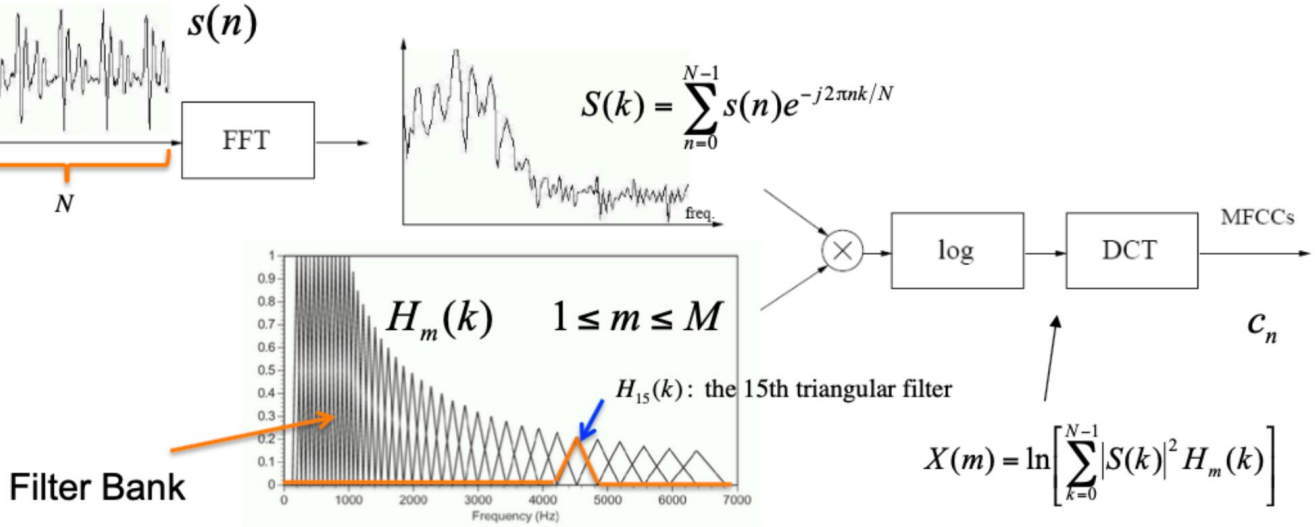
Future horizons

Disentangled speech representation learning

General speech and language understanding (e.g. intonation and intention)

Human-computer interaction with speech

Mel-Frequency Cepstral Coefficients (MFCCs)



$$\text{MFCC}_n = c_n = \sum_{m=1}^{M} \cos\left[n\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right] X(m) = \mathbf{d}_n^T \vec{X}, \qquad 0 \le n \le P$$

$$\Rightarrow \text{MFCC vector} = \begin{bmatrix} c_0 & c_1 & \cdots & c_P \end{bmatrix}^T = \mathbf{D}\vec{X}$$

D : DCT Transformation matrix $[P \times M]$

M : No. of triangular filters in the filter bank, typically $20 \sim 30$

P : No. of cepstral coefficients, typically 12

c_0 : Logarithm of energy of the current frame

Benefits of MFCC

Approximates human hearing

Dimensionality reduction

Good at distinguishing between different phonemes

Directly Learning from Raw Waveforms

STFT: temporal and frequency resolution tradeoff

CNN: Temporal resolution – stride size;

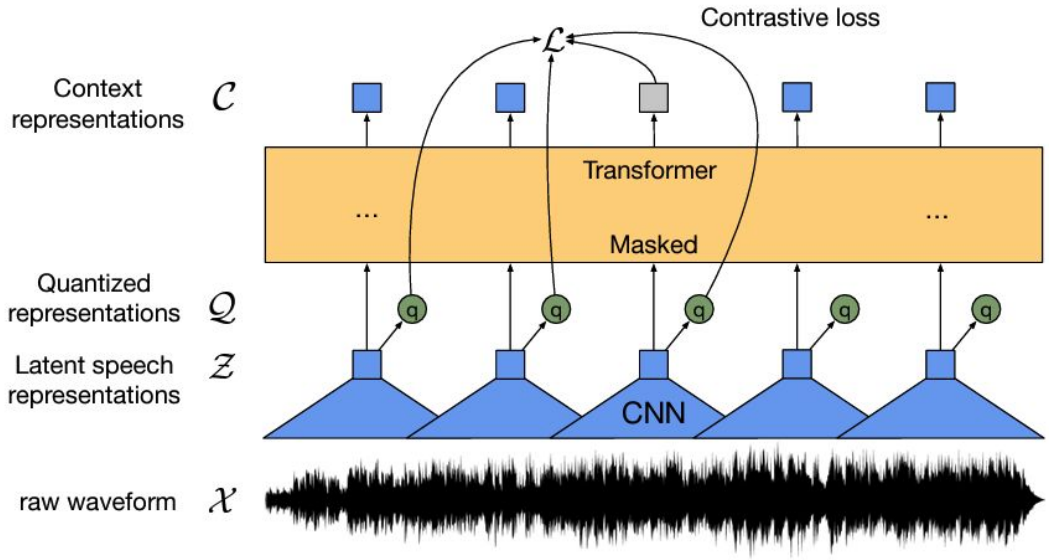
Frequency resolution – number of channels

Frequency component – kernel size

Phase information is kept in raw waveform.

Refer to SincNet, RawNet if interested.

Self-supervised Learning Features



Refer to wav2vec2, HuBERT, WavLM if interested.

1D convolution