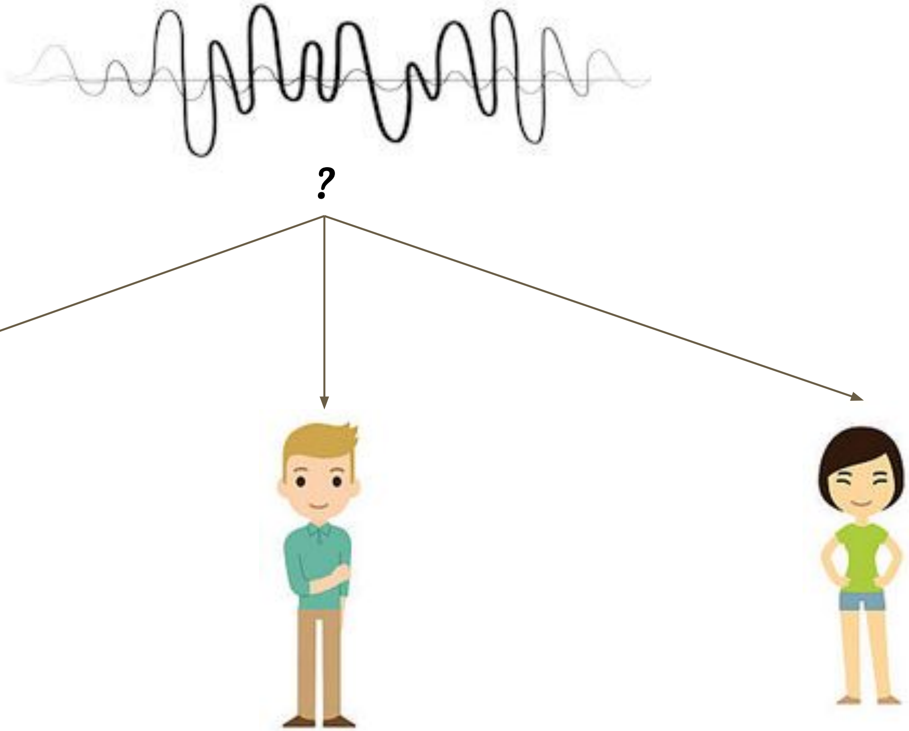
Project Presentation:

Speaker Identification

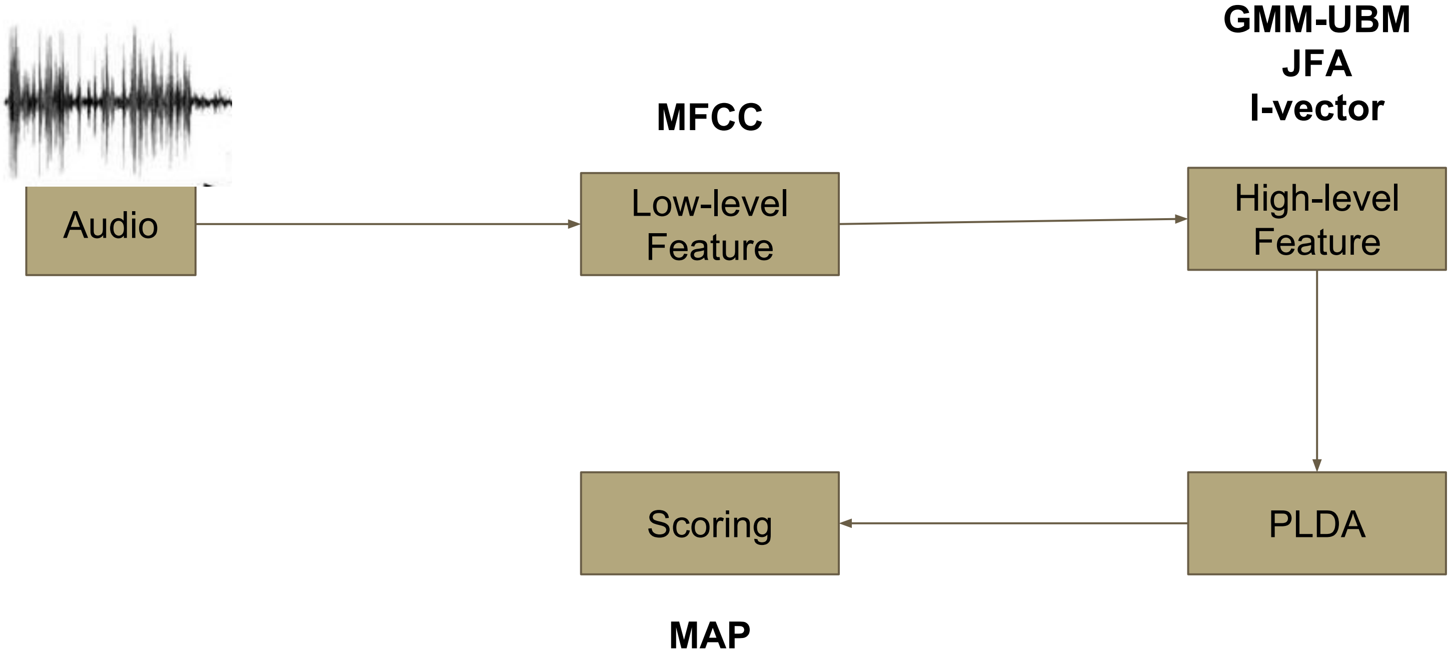
Le Ngoc Tuan Khang

What is Speaker Identification?

Whose voice is this?



Overall Pipeline



Audio

MFCC

Low-level Feature

GMM-UBM
JFA
I-vector

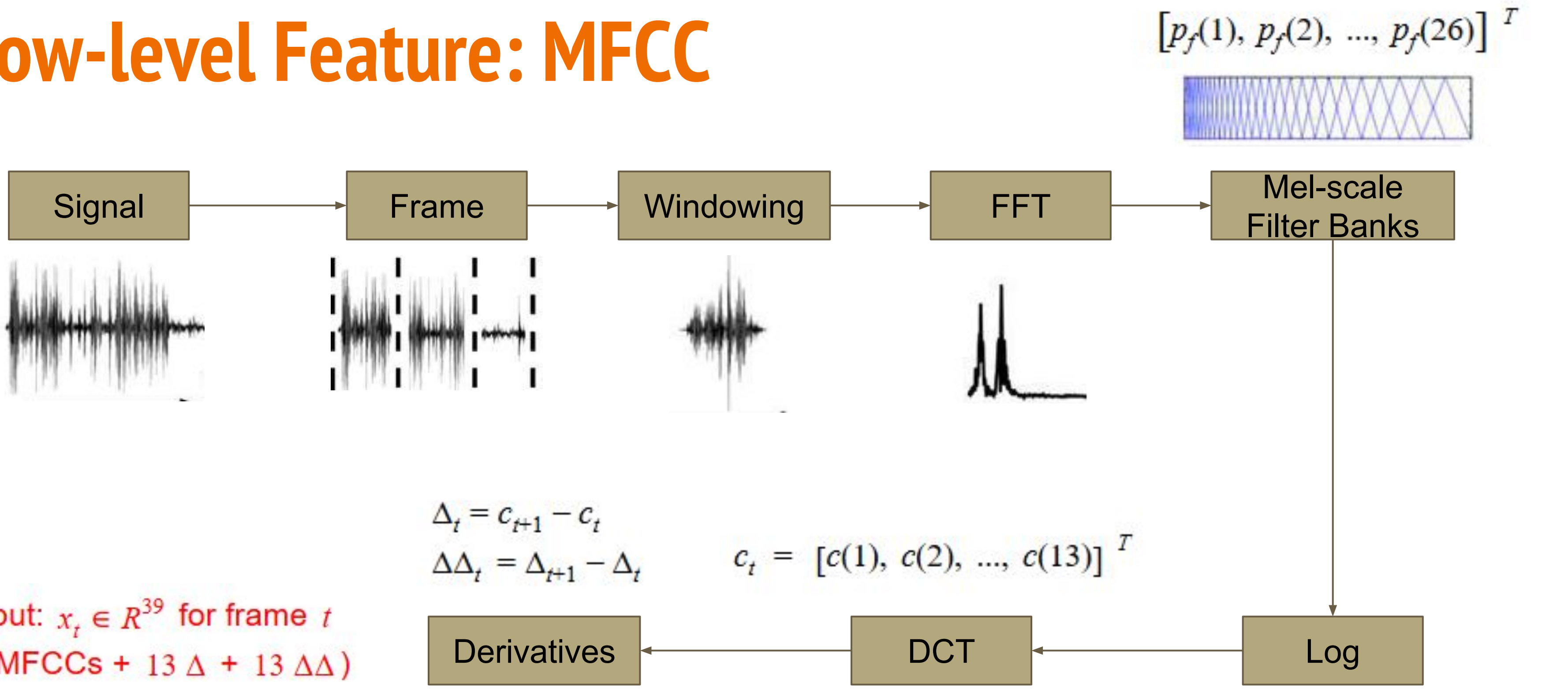
High-level Feature

PLDA

Scoring

MAP

Low-level Feature: MFCC



$$[p_f(1),\ p_f(2),\ \dots,\ p_f(26)]^T$$

$$[\log p_f(1),\ \log p_f(2),\ \dots,\ \log p_f(26)]^T$$

$$c_t = [c(1),\ c(2),\ \dots,\ c(13)]^T$$

$$\Delta_t = c_{t+1} - c_t$$

$$\Delta\Delta_t = \Delta_{t+1} - \Delta_t$$

Output: $x_t \in R^{39}$ for frame t

(13 MFCCs + 13 Δ + 13 $\Delta\Delta$)

Mel Frequency Cepstral Coefficient

High-level Feature: Gaussian Mixture Models

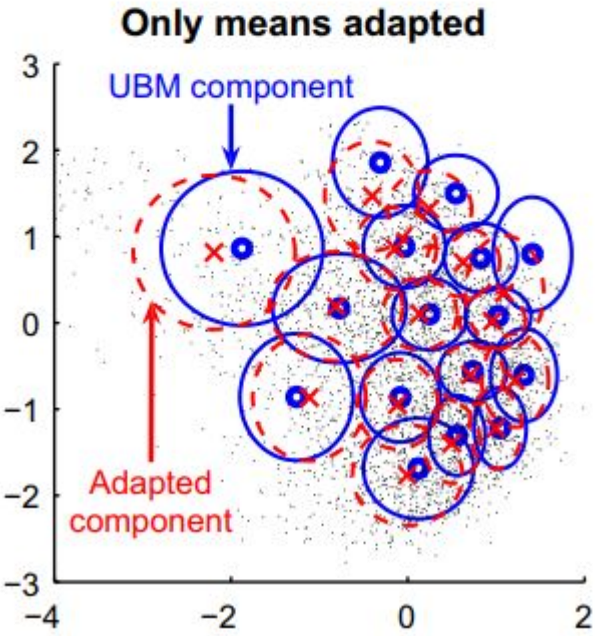
Assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

$$Pr(X) = \sum_{i=1}^{k} \pi_k N(X \mid \mu_k, \Sigma_k)$$

The parameter **ϴ = {π, μ, Σ}** (under which the data is **most likely)** can be learned by **Expectation-Maximization** algorithm.

High-level Feature: UBM-GMM

- UBM: Universal Background Model
- UBM is a GMM which trained on a large dataset.
- For each speaker in the identification set, UBM is **adapted** to represent that speaker.
- Adapted GMM means are stacked into a **supervector**.



High-level Feature: I-vector

- I-vector (GMM supervectors + factor analysis)

$$M = m + T\Phi$$

Giving T, i-vector can be extracted using Baum-Welch statistics.

M: speaker supervector

m: UBM supervector

T: **total variability matrix** (trained from training dataset)

Φ: random vector (called **i-vector**) ~ N(0, 1)

PLDA: Probabilistic Linear Discriminant Analysis

- PLDA: Jointly model **within-speaker** and **between-speaker** variabilities.
- For each i-vector of a speaker:

$$\Phi = \mu + s + c$$

Φ:	i-vector
μ:	overall mean of the training dataset
s:	speaker component
c:	channel / within-speaker variabilities component

PLDA model

$$
\begin{aligned}
Pr(x \mid \theta) &= N(x \mid \mu + s + c, \Sigma) \\
Pr(s) &= N(s \mid 0, B) \\
Pr(c) &= N(c \mid 0, M)
\end{aligned}
$$

The PLDA model is parameterized by **ϴ = {μ, B, M, Σ}**.

In the training phase, parameter **ϴ** (under which the data is **most likely)** is learned by **Expectation-Maximization** algorithm.

Scoring

Maximum a posteriori (MAP):

Choose the speaker corresponding to the model that gives highest probability:

$$Pr(M_i \mid x) = \frac{Pr(x \mid M_i)\,Pr(M_i)}{\sum_{j=1}^{R} Pr(x \mid M_j)\,Pr(M_j)}$$

References

1. https://blogs.technet.microsoft.com/machinelearning/2015/12/14/now-available-speaker-video-apis-from-microsoft-project-oxford/
2. *“Probabilistic Linear Discriminant Analysis for Inferences About Identity”* - Simon J.D. Prince, James H. Elder
3. *“An overview of text-independent speaker recognition: From features to supervectors”* - Tomi Kinnunen, Haizhou Li