Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

**SEEM3510**

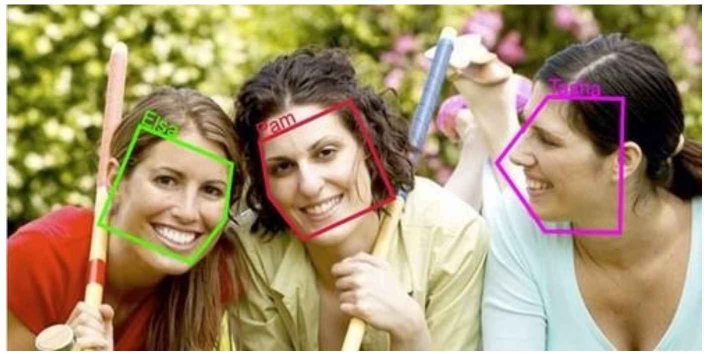
Lingwei Meng, Levi

# Introduction of speaker recognition

# Biometric recognition

- Face
- Fingerprint
- Palmprint
- DNA
- Iris
- Signatures
- Voiceprint

…



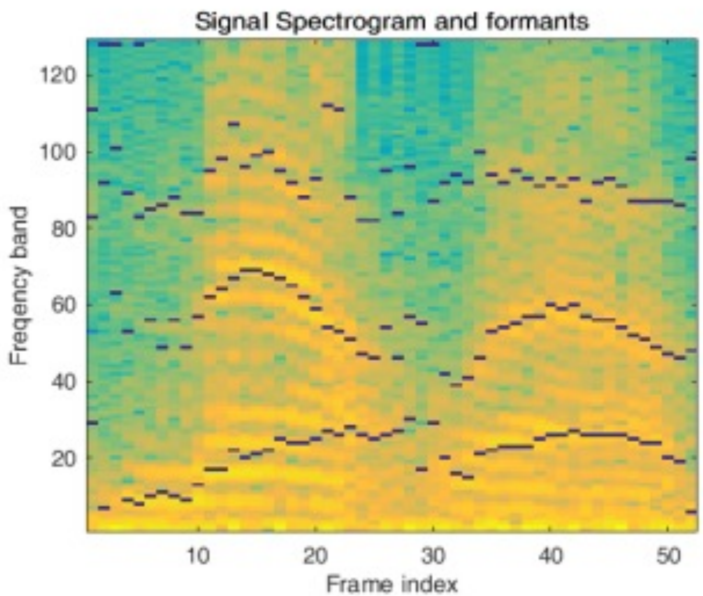


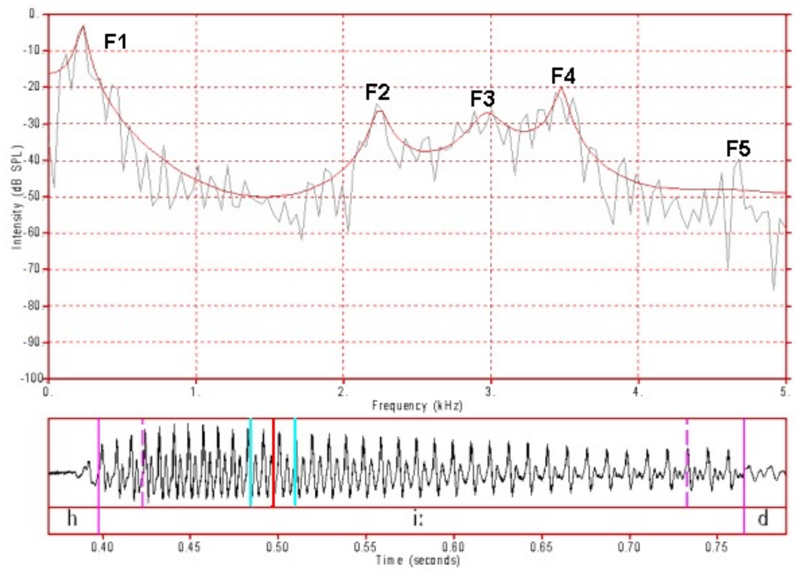
# Why voice biometric

- An efficient and more natural choice for remote authentication
  - Local: fingerprint, face
- A low-cost and convenient approach to authentication
- Voice is the most natural signal not involving privacy issues

# Spectrogram of the speech signals

- How to detect a word("matlab")? $spectrum(t, f) = STFT(wav(t))$



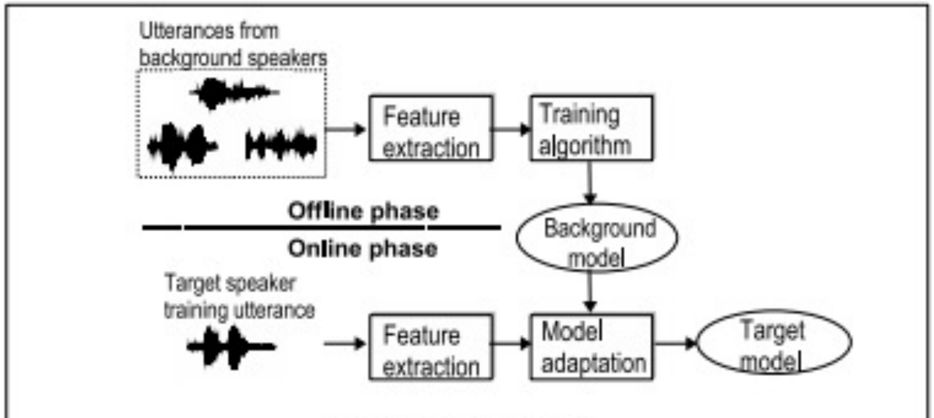

  - Convert Wave from time domain into frequency domain, using short Fourier fast transformation (STFT).
    - Resonance peaks
    - Harmonics frequencies - fundamental frequency

# Speaker recognition tasks

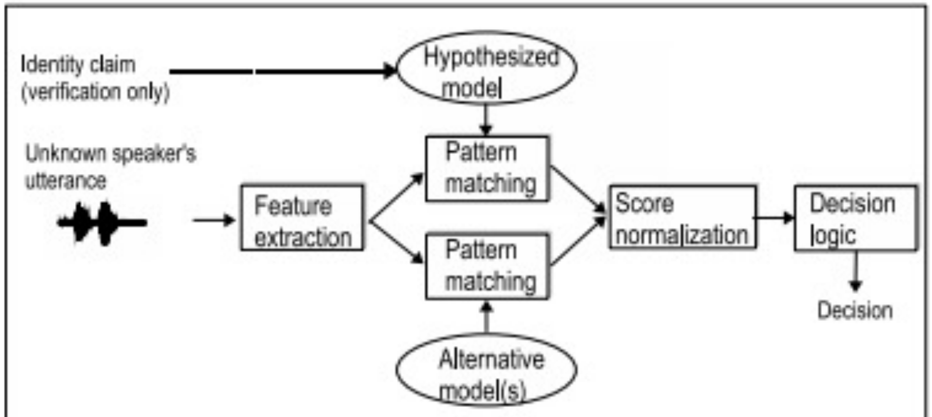
- Speaker recognition is the identification of a person from characteristics of voices (voice biometrics)
  - ***Verification* or *authentication:*** If the speaker claims to be of a certain identity and the voice is used to verify this claim, which is a 1:1 match task.
  - ***Identification*:** Determine an unknown speaker's identity with an utterance, which is a 1:N  match task.


# Speaker recognition system

- Speaker enrollment (training)
  - Collect target speakers' information from their utterances to train the target recognizer models.
  - Train a background model based on all speakers' information.
- Speaker verification (test)
  - The feature vectors extracted from the unknown speaker's utterance are compared against the target speaker model(s) to give a similarity score.

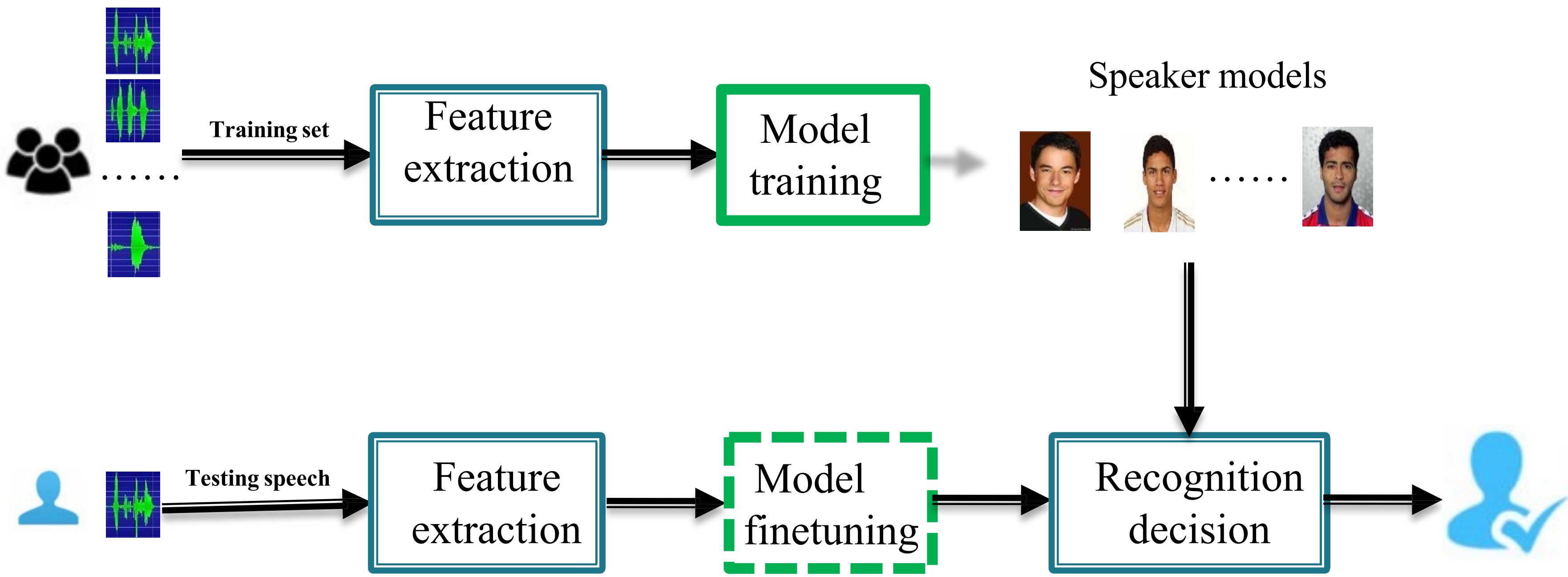



Speaker enrollment



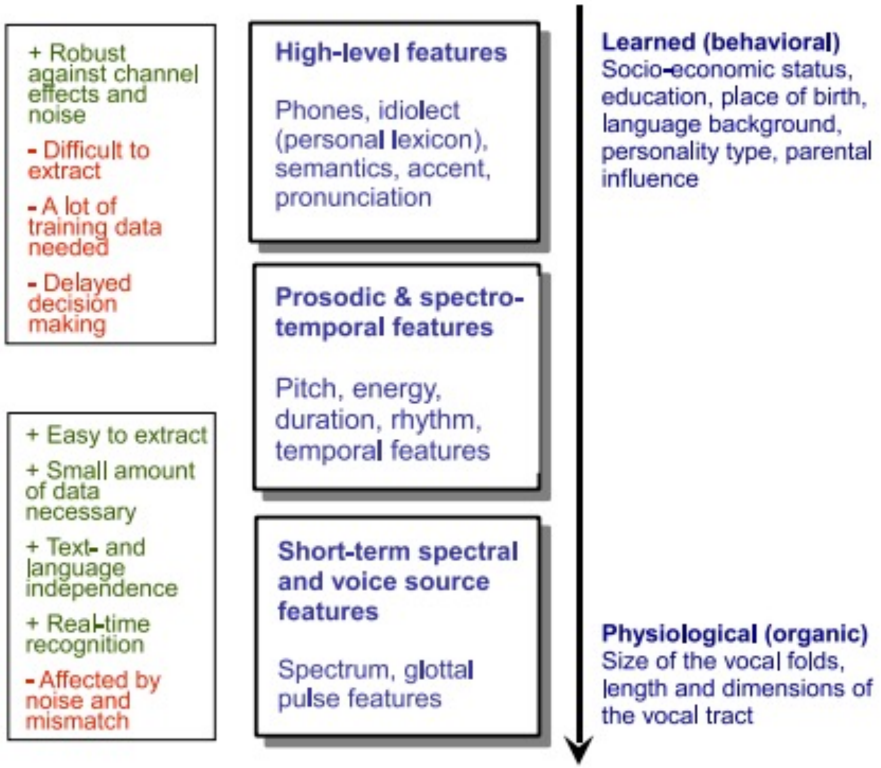

Speaker verification/identification

# Framework of speaker recognition



# Features for speaker recognition

- High-level features
  - Phones, idiolect, accent, pronunciation
    - "uh-huh", "you know", "oh yeah", etc.
    - Can be learned from familiar people
- Prosodic feature
  - Pitch, rhythm
- Short-term spectral
  - Spectrum (e.g. MFCC, filter bank)
    - Easy to extract, small amount
    - Vulnerable to noise and channels




Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, Volume 52, Issue 1, 2010,

# Universal background model combined with Gaussian Mixture model: GMM-UBM

- Universal background model (UBM) is first trained with the Expectation Maximization (EM) algorithm from a large number of speakers as GMMs.
  - When enrolling a new speaker to the system, the parameters of the UBM are adapted to the feature distribution of the new speaker.
  - The adapted model

$$P\left(X \mid \lambda^{ubm}\right) = \sum_{i=1}^{C} \bar{w}_i \mathrm{N}\left(X \mid \bar{\mu}_i, \bar{\Sigma}_i\right)$$

$$\lambda^{ubm} = \left\{\bar{w}_i, \bar{\mu}_i, \bar{\Sigma}_i\right\}$$

$X$ : input features

$i$ : gaussian component


Reynolds, D., Quatieri, T., and Dunn, R. Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10, 1 (January 2000), 19–41

# GMM-UBM: speaker adaptation

- Given the enrollment samples with T frames, and UBM parameters

$$X^k = \left[ X_1^k, X_2^k, \ldots, X_T^k \right] \qquad \lambda^{ubm} = \left\{ \bar{w}_i, \bar{\mu}_i, \bar{\Sigma}_i \right\}$$

- Using *maximum a posteriori* (MAP) method
  - Align each frame into different Gaussian components

$$P\left(\lambda_i^{ubm} \mid X_t^k\right) = \frac{\bar{w}_i \mathrm{N}\left(X_t^k \mid \bar{\mu}_i, \bar{\Sigma}_i\right)}{\sum_{i'=1}^{M} \bar{w}_{i'} \mathrm{N}\left(X_{i'}^k \mid \bar{\mu}_{i'}, \bar{\Sigma}_{i'}\right)}, i = 1, \ldots, C$$

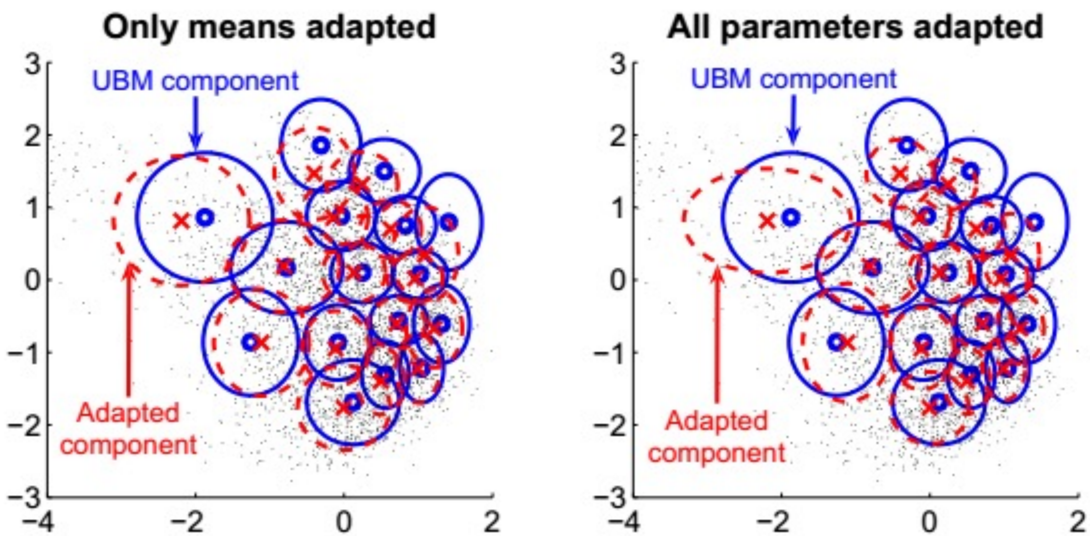
  - Compute the zero, first, and second statistics, and then update the means of the Gaussian Components as the target model

$$\tilde{\mu}_i^k = \frac{\sum_{t=1}^{T} P\left(\lambda_i \mid X_t^k\right) X_t^k}{\sum_{t=1}^{T} P\left(\lambda_i \mid X_t^k\right)}; \quad \mu_i^k = \left(1 - \alpha\right) \bar{\mu}_i + \alpha \tilde{\mu}_i^k, i = 1, \ldots, C$$

# Super-vector Methods

- *Gaussian supervector:* By stacking the *d*-dimensional ***mean*** *vectors* of a *K*-component adapted GMM into a *vector*.
  - it becomes possible to directly quantify and *remove the unwanted variability* from the supervectors.
  - **Data dimension reduction**: PCA,NAP, LDA, PLDA…

$$M = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_C \end{bmatrix}$$



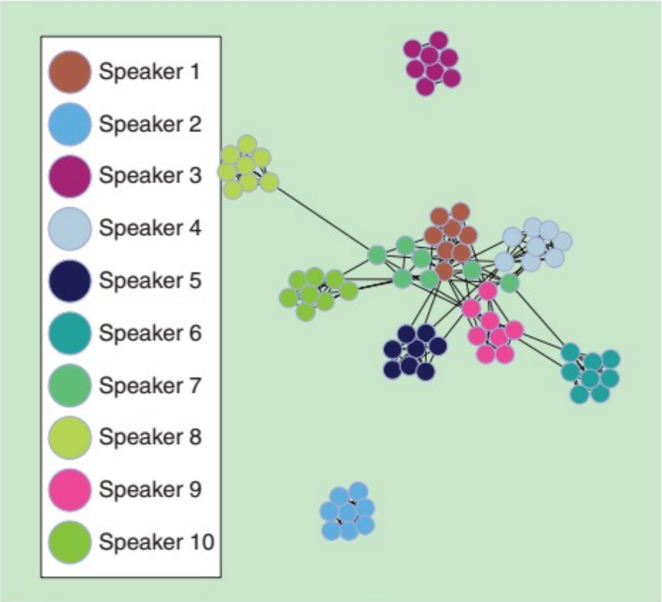
# i-vector and total variability space

- For i-vector, observing the fact that the channel factors also contain speaker-dependent information, the speaker and channel factors were combined into a single space termed the *total variability space*.

$$M = m + Tw$$

  - T: the total variable space
  - w: the i-vector
  - m is the UBM supervector
  - Cosine Similarity

$$D(w_1, w_2) = \frac{w_1^T w_2}{\|w_1\| \, \|w_2\|}$$




[FIG9] A graphical representation of 79 utterances spoken by ten individuals collected from the NIST SRE 2004 corpus. The i-vector representation is used for each segment; the plot is generated using GUESS, an open-source graph exploration software [123] that can visualize higher-dimensional data using distance measures between samples.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, 2011.
Some materials in Chinese

# Different tasks:

- **Text independent**: no constraint on the text
- **Text dependent**: fixed passphrase,
  - e.g. 'Hey Siri'
- **Text constrained**: fixed vocabulary,
  - e.g. randomly prompted *digit strings* (e.g. Please say 672193)

- Advances during last decade have enabled reliable authentication:
  - Text dependent with short fixed passphrases in clean scenarios
  - Text independent with relative long utterances (tens of seconds)

# Speaker recognition applications

- Military intelligence: initial application
- Criminal verification:
  - corroborate the voice samples for  forensic analysis
- Transaction authentication
  - Telephone banking: China Construction Bank (CCB)
  - Call center: Voicevault
  - Mobile app: log in with voiceprint

- e-Signatures: Call Center integration
  - https://vimeo.com/134044022

# Example with Echo

- Dialogues, Meeting minutes,
  - Who speak
  - [Teach Amazon Echo to Recognize Your Voice - YouTube]

# Challenges for speaker recognition / verification

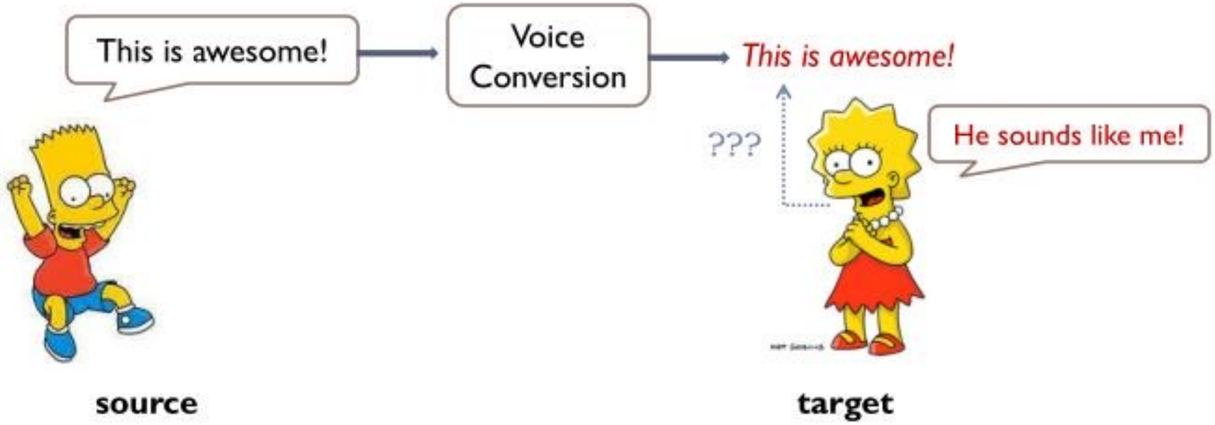
- Multi-channel
- Multi-speaker
- Noisy Environment
- Time-varying (or ageing)
  - Age or health problems
  - Emotions
- Short utterance

- *Speech Spoofing*
  - the impersonation of another person or the presentation of a pre-recorded or synthesized speech signals

# Challenges from speech spoofing

- How about replay attack?


- or speech synthesis, or *voice conversion*.
- [This AI Clones Your Voice After Listening for 5 Seconds 🤐 – YouTube]
- [MOS Test 1 (liusongxiang.github.io)]




# Thank you.