EE6424 Digital Audio Signal Processing
Part 2
Lecture 6:
Speech and Speaker Recognition –
A brief Introduction

Outline of lecture

- Feature extraction
- Automatic speech recognition
  - Acoustic model
  - Pronunciation dictionary
  - Language model
- Speaker Recognition
  - Speaker model
  - Verification and identification

EE6424 Part 2: Lecture 6.1

FEATURE EXTRACTION

Feature extraction

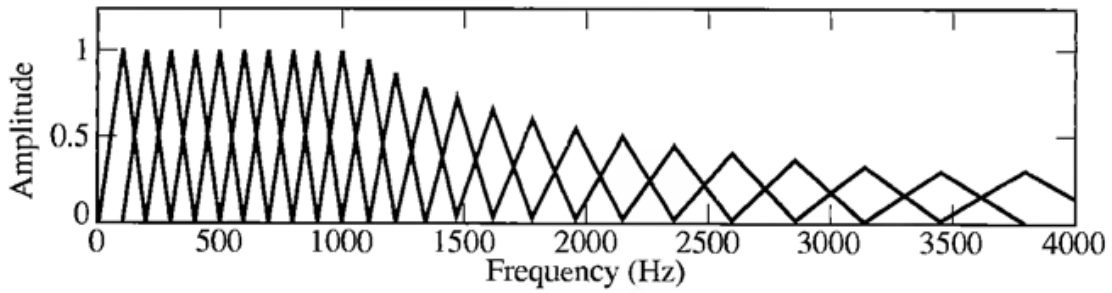
- The purpose of feature extraction from speech signal is to
  - **Retain** useful information (the linguistic information for speech recognition, speaker characteristic for speaker recognition).
  - **Remove** noise and other nuisance information.
- Various combinations of **acoustic**, **articulatory**, and **auditory** features have been utilized for speech and speaker recognition.
- The most popular acoustic features have been the **Mel-Frequency Cepstrum Coefficients** (MFCC). Other useful features include:
  - Linear Predictive Cepstral Coefficients (LPCC)
  - Perceptual Linear Prediction (PLP) Coefficients

Mel cepstrum

- The speech waveform $s(n)$ is first windowed with analysis window $w(n)$ and the discrete STFT is computed

$$S(n, \omega_k) = \sum_{m=-\infty}^{\infty} s(m) w(n-m) e^{-j\omega_k m}$$

- The magnitude of $S(n, \omega_k)$ is then weighted by a set of mel-scale filters, collectively referred as a mel-scale filter bank. Here, $\omega_k = \frac{2\pi}{N} k$ denotes discrete frequency with N being the DFT length.



- The mel-scale filter bank exploits auditory principles. This filter bank with triangularly-shaped frequency responses is a rough approximation to auditory critical-band filters.
- The center frequencies and bandwidth of mel-scale filters are linear for low frequency up to 1000 Hz and logarithmically increase with increasing frequency.

- Denote the frequency response of the lth mel-scale filter as $V_l(\omega_k)$. The output energy of the mel-scale filters, for $l = 1,2, \dots, R$, are computed as follows

$$E(n,l) = \frac{1}{A_l} \sum_{k=L_l}^{U_l} |V_l(\omega_k) S(n,\omega_k)|^2$$

  - L_l and U_l denote the lower and upper frequency indices over which the filter $V_l(\omega_k)$ is non-zero.
  - The factor A_l normalizes the filter $V_l(\omega_k)$ to unit energy

$$A_l = \sum_{k=L_l}^{U_l} |V_l(\omega_k)|^2$$

- The **mel-cepstrum** is computed for each speech frame at time n via Discrete Cosine Transform (DCT) as follows

$$C(n,m) = \frac{1}{R}\sum_{l=0}^{R-1} \log\{E(n,l)\} \cos\left(\frac{2\pi}{R}\right) lm$$

- Taking the log of the mel-scale filter energies before DCT leads to the name of **ceps**trum instead of **spec**trum. The term $\log\{E(n,l)\}$ is referred to as the mel-scale filter log energy.

- The end result of feature extraction:
  - Each frame is represented as a vector of M mel-frequency cepstral coefficients (MFCC), $C(n,m)$, where $M < R$. The number of filters R is usually in the range from 20 to 30, while the number of coefficients per frames, M, is usually from 10 to 20.
  - The speech signal $s(n)$ is represented as a sequence of MFCC vectors.

EE6424 Part 2: Lecture 6.2

INTRODUCING SPEECH RECOGNITION

Automatic speech recognition

- The goal of automatic speech recognition (ASR), or machine recognition of speech, is to convert a speech signal into a text message transcription of the spoken words.
- An automatic speech recognizer is in essence a digital equivalent of the **speech perception** part of the **speech chain** (see EE6424 Part 2 Lecture 2).
  - **Speech production**

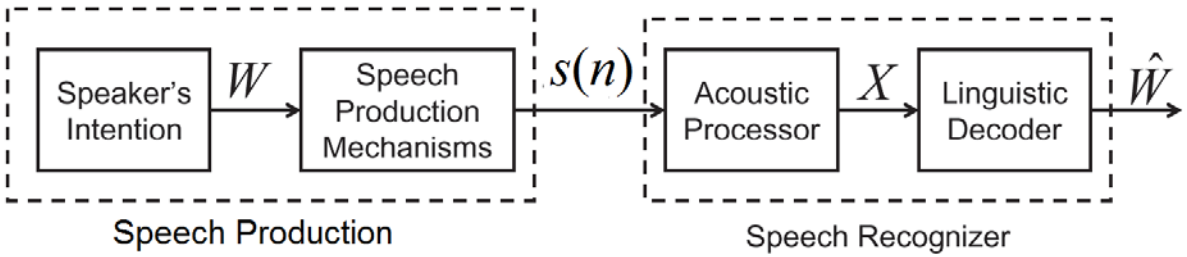
    A speaker expresses his **intention** or thought by composing a **linguistically** meaningful sentence W in the form of words.

    Appropriate control signals are then sent to the vocal organs to produce the speech **sounds** resulting in the speech waveform $s(n)$.

– **Speech recognizer**

An **acoustic** processor analyzes the speech signal (much similar to the ear) and converts it into a sequence of feature vectors, X (e.g., MFCC)

A **linguistic decoding** process (much similar to the auditory cortex) makes the best **estimate** of the spoken words, resulting in the recognized sentence $\widehat{W}$.



Major challenges

- ASR systems fall short of human speech perception in all but the simplest constrained tasks.
- Major factors that limit the performance:
  - **Channel** variability – differences in the microphones used to record the speech signal
  - **Acoustic** variability – differences in the acoustic environment in which the speaker is located (e.g., quiet office, noisy room, outdoors)
  - **Inter-speaker** variability – differences between speakers, e.g., speaker's accent
  - **Intra-speaker** variability – differences between instances where the same words are pronounced by a speaker.

Mathematical formulation of ASR

- A speech recognizer seeks to find the most likely sentence (word string) $\widehat{W}$ associated with a measured sequence of feature vectors, X, representing an input speech:

$$\widehat{W} = \underset{W_l}{\operatorname{argmax}}\, P(W_l|X)$$

- Using **Bayes rule**, we have

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)}$$

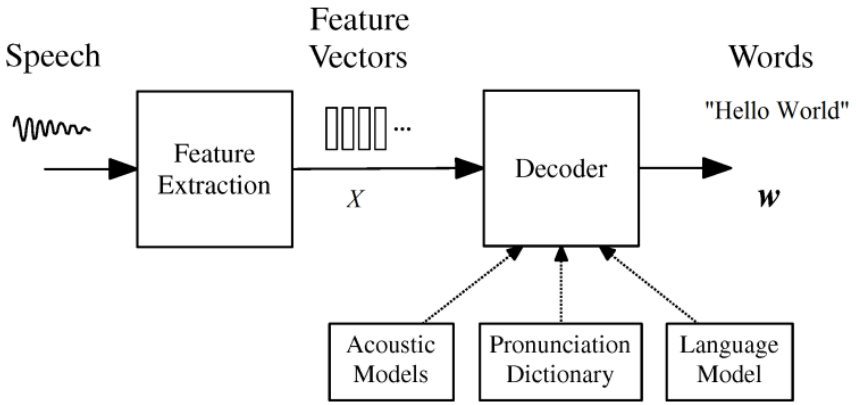
- As the probability $P(W|X)$ is difficult to model directly, we use the Bayes rule to transform the above optimization problem into an equivalent problem of finding

$$\widehat{W} = \underset{W_l}{\operatorname{argmax}}\, P(X|W_l)P(W_l)$$

- The probability $P(X|W)$ is determined by the **acoustic model**. It gives the likelihood that the word string W would produce the sequence X.
- The probability $P(W)$ is determined by the **language model**. It gives the probability of specific words occurring one after another in a certain language.
- The denominator $P(X)$ is ignored since it is independent of the word sequence W.
- The acoustic and language models are learned from a set of training data that has been labelled, usually by a human expert.

Components of an ASR system

- A typical speech recognition system consists of
  - Feature extraction front-end
  - Decoder
    - **Acoustic models** – represents the sound (acoustic) units
    - **Language models** – represents the linguistic rule (order of words)
    - **Pronunciation dictionary** – maps words to their pronunciations in terms of phonemes



- The speech recognition process is as follows:
  - The input speech signal is converted to a sequence of feature vectors X by the feature extraction front-end.
  - The decoder uses the set of acoustic models and pronunciation dictionary to provide a match score $P(X|W_l)$ for each possible sequence of words W_l. The language model is used to compute a score $P(W_l)$ for each sequence of words W_l.
  - The decoder operates by searching through all possible word sequences using pruning to remove unlikely hypothesis to keep the search tractable.
  - The decoder picks the most likely sequence of words $\widehat{W}$ that could have produced the input sequence of feature vectors X.

$$\widehat{W} = \underset{W_l}{\operatorname{argmax}}\, P(X|W_l)P(W_l)$$

Acoustic model

- The decoder matches the **feature vectors** with **reference patterns**, which are called **acoustic models**.
- The reference patterns are usually **Hidden Markov Models** (HMMs) trained for **words** or, more often, for **phones** as linguistic units.
- HMMs cope with **temporal variation**, which is important since the duration of individual phones may differ between the reference speech signal and the speech signal to be recognized.
- A linear normalization of the time axis is not sufficient here, since not all phones are expanded or compressed over time in the same way. For instance, stop consonants (/d/, /t/, /g/, /k/, /b/, and /p/) do not change their lengths much, whereas the lengths of vowels strongly depend on the overall speaking rate.

- Acoustic models for whole **words** are used only for speech recognizer with a very small vocabulary (much less than 1000).
- For English, the number of phones is around 40. For example, the word "bat" is composed of three phones /b/, /ae/, and /t/. This makes it practical to model the **basic linguistic unit** of speech as **phones**.
- During decoding, a word model is made by concatenating phone models. The composition of a word from phones is specified in the pronunciation dictionary.
- For any hypothesized word stream W_l, the corresponding acoustic model $P_A(X|W_l)$ is synthesized by concatenating words.
- The synthesized models are used to assign probabilities to the acoustic measurement X (i.e., matching the feature vector sequence with the reference patterns).

Pronunciation dictionary

- The **pronunciation** dictionary defines which combination of phones gives valid **words** for the recognition. It can contain information about different pronunciation variants of the same word.
- A **pronunciation dictionary** is essentially a **table** as is shown below. The words in the left column are related to their pronunciation (phones) in the right column.

word	*pronunciation*
INCREASE	ih n k r iy s
INCREASED	ih n k r iy s t
INCREASES	ih n k r iy s ah z
INCREASING	ih n k r iy s ih ng
INCREASINGLY	ih n k r iy s ih ng l iy
INCREDIBLE	ih n k r eh d ah b ah l

- The size of the vocabulary determines the complexity (the search space of the decoder) of the speech recognition task:
  - **Small** vocabulary: less than 1000 words
  - **Medium** vocabulary: between 1000 to 10,000 words
  - **Large** vocabulary: more than 10,000 words
- There are words which are not covered by the dictionary. These are known as the **out-of-vocabulary** (OOV) words.

Language model

- A language model assigns a **probability** to a sequence of **words**.
- The most widely used language model is the **n-gram** language models. An n-gram language model approximates the conditional probability of a given word w_k as follows

$$P(w_k|w_1, w_2, \dots, w_{k-1}) \approx P(w_k|w_{k-n+1}, w_{k-n+2}, \dots, w_{k-1})$$

- The probability of a word w_k given all the previous words can be approximated by the probability given only $n-1$ previous words.
- Commonly used n-gram language models are **unigram**, **bigram**, **trigram**, and **quadrigram**, for $n = 1$, 2, 3, and 4, respectively.
- For **unigram**, the probability of word is simply given by the frequency of the word itself, being independent of the word history.

- n-gram models can be trained by counting and normalizing (i.e., dividing by total count so that the resulting probabilities fall legally between 0 and 1) occurrences of words in training corpus (a text corpus, for example, paragraphs in newspaper).

$$
\begin{aligned}
&P(\text{rabbit}|\text{Just the other day I saw a}) \\
&\approx P(\text{rabbit}|\text{day I saw a}) \\
&\approx P(\text{rabbit}|\text{ I saw a}) \\
&\approx P(\text{rabbit}|\text{saw a}) \\
&\approx P(\text{rabbit}|\text{ a}) \\
&\approx P(\text{rabbit})
\end{aligned}
$$

Example 1

- Vocabulary: {A, BOOK, BUY, CAR, HAVE, I, NEW, RED, THEY}
- Unigram language model:

$P(\mathrm{A})$	$3/15$	$P(\mathrm{BOOK})$	$1/15$	$P(\mathrm{BUY})$	$1/15$
$P(\mathrm{CAR})$	$2/15$	$P(\mathrm{HAVE})$	$2/15$	$P(\mathrm{I})$	$2/15$
$P(\mathrm{NEW})$	$2/15$	$P(\mathrm{RED})$	$1/15$	$P(\mathrm{THEY})$	$1/15$

  - Test sentence:
    - <s> I BUY A NEW BOOK </s>
    - <s> I BUY A NEW CAR </s>
  - Note: <s> and </s> indicates START and END of a sentence.

- Probability:

$$P(\text{I BUY A NEW BOOK}) \approx P(\text{I})P(\text{BUY})P(\text{A})P(\text{NEW})P(\text{BOOK})$$
$$= \frac{2}{15} \times \frac{1}{15} \times \frac{3}{15} \times \frac{2}{15} \times \frac{1}{15} = \frac{12}{15^5} = 1.58 \times 10^{-5}$$

$$P(\text{I BUY A NEW CAR}) \approx P(\text{I})P(\text{BUY})P(\text{A})P(\text{NEW})P(\text{CAR})$$
$$= \frac{2}{15} \times \frac{1}{15} \times \frac{3}{15} \times \frac{2}{15} \times \frac{2}{15} = \frac{24}{15^5} = 3.16 \times 10^{-5}$$

- Given the same number of words in the two sentences, the second sentence is more likely to occur compared to the first sentence.
- Both are grammatically correct, it reflects the facts that language model is strongly dependent on the training corpus.

Example 2

- **Bi-gram** language model:

$P(\mathrm{I}\mid <\mathrm{S}>)$	2/3	$P(\mathrm{THEY}\mid <\mathrm{S}>)$	1/3	$P(\mathrm{NEW}\mid\mathrm{A})$	2/3
$P(\mathrm{RED}\mid\mathrm{A})$	1/3	$P(\mathrm{A}\mid\mathrm{BUY})$	1/1	$P(\mathrm{A}\mid\mathrm{HAVE})$	2/2
$P(\mathrm{HAVE}\mid\mathrm{I})$	1/2	$P(\mathrm{BUY}\mid\mathrm{I})$	1/2	$P(\mathrm{CAR}\mid\mathrm{NEW})$	1/2
$P(\mathrm{BOOK}\mid\mathrm{NEW})$	1/2	$P(\mathrm{CAR}\mid\mathrm{RED})$	1/1	$P(\mathrm{HAVE}\mid\mathrm{THEY})$	1/1

- Probability:

$$P(\mathrm{I\ BUY\ A\ NEW\ BOOK}) \approx P(\mathrm{I}\mid <s>)P(\mathrm{BUY}\mid\mathrm{I})P(\mathrm{A}\mid\mathrm{BUY})P(\mathrm{NEW}\mid\mathrm{A})P(\mathrm{BOOK}\mid\mathrm{NEW})$$
$$= \frac{2}{3}\times\frac{1}{2}\times\frac{1}{1}\times\frac{2}{3}\times\frac{1}{2} = \frac{4}{36} = \frac{1}{9} = 1.11\times 10^{-1}$$

$$P(\mathrm{I\ BUY\ A\ NEW\ CAR}) \approx P(\mathrm{I}\mid <s>)P(\mathrm{BUY}\mid\mathrm{I})P(\mathrm{A}\mid\mathrm{BUY})P(\mathrm{NEW}\mid\mathrm{A})P(\mathrm{CAR}\mid\mathrm{NEW})$$
$$= \frac{2}{3}\times\frac{1}{2}\times\frac{1}{1}\times\frac{2}{3}\times\frac{1}{2} = \frac{4}{36} = \frac{1}{9} = 1.11\times 10^{-1}$$

- A bi-gram language model gives a better approximation than a unigram.

EE6424 Part 2: Lecture 6.3

INTRODUCING SPEAKER RECOGNITION

Individual characteristics

- Speaker recognition refers to the automatic recognition of a speaker (or talker) through measurements of characteristics arising in the speaker's voice signal.
- A spoken message conveys information about the speaker in addition to the meaning of the message.
- Individual speaker is characterized by a variety of voice attributes:
  - **High-level** attributes:
    - Prosody (i.e., pitch intonation)
    - Accents
    - Choice of words

- **Low-level** attributes:
  - Vocal tract spectrum
  - Pitch period
  - Formant trajectory

- High-level attributes are related to the behavioral differences in the manner of speaking and are difficult to extract by machine for automatic speaker recognition (though fairly easy for human)

- Low-level attributes are related to physiological aspect of vocal organ (mostly the vocal tract) and are more measurable given their acoustic nature.

Input and decision modes

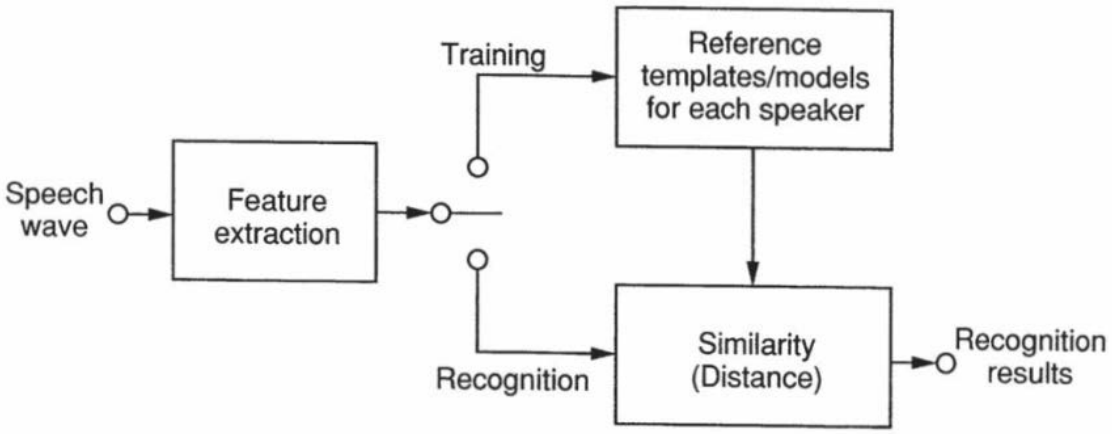
- **Input** modes
  - **Text dependent**:
    - The content of the speech is known (e.g., speaker prompted with text to speak)
    - More accurate
  - **Text independent**:
    - Unconstrained but less accurate
    - Can be applied to ‘found’ speech (no control over the content)
- **Decision** modes
  - **Identification**
    - Speech sample from an unknown speaker is compared with models of registered speakers. The unknown speaker is identified as the speaker whose model best matches the input speech sample.
    - One-to-many matching

– **Verification**

- An identity claim is made by an unknown speaker. Speech sample from the speaker is compared with the model of the claimed identity. The identify claim is accepted if the match is good enough (i.e., passes a given threshold).
- One-to-one matching

Structure of speaker recognition system

- Feature parameters extracted from a speech wave are compared with stored **models** (or reference templates) for each registered speakers.
- The recognition decision is made according to the **distance** (or similarity) values. The distance measure is closely related to the type of model and algorithm used.



Gaussian mixture model

- A speaker **model** (or reference template) is constructed using **enrollment** utterances from that speaker. Each enrollment utterance X is a **sequence of feature vectors** $\{x_t\}_{t=1}^{T}$ generated by the feature extraction front-end.
- For the case of text-independent speaker recognition, where the system has no prior knowledge of the text of speaker's utterance, **Gaussian mixture models** (GMMs) have proven to be effective.

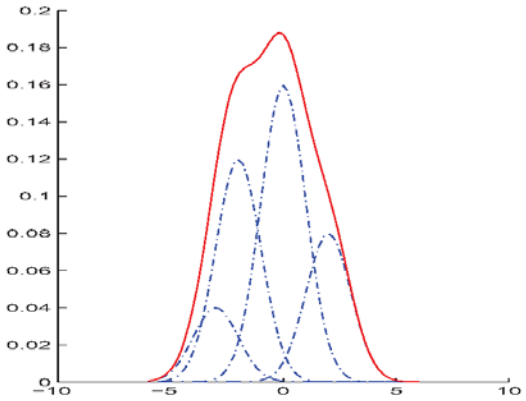
$$p(\mathbf{x}\,|\,\theta) = \sum_{k=1}^{K} w_k \mathcal{N}\left(\mathbf{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

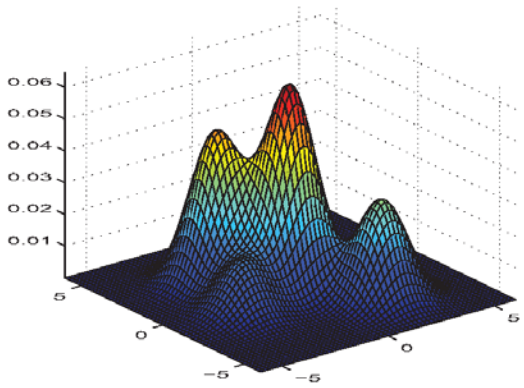
  - The **weights** w_k, for $k = 1,2,\ldots,K$, always sum to 1 such that the resulting mixture $p(\mathbf{x}|\theta)$ is a legitimate probability distribution.
  - The set $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k\}_{k=1}^{K}$ represents the **parameters** of the distribution

- The Gaussian or normal density is given by

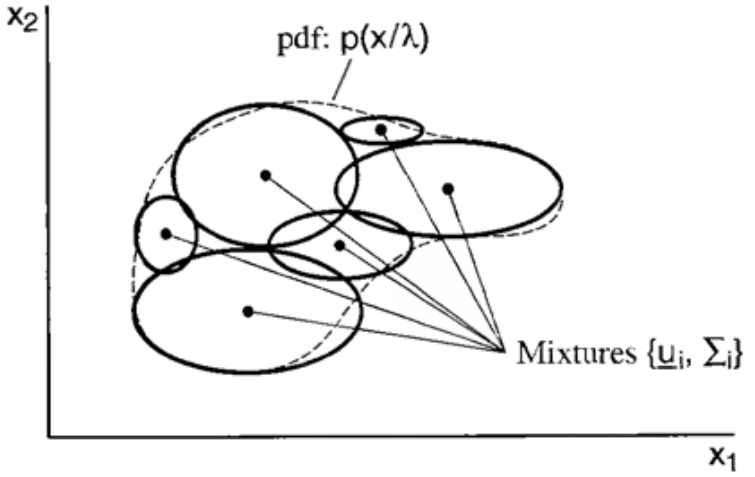
$$\mathcal{N}\left(\mathbf{x}\,|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)=\frac{1}{\sqrt{\left(2\pi\right)^D\left|\boldsymbol{\Sigma}_k\right|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)\right]$$

- Feature vectors of the enrollment utterance $\{\mathbf{x}_t\}_{t=1}^{T}$ are assumed to be drawn from a probability density function (pdf) that is a **mixture** of K Gaussians.





- A GMM can be interpreted as representation of various **acoustic classes** that make up the sounds of a speaker.
- Each **component** density can be thought of as an **acoustic class**, each representing one speech sound (e.g., a particular phoneme) or a set of speech sounds (voiced, unvoiced, fricative, diphthong etc.).



Expectation maximization (EM) for GMM

- Given the enrollment data $\{\mathbf{x}_t\}_{t=1}^{T}$, the maximum likelihood estimate of $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k\}_{k=1}^{K}$ can be obtained using the EM algorithm.
- **E-step**: compute the membership of each feature vector to the K Gaussians

$$\lambda_i(\mathbf{x}_t) = \frac{\mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot w_k}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot w_k}$$

- **M-step**: Update the mean $\boldsymbol{\mu}_k$, covariance matrices $\boldsymbol{\Sigma}_k$, and weights w_i based on the membership information $\lambda_k(\mathbf{x}_t)$ of each frame

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \times \sum_{t=1}^{T} \lambda_k(\mathbf{x}_t) \cdot \mathbf{x}_t$$

$$\boldsymbol{\Sigma}_k = \frac{1}{n_k} \times \sum_{t=1}^{T} \lambda_k(\mathbf{x}_t) \cdot (\mathbf{x}_t - \boldsymbol{\mu}_k)(\mathbf{x}_t - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$w_k = \frac{1}{T} \underbrace{\sum_{t=1}^{T} \lambda_k(\mathbf{x}_t)}_{n_k}$$

Identification versus verification

- The **quality of match** of a test utterance to a speaker **model** θ is taken as the **average log-likelihood** given by

$$s(Y \mid \theta) = \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{y}_t \mid \theta)$$

- For an **identification** task, speaker model with maximum score is selected.
- For a **verification** task, the decision is based on log-likelihood ratio of the following form

$$\Lambda(Y) = s(Y \mid \theta) - s(Y \mid \theta_{bg})$$

  - θ is the speaker model while θ_{bg} represents a **background** model.
  - The log-likelihood ratio $\Lambda(Y)$ is to be compared to a threshold α so as to **accept** (if $\Lambda(Y) \geq \alpha$) or **reject** (if $\Lambda(Y) < \alpha$) the identity claim.

Summary

- The main objective of spoken **communication** is centered on the spoken **message**.
- Variety due to the **speaker** (i.e., the transmitter of the message) is seen as **unwanted variability** and is compensated for by the listener.
- A key goal in speech recognition is to **neutralize inter-speaker variability**.
- In speaker recognition the focus is put on the characteristics of the speaker.
  - **Inter-speaker** variability factors become essential to characterize the voice of a particular speaker.
  - **Intra-speaker** variability, due to noise, microphone or channel variations, becomes predominant and plays a limiting role.