CS578- Speech Signal Processing

Lecture 9: Speaker Recognition

Yannis Stylianou

University of Crete, Computer Science Dept., Multimedia Informatics Lab
yannis@csd.uoc.gr

Univ. of Crete

Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

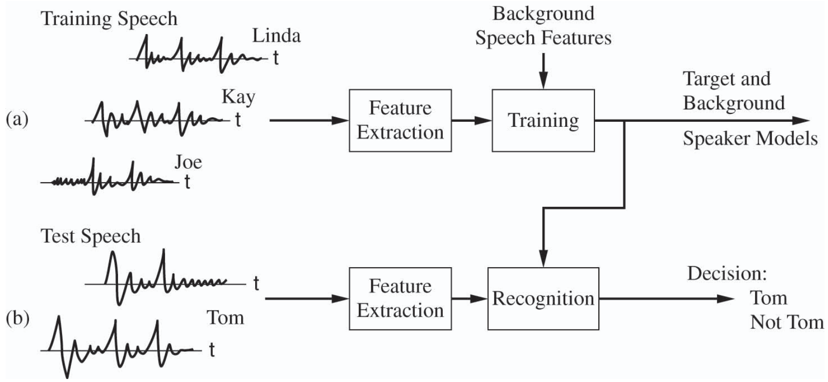
Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

Terms

- Speaker identification
- Speaker verification
- Claimant (Target speaker)
- Imposter (Background speaker)
- False acceptances/false rejections
- Features
- Training stage/testing stage
- Mismatch conditions

Overview Of A Speaker Verification System



OUTLINE

Cues for recognition: High Level

- Clarity
- Roughness
- Animation
- Magnitude
- Pitch intonation
- Articulation rate
- Dialect

Cues for recognition: High Level

- Vocal tract spectrum
- Instantaneous pitch
- Glottal flow excitation
- Modulations in formant trajectories

Mel-Cepstrum

- Compute STFT:

$$X(n, \omega_k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega_k m}$$

where $\omega_k = \frac{2\pi}{N}k$ with N the DFT length

- Apply *mel-scale filters* $V_l(\omega_k)$ on $|X(n, \omega_k)|$:

$$|V_l(\omega_k)\, X(n, \omega_k)|$$

- Compute the energy in each mel-frequency band:

$$E_{mel}(n, l) = \frac{1}{A_k} \sum_{k=L_l}^{U_l} |V_l(\omega_k)\, X(n, \omega_k)|^2$$

where L_l and U_l denote the lower and upper limit of the lth filter and

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

- Compute *mel-cepstrum*:

$$C_{mel}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log\left(E_{mel}(n, l)\right) \cos\left(\frac{2\pi}{R} lm\right)$$

where R is the number of filters.

Mel-Cepstrum

- Compute STFT:

$$X(n, \omega_k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega_k m}$$

where $\omega_k = \frac{2\pi}{N} k$ with N the DFT length

- Apply *mel-scale filters* $V_l(\omega_k)$ on $|X(n, \omega_k)|$:

$$|V_l(\omega_k)\, X(n, \omega_k)|$$

- Compute the energy in each mel-frequency band:

$$E_{mel}(n, l) = \frac{1}{A_k} \sum_{k=L_l}^{U_l} |V_l(\omega_k)\, X(n, \omega_k)|^2$$

where L_l and U_l denote the lower and upper limit of the lth filter and

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

- Compute *mel-cepstrum*:

$$C_{mel}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log\left(E_{mel}(n, l)\right) \cos\left(\frac{2\pi}{R} lm\right)$$

where R is the number of filters.

Mel-Cepstrum

- Compute STFT:

$$X(n, \omega_k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega_k m}$$

where $\omega_k = \frac{2\pi}{N} k$ with N the DFT length

- Apply *mel-scale filters* $V_l(\omega_k)$ on $|X(n, \omega_k)|$:

$$|V_l(\omega_k)\, X(n, \omega_k)|$$

- Compute the energy in each mel-frequency band:

$$E_{mel}(n, l) = \frac{1}{A_k} \sum_{k=L_l}^{U_l} |V_l(\omega_k)\, X(n, \omega_k)|^2$$

where L_l and U_l denote the lower and upper limit of the lth filter and

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

- Compute *mel-cepstrum*:

$$C_{mel}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log\left(E_{mel}(n, l)\right) \cos\left(\frac{2\pi}{R} lm\right)$$

where R is the number of filters.

Mel-Cepstrum

- Compute STFT:

$$X(n, \omega_k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega_k m}$$

  where $\omega_k = \frac{2\pi}{N} k$ with N the DFT length

- Apply *mel-scale filters* $V_l(\omega_k)$ on $|X(n, \omega_k)|$:

$$|V_l(\omega_k)\, X(n, \omega_k)|$$

- Compute the energy in each mel-frequency band:

$$E_{mel}(n, l) = \frac{1}{A_k} \sum_{k=L_l}^{U_l} |V_l(\omega_k)\, X(n, \omega_k)|^2$$

  where L_l and U_l denote the lower and upper limit of the lth filter and

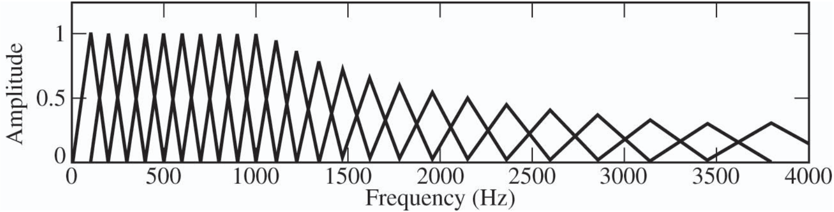
$$A_l = \sum_{k=L_l}^{U_l} |V_l(\omega_k)|^2$$

- Compute *mel-cepstrum*:

$$C_{mel}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log\left(E_{mel}(n, l)\right) \cos\left(\frac{2\pi}{R} lm\right)$$

  where R is the number of filters.

Triangular mel-scale filter bank



Sub-Cepstrum

- Convolve mel-scale filter impulse response $u_l[n]$ (*subband filter*) with $x[n]$:

$$\tilde{X}(n, \omega_l) = x[n] \star u_l[n]$$

- Compute energy:

$$E_{sub}(n, l) = \sum_{m=-N/2}^{N/2} p[n-m] |\tilde{X}(n, \omega_l)|^2$$

where $p[n]$ is a smoothing filter.

- Compute *subband cepstrum*:

$$C_{sub}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log\left(E_{sub}(n, l)\right) \cos\left(\frac{2\pi}{R} lm\right)$$

Sub-Cepstrum

- Convolve mel-scale filter impulse response $u_l[n]$ (*subband filter*) with $x[n]$:

$$\tilde{X}(n, \omega_l) = x[n] \star u_l[n]$$

- Compute energy:

$$E_{sub}(n, l) = \sum_{m=-N/2}^{N/2} p[n-m] |\tilde{X}(n, \omega_l)|^2$$

where $p[n]$ is a smoothing filter.

- Compute *subband cepstrum*:

$$C_{sub}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log \left(E_{sub}(n, l) \right) \cos \left(\frac{2\pi}{R} lm \right)$$

Sub-Cepstrum

- Convolve mel-scale filter impulse response $u_l[n]$ (*subband filter*) with $x[n]$:

$$\tilde{X}(n, \omega_l) = x[n] \star u_l[n]$$

- Compute energy:

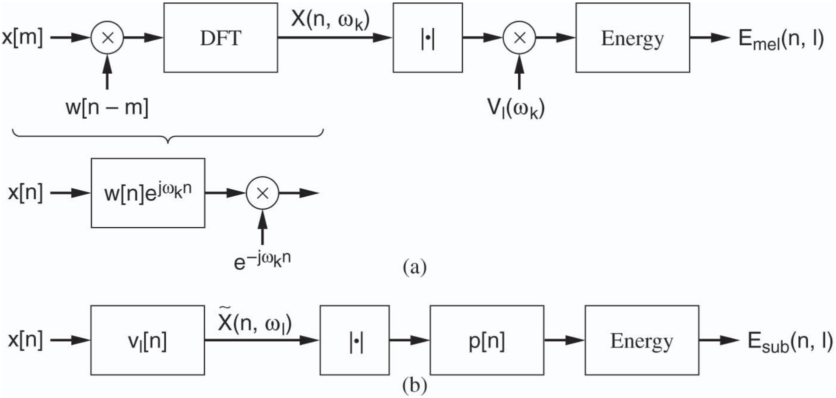
$$E_{sub}(n, l) = \sum_{m=-N/2}^{N/2} p[n-m] |\tilde{X}(n, \omega_l)|^2$$

where $p[n]$ is a smoothing filter.

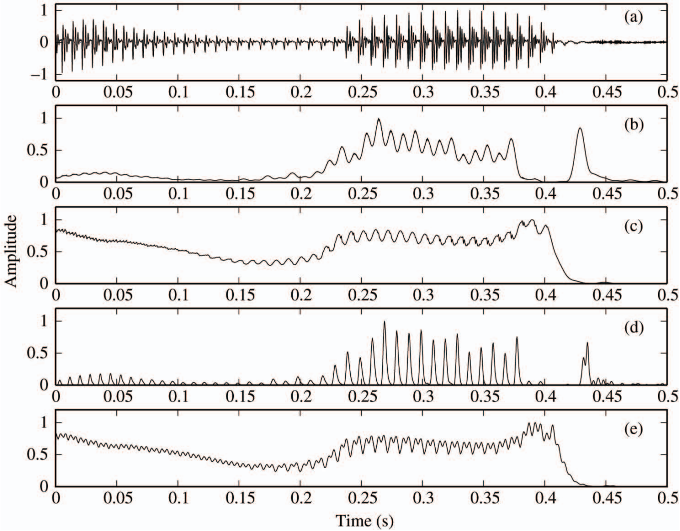
- Compute *subband cepstrum*:

$$C_{sub}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log\left(E_{sub}(n, l)\right) \cos\left(\frac{2\pi}{R} lm\right)$$

Comparing Mel-Cepstrum and Sub-Cepstrum



Energies from mel-scale and subband filter banks



Outline

Minimum-Distance Classifier

- Compute the average (mel or subband) cepstral features for the training data:

$$\bar{C}^{tr}[n] = \frac{1}{M}\sum_{m=1}^{M} C^{tr}[mL, n]$$

where L denotes the frame length.

- Compute the average cepstral features for the testing data:

$$\bar{C}^{ts}[n] = \frac{1}{M'}\sum_{m=1}^{M'} C^{ts}[mL, n]$$

- Compute a distance:

$$D = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}^{tr}[n] - \bar{C}^{ts}[n])^2$$

- For speaker verification:

$$\text{if } D > \text{ Threshold, then speaker is verified}$$

Minimum-Distance Classifier

- Compute the average (mel or subband) cepstral features for the training data:

$$\bar{C}^{tr}[n] = \frac{1}{M}\sum_{m=1}^{M} C^{tr}[mL, n]$$

  where L denotes the frame length.

- Compute the average cepstral features for the testing data:

$$\bar{C}^{ts}[n] = \frac{1}{M'}\sum_{m=1}^{M'} C^{ts}[mL, n]$$

- Compute a distance:

$$D = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}^{tr}[n] - \bar{C}^{ts}[n])^2$$

- For speaker verification:

$$\text{if } D > \text{ Threshold, then speaker is verified}$$

Minimum-Distance Classifier

- Compute the average (mel or subband) cepstral features for the training data:

$$\bar{C}^{tr}[n] = \frac{1}{M}\sum_{m=1}^{M} C^{tr}[mL, n]$$

where L denotes the frame length.

- Compute the average cepstral features for the testing data:

$$\bar{C}^{ts}[n] = \frac{1}{M'}\sum_{m=1}^{M'} C^{ts}[mL, n]$$

- Compute a distance:

$$D = \frac{1}{R-1}\sum_{n=1}^{R-1} (\bar{C}^{tr}[n] - \bar{C}^{ts}[n])^2$$

- For speaker verification:

$$\text{if } D > \text{ Threshold, then speaker is verified}$$

Minimum-Distance Classifier

- Compute the average (mel or subband) cepstral features for the training data:

$$\bar{C}^{tr}[n] = \frac{1}{M}\sum_{m=1}^{M} C^{tr}[mL, n]$$

where L denotes the frame length.

- Compute the average cepstral features for the testing data:

$$\bar{C}^{ts}[n] = \frac{1}{M'}\sum_{m=1}^{M'} C^{ts}[mL, n]$$

- Compute a distance:

$$D = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}^{tr}[n] - \bar{C}^{ts}[n])^2$$

- For speaker verification:

if $D >$ Threshold, then speaker is verified

Using Acoustic Classes

- Let assume we know the acoustic class of each speech segment
- For each class i compute the mean:

$$\begin{aligned} \bar{C}_i^{tr}[n] &= \frac{1}{M}\sum_{m=1}^{M} C_i^{tr}[mL, n] \\ \bar{C}_i^{ts}[n] &= \frac{1}{M'}\sum_{m=1}^{M'} C_i^{ts}[mL, n] \end{aligned}$$

- Compute the Euclidean distance in each class:

$$D_i = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}_i^{tr}[n] - \bar{C}_i^{ts}[n])^2$$

- Average over all classes:

$$D = \frac{1}{I}\sum_{i=1}^{I} D_i$$

- Use D as previously for speaker verification (or identification)

Using Acoustic Classes

- Let assume we know the acoustic class of each speech segment
- For each class i compute the mean:

$$\begin{aligned} \bar{C}_i^{tr}[n] &= \frac{1}{M}\sum_{m=1}^{M} C_i^{tr}[mL, n] \\ \bar{C}_i^{ts}[n] &= \frac{1}{M'}\sum_{m=1}^{M'} C_i^{ts}[mL, n] \end{aligned}$$

- Compute the Euclidean distance in each class:

$$D_i = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}_i^{tr}[n] - \bar{C}_i^{ts}[n])^2$$

- Average over all classes:

$$D = \frac{1}{I}\sum_{i=1}^{I} D_i$$

- Use D as previously for speaker verification (or identification)

Using Acoustic Classes

- Let assume we know the acoustic class of each speech segment
- For each class i compute the mean:

$$\begin{aligned}\bar{C}_i^{tr}[n] &= \tfrac{1}{M}\sum_{m=1}^{M} C_i^{tr}[mL, n] \\ \bar{C}_i^{ts}[n] &= \tfrac{1}{M'}\sum_{m=1}^{M'} C_i^{ts}[mL, n]\end{aligned}$$

- Compute the Euclidean distance in each class:

$$D_i = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}_i^{tr}[n] - \bar{C}_i^{ts}[n])^2$$

- Average over all classes:

$$D = \frac{1}{I}\sum_{i=1}^{I} D_i$$

- Use D as previously for speaker verification (or identification)

Using Acoustic Classes

- Let assume we know the acoustic class of each speech segment
- For each class i compute the mean:

$$\begin{aligned}\bar{C}_i^{tr}[n] &= \tfrac{1}{M}\sum_{m=1}^{M} C_i^{tr}[mL, n] \\ \bar{C}_i^{ts}[n] &= \tfrac{1}{M'}\sum_{m=1}^{M'} C_i^{ts}[mL, n]\end{aligned}$$

- Compute the Euclidean distance in each class:

$$D_i = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}_i^{tr}[n] - \bar{C}_i^{ts}[n])^2$$

- Average over all classes:

$$D = \frac{1}{I}\sum_{i=1}^{I} D_i$$

- Use D as previously for speaker verification (or identification)

Using Acoustic Classes

- Let assume we know the acoustic class of each speech segment
- For each class i compute the mean:

$$\begin{aligned} \bar{C}_i^{tr}[n] &= \frac{1}{M}\sum_{m=1}^{M} C_i^{tr}[mL,n] \\ \bar{C}_i^{ts}[n] &= \frac{1}{M'}\sum_{m=1}^{M'} C_i^{ts}[mL,n] \end{aligned}$$

- Compute the Euclidean distance in each class:

$$D_i = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}_i^{tr}[n] - \bar{C}_i^{ts}[n])^2$$

- Average over all classes:

$$D = \frac{1}{I}\sum_{i=1}^{I} D_i$$

- Use D as previously for speaker verification (or identification)

Multivariate Gaussian pdf

Let $\mathbf{x}$ be a $d \times 1$ vector

- Gaussian pdf:

$$g_{\mu,\mathbf{\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^d \sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

  where μ is the mean vector and $\mathbf{\Sigma}$ the covariance matrix.

- Estimation of the mean:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

- Estimation of the (unbiased) covariance matrix:

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

Multivariate Gaussian pdf

Let $\mathbf{x}$ be a $d \times 1$ vector

- Gaussian pdf:

$$g_{\mu,\mathbf{\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^d \sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu)}$$

  where μ is the mean vector and $\mathbf{\Sigma}$ the covariance matrix.

- Estimation of the mean:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

- Estimation of the (unbiased) covariance matrix:

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

Multivariate Gaussian pdf

Let $\mathbf{x}$ be a $d \times 1$ vector

- Gaussian pdf:

$$g_{\mu,\mathbf{\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^d \sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

  where μ is the mean vector and $\mathbf{\Sigma}$ the covariance matrix.

- Estimation of the mean:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

- Estimation of the (unbiased) covariance matrix:

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

Gaussian Mixture Model - GMM

- Mixture of Gaussian PDFs

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} p(q_k|\mathbf{x}) g_{\mu_k, \mathbf{\Sigma}_k}(\mathbf{x})$$

  where

$$\sum_{k=1}^{K} p(q_k|\mathbf{x}) = 1$$

- Speaker model, θ

$$\theta = \{p_k, \mu_k, \mathbf{\Sigma}_k\}$$

  for $k = 1, 2, \cdots, K$

Speaker Identification

- If we have estimated S target speaker models θ_j with $j = 1, 2, \cdots, S$.
- *Maximum a posteriori probability classification*:

$$\max_{\theta_j} P(\theta_j|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\theta_j)P(\theta_j)}{\sum_{j=1}^{S} p(\mathbf{x}_i|\theta_j)}$$

- *Maximum Likelihood*:

$$\max_{\theta_j} p(\mathbf{x}_i|\theta_j)$$

- if $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots \mathbf{x}_{M-1}\}$ and assuming frames are independent:

$$p(\mathbf{X}|\theta_j) = \prod_{i=0}^{M-1} p(\mathbf{x}_i|\theta_j)$$

- Speaker identification:

$$\hat{S} = \max_{1 \leq j \leq S} \sum_{i=0}^{M-1} \log\left[p(\mathbf{x}_i|\theta_j)\right]$$

Speaker Identification

- If we have estimated S target speaker models θ_j with $j = 1, 2, \cdots, S$.
- *Maximum a posteriori probability classification*:

$$\max_{\theta_j} P(\theta_j|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\theta_j)P(\theta_j)}{\sum_{j=1}^{S} p(\mathbf{x}_i|\theta_j)}$$

- *Maximum Likelihood*:

$$\max_{\theta_j} p(\mathbf{x}_i|\theta_j)$$

- if $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots \mathbf{x}_{M-1}\}$ and assuming frames are independent:

$$p(\mathbf{X}|\theta_j) = \prod_{i=0}^{M-1} p(\mathbf{x}_i|\theta_j)$$

- Speaker identification:

$$\hat{S} = \max_{1 \leq j \leq S} \sum_{i=0}^{M-1} \log\left[p(\mathbf{x}_i|\theta_j)\right]$$

SPEAKER IDENTIFICATION

- If we have estimated S target speaker models θ_j with $j = 1, 2, \cdots, S$.
- *Maximum a posteriori probability classification*:

$$\max_{\theta_j} P(\theta_j|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\theta_j)P(\theta_j)}{\sum_{j=1}^{S} p(\mathbf{x}_i|\theta_j)}$$

- *Maximum Likelihood*:

$$\max_{\theta_j} p(\mathbf{x}_i|\theta_j)$$

- if $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots \mathbf{x}_{M-1}\}$ and assuming frames are independent:

$$p(\mathbf{X}|\theta_j) = \prod_{i=0}^{M-1} p(\mathbf{x}_i|\theta_j)$$

- Speaker identification:

$$\hat{S} = \max_{1 \le j \le S} \sum_{i=0}^{M-1} \log\left[p(\mathbf{x}_i|\theta_j)\right]$$

Speaker Identification

- If we have estimated S target speaker models θ_j with $j = 1, 2, \cdots, S$.
- *Maximum a posteriori probability classification*:

$$\max_{\theta_j} P(\theta_j|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\theta_j)P(\theta_j)}{\sum_{j=1}^{S} p(\mathbf{x}_i|\theta_j)}$$

- *Maximum Likelihood*:

$$\max_{\theta_j} p(\mathbf{x}_i|\theta_j)$$

- if $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots \mathbf{x}_{M-1}\}$ and assuming frames are independent:

$$p(\mathbf{X}|\theta_j) = \prod_{i=0}^{M-1} p(\mathbf{x}_i|\theta_j)$$

- Speaker identification:

$$\hat{S} = \max_{1 \leq j \leq S} \sum_{i=0}^{M-1} \log\left[p(\mathbf{x}_i|\theta_j)\right]$$

Speaker Identification

- If we have estimated S target speaker models θ_j with $j = 1, 2, \cdots, S$.
- *Maximum a posteriori probability classification*:

$$\max_{\theta_j} P(\theta_j|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\theta_j)P(\theta_j)}{\sum_{j=1}^{S} p(\mathbf{x}_i|\theta_j)}$$

- *Maximum Likelihood*:

$$\max_{\theta_j} p(\mathbf{x}_i|\theta_j)$$

- if $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots \mathbf{x}_{M-1}\}$ and assuming frames are independent:

$$p(\mathbf{X}|\theta_j) = \prod_{i=0}^{M-1} p(\mathbf{x}_i|\theta_j)$$

- Speaker identification:

$$\hat{S} = \max_{1 \le j \le S} \sum_{i=0}^{M-1} \log\left[p(\mathbf{x}_i|\theta_j)\right]$$

Speaker Verification

- If we have estimated a GMM for the target speaker θ_t and a GMM for a collection of imposters (*background model*), θ_{BUM}
- Compute the ratio:

$$\frac{P(\theta_t|\mathbf{X})}{P(\theta_{BUM}|\mathbf{X})} = \frac{p(\mathbf{X}|\theta_t)P(\theta_t)}{p(\mathbf{X}|\theta_{BUM})P(\theta_{BUM})}$$

- Compute the *log-likelihood ratio*:

$$\Lambda(\mathbf{X}) = \log\left[p(\mathbf{X}|\theta_t)\right] - \log\left[p(\mathbf{X}|\theta_{BUM})\right]$$

- Compare with a threshold

$$\begin{aligned} \Lambda(\mathbf{X}) &\geq \lambda, \quad \text{accept} \\ \Lambda(\mathbf{X}) &< \lambda, \quad \text{reject} \end{aligned}$$

Speaker Verification

- If we have estimated a GMM for the target speaker θ_t and a GMM for a collection of imposters (*background model*), θ_{BUM}
- Compute the ratio:

$$\frac{P(\theta_t|\mathbf{X})}{P(\theta_{BUM}|\mathbf{X})} = \frac{p(\mathbf{X}|\theta_t)P(\theta_t)}{p(\mathbf{X}|\theta_{BUM})P(\theta_{BUM})}$$

- Compute the *log-likelihood ratio*:

$$\Lambda(\mathbf{X}) = \log\left[p(\mathbf{X}|\theta_t)\right] - \log\left[p(\mathbf{X}|\theta_{BUM})\right]$$

- Compare with a threshold

$$\begin{aligned} \Lambda(\mathbf{X}) &\geq \lambda, \quad \text{accept} \\ \Lambda(\mathbf{X}) &< \lambda, \quad \text{reject} \end{aligned}$$

Speaker Verification

- If we have estimated a GMM for the target speaker θ_t and a GMM for a collection of imposters (*background model*), θ_{BUM}
- Compute the ratio:

$$\frac{P(\theta_t|\mathbf{X})}{P(\theta_{BUM}|\mathbf{X})} = \frac{p(\mathbf{X}|\theta_t)P(\theta_t)}{p(\mathbf{X}|\theta_{BUM})P(\theta_{BUM})}$$

- Compute the *log-likelihood ratio*:

$$\Lambda(\mathbf{X}) = \log\left[p(\mathbf{X}|\theta_t)\right] - \log\left[p(\mathbf{X}|\theta_{BUM})\right]$$

- Compare with a threshold

$$\begin{aligned} \Lambda(\mathbf{X}) &\geq \lambda, \quad \text{accept} \\ \Lambda(\mathbf{X}) &< \lambda, \quad \text{reject} \end{aligned}$$

Speaker Verification

- If we have estimated a GMM for the target speaker θ_t and a GMM for a collection of imposters (*background model*), θ_{BUM}
- Compute the ratio:

$$\frac{P(\theta_t|\mathbf{X})}{P(\theta_{BUM}|\mathbf{X})} = \frac{p(\mathbf{X}|\theta_t)P(\theta_t)}{p(\mathbf{X}|\theta_{BUM})P(\theta_{BUM})}$$

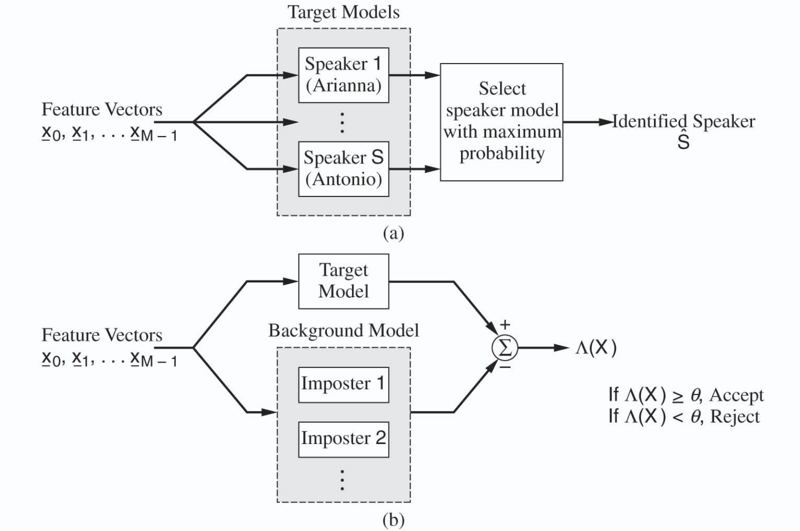
- Compute the *log-likelihood ratio*:

$$\Lambda(\mathbf{X}) = \log\left[p(\mathbf{X}|\theta_t)\right] - \log\left[p(\mathbf{X}|\theta_{BUM})\right]$$

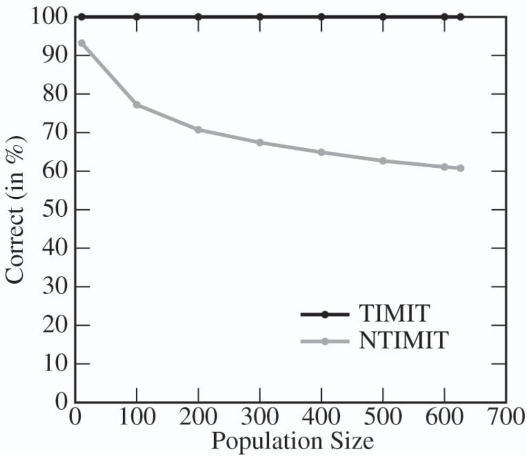
- Compare with a threshold

$$\begin{array}{lll} \Lambda(\mathbf{X}) & \geq & \lambda, \quad \text{accept} \\ \Lambda(\mathbf{X}) & < & \lambda, \quad \text{reject} \end{array}$$

GMM-based recognition systems



Performance of GMM-based recognition systems

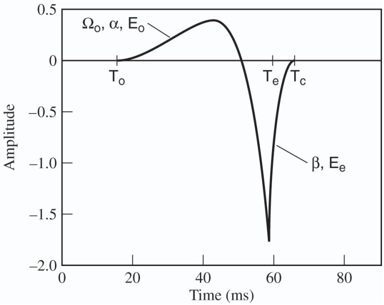


▷ 19 mel-scale coeff (24-1-2-2), 8-component GMM with diagonal covariance matrix.

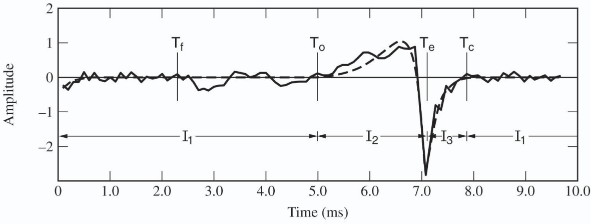
Outline

7-parameters LF model:

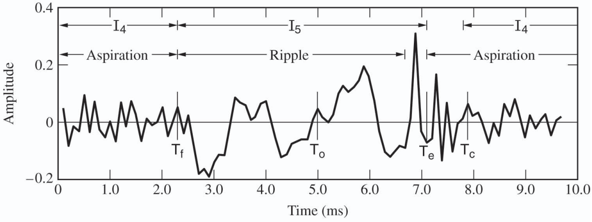
$$
\begin{aligned}
u_{LF}(t) &= 0, && 0 \leq t < T_o \\
&= E_o e^{\alpha(t-T_o)} \sin\left[\Omega_0(t - T_o)\right], && T_o \leq t < T_e \\
&= -E_1\left[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}\right], && T_e \leq t < T_c
\end{aligned}
$$



Example of a glottal flow derivative estimate [1]

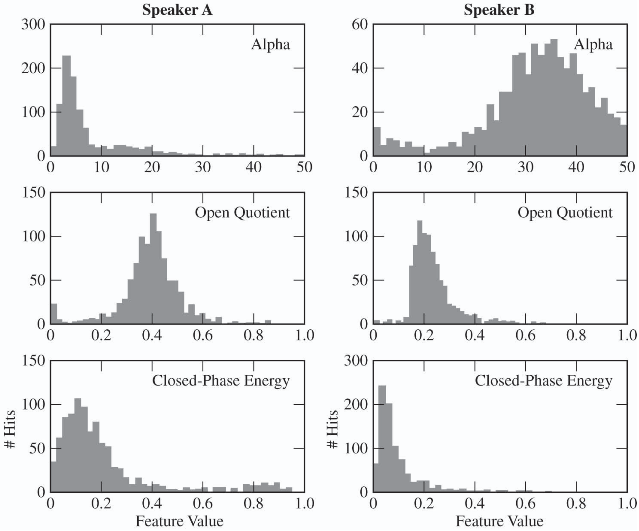


(a)



(b)

Comparing histograms for two speakers based on GFD estimates [1]



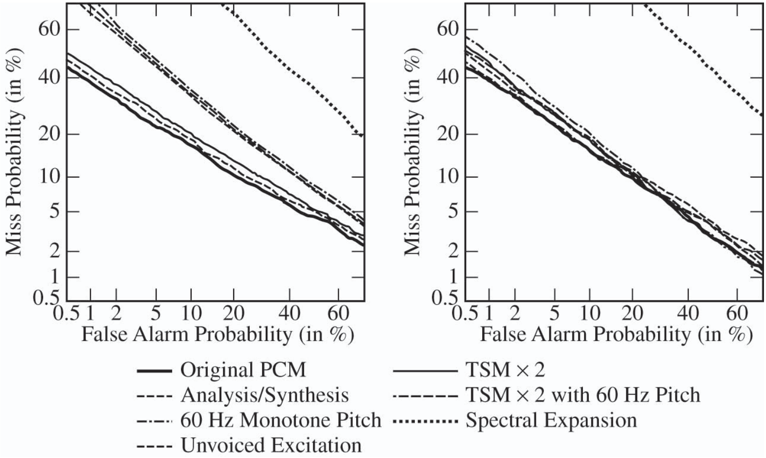
Speaker identification performance using GFD parameters [1]

Table: Using GFD estimates

Features		Male	Female
Coarse:	7 LF	58.3%	68.2%
Fine:	5 energy	39.5%	41.8%
Source:	12 LF & energy	69.1%	73.6%

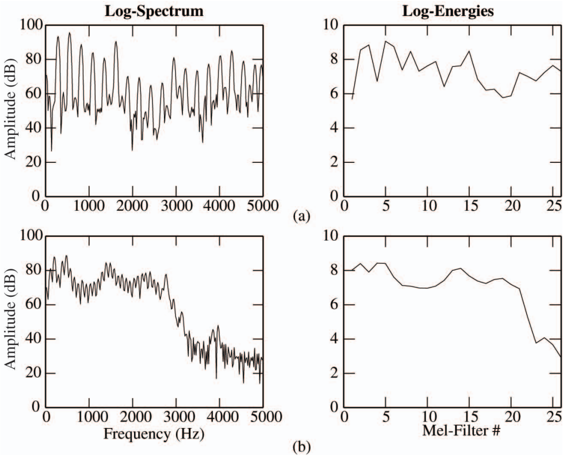
Table: Using mel-cepstrum on GFD estimates

Features	Male	Female
Modeled GFD:	41.1%	51.8%
GFD:	95.1%	95.5%



▷ Left: females, Right: males

Explaining the performance of prosody for SID



Outline

Acknowledgments

Most, if not all, figures in this lecture are coming from the book:

T. F. Quatieri: Discrete-Time Speech Signal Processing,
principles and practice
2002, Prentice Hall

and have been used after permission from Prentice Hall

Outline

M. Plumpe, T. F. Quatieri, and D. Reynolds, “Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 1, pp. 569–586, Sept. 1999.