Feature Selection in Speaker Verification Systems

Arnon Cohen and Yaniv Zigel
Electrical and Computer Eng. Dept.
Ben-Gurion University
Beer-Sheva, Israel
Speaker Recognition

- Speaker Identification (closed and open sets)
- Speaker Verification
- Speaker Spotting
- Group Recognition  Gender, Accent, Age,……
- Speaker Tracking

- Text Dependent / Text Independent
- Supervised / Unsupervised
Applications

• Commercial
  – Access Control (supervised verification, Text depen./Indep.))
  – Segmentation   (Unsupervised, Text Independent)
  – ................

• Military
  – Spotting
  – Tracking

• Forensic
  – Identification, Verification, Supervised, Unsupervised
Speaker Recognition - Constraints

• **Natural constraints:**

  • Relatively small amount of information on speaker’s identity in the speech signal (as compare to the message).

  • Features change in time (sensitivity to the train-test time difference).

  • The Human voice is highly sensitive to:

    • speaker’s physiologic and psychological states,

    • environment conditions (heat, coldness),
**Speaker Recognition - Constraints**

- **Environmental constraints:**
  - different SNRs,
  - channel transmission (telephone, internet),
  - microphone transmission (different handsets, distances, angles),

- **System’s and Application Constraints**
  - Size of training database
  - Duration of test utterance
  - Time between Training and Testing
  - Memory, MIPS.
Supervised Speaker Recognition

**Training (Enrollment)**

- Utterance
- Data Acquisition
- Feature Extraction
- Model Parameters Estimation
- Model of ith person
- Cohorts Selection Policy

**Recognition**

- Utterance
- Data Acquisition
- Feature Extraction
- Matching Strategy
- Rejection Strategy

Targets & Impostors
Speaker Verification

\( T - \text{Target} \quad \rightarrow \quad \lambda_T \text{ model of Target Speaker} \)

\( I - \text{Impostor} \quad \rightarrow \quad \lambda_I \text{ model of Impostor Speaker} \)

\( X - \text{Tested Speaker} \)

\( C - \text{Cohort} \quad \rightarrow \quad \lambda_c \text{ model of Cohort Speaker} \)

\( O - \text{Observation} \quad \text{Sequence of feature vectors, corresponding to the utterance} \)

\[ \log\left\{ P(O_x | \lambda_T) \right\} - \{ f\left(P(O_x | \lambda_C)\right) \right\} \]

\( \begin{cases} 
> \xi & \text{accept} \\
\leq \xi; > \psi & \text{repeat} \\
\leq \psi & \text{reject}
\end{cases} \)

Score Normalization (Dynamic Threshold)
Open Questions

- What (and how many) features to use?
- What recognition engine to use? (*Type, Architecture, order*)
- Rejection policy – Cohorts? How many, how to choose?
- What normalization scheme should be used?
- How to deal with the channel problem? (*robust features, normalization, adaptation…*)
- How to deal with the noise problem?
- How to deal with variations in speaker characteristics (*time, physiology, psychology*)
- How to deal with mimicry and falsification?
Features for Speaker Recognition

- Efficient in representing the speaker dependent information.
- Easy to measure
- Stable over time
- Occur naturally and frequently in speech
- Change little from one speaking environment to another
- Insensitive to mimicry and falsification
- Insensitive to noise and bandwidth limitations
- Insensitive to speaking state

*Such features do not exist!*
Features used in Speaker Recognition Systems

- **Vocal Tract model features**
  - Autocorrelation coefficients (COR)
  - Linear Prediction Coefficients (LPC)
  - Partial Correlation coefficients (PARCOR)
  - Log Area Ratio coefficients (LAR)
  - Perceptual Linear Prediction (PLP)

- **Spectral and Cepstral features**
  - Line Spectrum Pairs (LSP)
  - Bank of filters (linear)
  - Bank of filters (Mel)
  - Mel Frequency Cepstral Coefficients (MFCC)

### Static Features ↔ Dynamic Features

\[ \Delta, \Delta \Delta \]
Other features:

- **Prosodic**
  - Pitch contours
  - Intonation
  - Stress

- **Phonetic (pronunciation)** [Andrews, Kohler & Campbell]
  - Phone features
    (requires large databases and Phone recognition system)

- **Idiolectal Features** [Doddington]
  - Unigrams - probability of word occurrence
  - Bigrams - probability of pairs of words
    (requires large databases and speech recognition systems)
The “Curse” of Dimensionality

Recognition Error

Expected behavior

Curse of dimensionality

Number of Features

Best dimension region
Recognition Engines

- Aural & spectrogram Matching
- Template Matching (DTW, VQ)
- HMM, GMM, ANN
- SVM?, Hybrid?, ????

Small Databases
- Cleaned, controlled speech

Large Databases
- Realistic, Unconstraint speech
Score Normalization

\[ s_1(O) = \log p(O | \lambda_T) \frac{\max_{c \in C(T)} \log p(O | \lambda_c)}{\log} \]

\[ s_4(O) = \log p(O | \lambda_T) - \frac{1}{C} \sum_{c=1}^{C} \log p(O | \lambda_c) \]

\[ s_2(O) = \log p(O | \lambda_T) - \max_{c \in C(T)} \log p(O | \lambda_c) \]

\[ s_5(O) = \frac{\log p(O | \lambda_T)}{\frac{1}{C} \sum_{c=1}^{C} \log p(O | \lambda_c)} \]

\[ s_3(O) = \log p(O | \lambda_T) - \log \left\{ \frac{1}{C} \sum_{c=1}^{C} p(O | \lambda_c) \right\} \]

\[ s_6(O) = \frac{\log p(O | \lambda_T)}{\log \left\{ \frac{1}{C} \sum_{c=1}^{C} p(O | \lambda_c) \right\}} \]
Detection Error Tradeoff (DET) curve
State of the Art

Text-Independent (read sentences)
Military radio, Multiple Radios & microphones
Moderate amount of training

Text-Independent (conversation)
Telephone data, Multiple Mics
Moderate amount of Training

Text-depended
Clean data
Single Microphone
Large amount of train/test data

Text-Depended
(Digits strings)
Telephone data, Multiple mics
Small amount of training data

Increased Constraints
25%

10%

1%

0.1%
Feature Selection

Common Feature Space

Speaker 1 - Feature Space

Speaker 2 - Feature Space
Feature Selection Problem

The method for feature selection can be specified in terms of two components:

I A performance criterion (effective criterion, discriminant function) for the selection of features from the input feature set.

II Selection procedure.

The problem of feature selection can be described as follows:

Given a set $y$ of $K$ features $y = \{y_i | i = 1, 2, \ldots, K\}$
select a subset $x$ of $k$ features ($k < K$) $x = \{x_i | i = 1, 2, \ldots, k, x_i \in y\}$
such that a criterion $J()$ is optimized.
Performance Criteria

F-ratio

\[ F = \frac{\text{variance of inter-speaker feature mean}}{\text{mean of intra-speaker feature variance}} \]

Scatter Matrices and Separability Criteria

\[ J_1 = \text{tr} \mathbf{S}_2^{-1} \mathbf{S}_1 \]
\[ J_2 = \ln |\mathbf{S}_2^{-1} \mathbf{S}_1| \]
\[ J_3 = \frac{\text{tr} \mathbf{S}_1}{\text{tr} \mathbf{S}_2} \]

Bhattacharyya Distance

\[ d_B = \frac{1}{2} \ln \left( \frac{\mathbf{W}_i + \mathbf{W}_j}{2 \sqrt{\|\mathbf{W}_i\| \|\mathbf{W}_j\|}} \right) + \frac{1}{8} (m_i - m_j)^T \left( \frac{\mathbf{W}_i + \mathbf{W}_j}{2} \right)^{-1} (m_i - m_j) \]

Bhattacharyya Shape

\[ d_{BS} = \ln \left( \frac{\mathbf{W}_i + \mathbf{W}_j}{2 \sqrt{\|\mathbf{W}_i\| \|\mathbf{W}_j\|}} \right) \]

Divergence Distance

\[ d_D = E \left[ \ln \frac{p (\omega, \mathbf{y}| \omega_j)}{p (\omega, \mathbf{y}| \omega_i)} \right] \]

Divergence Shape

\[ d_{Ds} = \text{tr} \left[ \mathbf{v}_j - \mathbf{W}_j \mathbf{v}_j^{-1} - \mathbf{W}_i^{-1} \right] \]
Performance Criteria

- **EER(O)**: Equal Error Rate  \[ FA = FR \]
- **E_{GM}(O)**: Geometric Mean Error  \[ E_{GM} = \sqrt{E_{FR}E_{FA}} \]
- **DCF(O)**: Decision Cost Function  \[ DCF(O) = \rho \Pr_{FR|O} + (1-\rho) \Pr_{FA|O} \]

\[ \rho = \text{desired weight of FR}; \quad 0 \leq \rho \leq 1 \]
9. Feature Selection

Feature optimization by a cost function
Feature Selection Problem

Feature Subset Selection Methods

- K-best Method
- Backward Selection
- Forward Selection
- The l-r Algorithm
- Sequential Floating Forward Search (SFFS)
- Sequential Floating Backwards Search (SFBS)
- Branch-and-Bound (BB)
- Dynamic Programming (DP)
- Exhaustive Search
- Random Walk
- Genetic Algorithms (GA)

Problem: \( \frac{K!}{k!(\alpha-k)!} \) searches
Performance criterion for Speaker Verification

In verification systems, the decision to accept or reject an identity claim is based on the comparison of a score with a threshold.

the score: \[ s(\text{O}) = \log p(\text{O} | \lambda_T) \]

- **O** - observations (from utterances)
- **\(\lambda_T\)** - target’s model
- **\(\tau\)** - the threshold
- **\(\text{O}_T\)** - target’s observations
- **\(\text{O}_I\)** - imposters’ observations
Performance criterion for Speaker Verification

- Evaluation of speaker verification systems: *Equal Error Rate (EER)*.

- EER as a criterion for feature selection – impractical:
  - Cost of computation,
  - Low resolution – Due to rough histograms.

- Solution – *Gaussian assumption*

\[
f[s(O)|O \in O_T] = \frac{1}{\sqrt{2\pi \sigma_T}} \exp \left[-\frac{(s(O) - \mu_T)^2}{2\sigma_T^2}\right]
\]
**Criterion for minimizing EER**

\[ P_{\text{miss}} = \int_{-\infty}^{T} f[s(O) | O \in O_T] \, ds = \int_{-\infty}^{T} \frac{1}{\sqrt{2\pi} \sigma_T} \exp \left[ -\frac{1}{2} \left( \frac{s - \mu_T}{\sigma_T} \right)^2 \right] \, ds = \text{erf} \left( \frac{T - \mu_T}{\sigma_T} \right) + \frac{1}{2} \]

\[ P_{\text{FA}} = \int_{T}^{\infty} f[s(O) | O \in O_I] \, ds = \int_{T}^{\infty} \frac{1}{\sqrt{2\pi} \sigma_I} \exp \left[ -\frac{1}{2} \left( \frac{s - \mu_I}{\sigma_I} \right)^2 \right] \, ds = -\text{erf} \left( \frac{T - \mu_I}{\sigma_I} \right) + \frac{1}{2} \]

\[ EER = P_{\text{miss}} = P_{\text{FA}} \Rightarrow \]

\[ \tau = \frac{\mu_I \sigma_T + \mu_T \sigma_I}{\sigma_I + \sigma_T} \]

\[ EER = \text{erf} \left( \frac{\mu_I - \mu_T}{\sigma_I + \sigma_T} \right) + \frac{1}{2} \]

\[ EER' = \frac{\mu_T - \mu_I}{\sigma_I + \sigma_T} \]
Experimental Setup

- The experiment was set for:
  - Text-dependent
  - Speaker-verification
  - CD-HMM: 5 states ; 2 Gaussians per state.

- The database
  - Hebrew word /hamesh/ (five) - HID database
  - High quality speech; sampled at 16KHz with 12 bits resolution
  - Three target speakers
  - 19 imposters for each target
  - Number of repetitions: - training: 20
                             - testing: 25 - 79
## Experimental Setup

### The features and their symbols

<table>
<thead>
<tr>
<th>#</th>
<th>Feature name</th>
<th>Order</th>
<th>Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mel Frequency Cepstral Coef. (MFCC)</td>
<td>12</td>
<td>( m_1 \div m_{12} )</td>
</tr>
<tr>
<td>2</td>
<td>Linear Prediction Cepstral Coef. (LPCC)</td>
<td>12</td>
<td>( c_1 \div c_{12} )</td>
</tr>
<tr>
<td>3</td>
<td>Log Area Ratio (LAR)</td>
<td>12</td>
<td>( a_1 \div a_{12} )</td>
</tr>
<tr>
<td>4</td>
<td>Linear Prediction Coef. (LPC)</td>
<td>12</td>
<td>( l_1 \div l_{12} )</td>
</tr>
<tr>
<td>5</td>
<td>Partial Correlation (PARCOR)</td>
<td>12</td>
<td>( p_1 \div p_{12} )</td>
</tr>
<tr>
<td>6</td>
<td>First diff of MFCC (( \Delta )-MFCC)</td>
<td>12</td>
<td>( \Delta m_1 \div \Delta m_{12} )</td>
</tr>
<tr>
<td>7</td>
<td>First diff of LPCC (( \Delta )-LPCC)</td>
<td>12</td>
<td>( \Delta c_1 \div \Delta c_{12} )</td>
</tr>
<tr>
<td>8</td>
<td>First diff of LAR (( \Delta )-LAR)</td>
<td>12</td>
<td>( \Delta a_1 \div \Delta a_{12} )</td>
</tr>
<tr>
<td>9</td>
<td>First diff of LPC (( \Delta )-LPC)</td>
<td>12</td>
<td>( \Delta l_1 \div \Delta l_{12} )</td>
</tr>
<tr>
<td>10</td>
<td>First diff of PARCOR (( \Delta )-PARCOR)</td>
<td>12</td>
<td>( \Delta p_1 \div \Delta p_{12} )</td>
</tr>
</tbody>
</table>

Total number of features: 120
Results

EER test results of the different selection methods in different feature space dimension (for speaker #1)
## Results

- Feature selection methods:  
  - Dynamic Programming  
  - Forward  
  - k – best

- Different selected feature set – for different speakers

- The Selected Features (Example for DP - 24 space dimension)

<table>
<thead>
<tr>
<th>Sp #</th>
<th>Selection method</th>
<th>Selected features</th>
</tr>
</thead>
</table>
| 1    | Dynamic Programming  
|      |                  | $m_2 m_3 m_4 m_8 m_9 m_{11} m_{12}$  
|      |                  | $a_5 a_{11} l_{11} l_{12} p_7 p_8 p_{11}$  
|      |                  | $\Delta m_1 \Delta m_5 \Delta m_8 \Delta m_{10}$  
|      |                  | $\Delta a_1 \Delta a_3 \Delta l_{10} \Delta p_3 \Delta p_6 \Delta p_{11}$  |
| 2    |                  | $m_8 m_{11} m_{12}$  
|      |                  | $a_8 a_{11} l_{11} l_{12} p_{11}$  
|      |                  | $\Delta m_2 \Delta m_5 \Delta m_8 \Delta m_{11} \Delta m_{12}$  
|      |                  | $\Delta a_5 \Delta a_9 \Delta a_{11} \Delta l_{11} \Delta p_3 \Delta p_5 \Delta p_8 \Delta p_9 \Delta p_{11}$  |
| 3    |                  | $m_4 m_7 m_9 m_{11}$  
|      |                  | $a_9 l_9 p_9$  
|      |                  | $\Delta m_1 \Delta m_4 \Delta m_8 \Delta m_9 \Delta m_{11} \Delta m_{12}$  
|      |                  | $\Delta c_1 \Delta a_1 \Delta a_2 \Delta a_5 \Delta l_1 \Delta l_3 \Delta l_4 \Delta l_8 \Delta p_1 \Delta p_5$  |
Results

Average DET curves of speaker verification results

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC space</td>
<td>4.78%</td>
</tr>
<tr>
<td>DP space</td>
<td>2.71%</td>
</tr>
</tbody>
</table>

Improvement of 43.3%
Conclusions – Where do we go from here?

- Current technology achieves accuracies on the order of 10% (EER) for realistic telephone quality, with HMM/GMM.

- Feature selection has a potential for increase in accuracy.

- Efficient cohort selection – a potential for increase in accuracy.

- A breakthrough is needed –

  SVM? , Hybrid? , new improved speech model? New features?
12. Unsupervised Recognition (Segmentation)

Example: Telephone conversation (two speakers)

Diagram:
- Input: Simultaneous Speech Detection
  - λ_{in}
  - λ_{real}

- Detection of Silence
  - λ_i

- Initial Condition

- Segmented data
  - A^0
  - B^0
  - S^0

- Train
- Closed Set Identification
- Convergence Test
  - λ_a
  - λ_s
  - λ_r

- New models