Integration and Imputation Methods for Unreliable Feature Compensation in GMM Based Speaker Verification

Andrzej Drygajlo and Mounir El-Maliki
andrzej.drygajlo@epfl.ch

Speech Processing Group (GTP)
Signal Processing Laboratory (LTS)
Swiss Federal Institute of Technology Lausanne (EPFL)
Table of Contents

• Robust Speaker Verification
• Speech Enhancement
• Missing Feature Detection
• Missing Feature Handling Approaches
• Performance Evaluation
• Conclusions and Possible Extensions
Context

Speaker Recognition

- Speaker Identification: Identify a speaker from N others
- Speaker verification: Determine if the claimed identity corresponds to the speaker of the utterance
  - Text-dependent
  - Text-prompted
  - Text-independent

Robust Speaker Verification

Real-life applications
Noisy environments

Performance degradation of speaker verification

- Robust feature extraction and modelling
- Speech enhancement in a pre-processing stage
- Missing (non-reliable, masked by noise) data detection and compensation
Human recognition with missing data

- The missing data condition occurs naturally. (e.g. interfering signals, band-restricted transmission, channel noise over telephone lines, ambient noise, …).

- Masked data is effectively missing data. Locally weaker sound components do not contribute to the neuronal output: they are masked and can be considered missing for the purposes of further processing.

- Redundancy in speech combats the missing data problem.
Objectives

- Robust speaker verification in the presence of additive noise
- Finding criteria for detecting missing features
- Handling missing features during the recognition process

Proposed Approach

Integration of three concepts into speaker verification system
Speaker Verification System: Overview

**Speaker verification**: Determines if the claimed identity corresponds to the speaker of the utterance.

**Text-independent**: No constraints on the uttered speech.

---

Model Training

- Input Speech
- Speech pre-processing: MFCC, Filter-bank
- Feature extraction
- Distance/Likelihood
- Speaker Modelling
- Decision: Accept, Reject
# Gaussian Mixture Model (GMM)

**Speaker Modelling**
- DTW template matching
- VQ
- GMM
- HMM

The GMM pdf is defined as:

$$p(x | \lambda) = \sum_{i=1}^{M} p_i \Phi(x, \mu_i, \Sigma_i)$$

where

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, \quad i = 1, \Lambda, M$$

$$\Phi(x) = \frac{\exp\{-\frac{1}{2}(x - \mu_i)^T(S_i)^{-1}(x - \mu_i)^T\}}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}$$

**Decision**

$$\frac{p(x | \lambda)}{p(x | \overline{\lambda})} > \text{Threshold} \quad \text{Accept}$$

where $\overline{\lambda}$ is the world model
Speech Enhancement

Spectral Subtraction

Speech corrupted by stationary additive noise:

\[ y(n) = s(n) + n(n) \]

Spectral estimate of the clean speech

\[ |\hat{S}_m(\omega)|^\gamma = \begin{cases} |Y_m(\omega)|^\gamma - |\overline{N}(\omega)|^\gamma & \text{if} \quad |Y_m(\omega)|^\gamma > |\overline{N}(\omega)|^\gamma \\ 0 & \text{otherwise} \end{cases} \]

\[ \gamma = 1 : \text{Magnitude spectral subtraction} \]

\[ \gamma = 2 : \text{Power spectral subtraction} \]

Estimated clean speech

\[ \hat{s}(n) = IDFT[|\hat{S}(\omega)| \cdot e^{(j \arg Y(\omega))}] \]

Detection of missing features

\[ |\hat{S}_m(\omega)|^\gamma = \begin{cases} |Y_m(\omega)|^\gamma - |\overline{N}(\omega)|^\gamma & \text{if} \quad |Y_m(\omega)|^\gamma > |\overline{N}(\omega)|^\gamma \\ \text{Missing feature} & \text{otherwise} \end{cases} \]
Missing feature detection

Generalized Spectral Subtraction: $\gamma=2$ : Power spectrum subtraction

$$D_m(\omega) = |Y_m(\omega)|^2 - \alpha \cdot |\overline{N}(\omega)|^2$$

$$|\hat{S}_m(\omega)|^2 = \begin{cases} D_m(\omega) & \text{if } D_m(\omega) > \beta \cdot |\overline{N}(\omega)|^2 \\ \beta \cdot |\overline{N}(\omega)|^2 & \text{otherwise} \end{cases} \quad \alpha \geq 1 \quad \text{overestimation} \quad 0 < \beta \ll 1 \quad \text{spectral flooring}$$

Spectral components below spectral floor non-reliable (missing)

$$|\hat{S}_m(\omega)|^2 = \begin{cases} D_m(\omega) & \text{if } D_m(\omega) > \beta \cdot |\overline{N}(\omega)|^2 \\ \text{Missing feature} & \text{otherwise} \end{cases}$$

- $\alpha = 1, \beta = 0$ : spectral subtraction
- $\alpha = \alpha_m(\omega), \beta = \beta_m(\omega)$ : Adaptive generalized spectral subtraction
Missing feature detection

The detection threshold of non-reliable spectral features

\[ \frac{|Y_m(\omega)|^\gamma}{|\hat{N}_m(\omega)|^\gamma} \leq (\alpha + \beta) \]

**Power Spectral Subtraction**

Features are non-reliable if

- \( \alpha = 1, \beta = 0 \)
- \( \gamma = 2 \)

\[ SNR_{post_m}(\omega) = \frac{|Y_m(\omega)|^2}{|\hat{N}_m(\omega)|^2} \leq 1 \]

**Generalized Spectral Subtraction**

Features are non-reliable if

- \( \alpha > 1, \beta << 1 \)
- \( \gamma = 2 \)

\[ SNR_{post_m}(\omega) = \frac{|Y_m(\omega)|^2}{|\hat{N}_m(\omega)|^2} \leq (\alpha + \beta) \]

**SNR criterion** (Cooke et al. 99)

Features are non-reliable if

- \( \alpha + \beta = 3.41 \)
- \( \gamma = 1 \)

\[ SNR_{prio} = \frac{|\hat{S}_m(\omega)|^2}{|\hat{N}_m(\omega)|^2} \leq 1 \]

- \( |\hat{S}_m(\omega)|^2 < \frac{1}{2} |Y_m(\omega)|^2 \)
Minimum Mean-Square Error Spectral Amplitude Estimator

Spectral amplitude estimation of the clean speech signal by minimizing the mean square error

\[ G(\omega) = \frac{1}{2} \sqrt{\pi} \frac{\sqrt{\nu_{\omega}}}{\text{SNR}_{\text{post}_{\omega}}} \exp\left( -\frac{\nu_{\omega}}{2} \right) \left[ (1 + \nu_{\omega})I_0\left(\frac{\nu_{\omega}}{2}\right) + \nu_{\omega}I_1\left(\frac{\nu_{\omega}}{2}\right) \right] \]

\[ G_{\omega}(m) = F[\text{SNR}_{\text{prio}_{\omega}}(m), \text{SNR}_{\text{post}_{\omega}}(m)] \]

where

\[ \nu_{\omega} = \frac{\text{SNR}_{\text{prio}_{\omega}}}{1 + \text{SNR}_{\text{prio}_{\omega}}} \frac{\text{SNR}_{\text{post}_{\omega}}}{} \]

Prior SNR

\[ \text{SNR}_{\text{prio}_{\omega}}(m) = \frac{E\{|S_{\omega}(m)|^2\}}{|N_{\omega}(m)|^2} \]

Posterior SNR

\[ \text{SNR}_{\text{post}_{\omega}}(m) = \frac{|Y_{\omega}(m)|^2}{|N_{\omega}(m)|^2} \]

\[ I_0 = 0\text{th order Bessel function} \]

\[ I_1 = 1\text{th order Bessel function} \]
A priori SNR Estimation by the Maximum Likelihood Approach

- The estimator of the a priori SNR is obtained in practice by a recursive averaging:

\[
SNR_{prio,\omega}(m) = \begin{cases} 
SNR_{post,\omega}(m) - 1 & \text{if non-negative} \\
0 & \text{otherwise}
\end{cases}
\]

where

\[
SNR_{post,\omega}(m) = p \cdot SNR_{post,\omega}(m-1) + (1 - p) \frac{SNR_{post,\omega}(m)}{\alpha}
\]

\[0 < p < 1\]

Features are non-reliable if

\[
SNR_{post,\omega}(m) \leq 1 \quad \alpha > 1
\]

\[\alpha = 1 \text{ and } p = 0\] Spectral Subtraction
Missing Feature Detection

• Missing Feature Proportions

Example: 0 dB signal-to-noise ratio

White Gaussian noise

\[ \text{GSS: } \alpha = 3, \beta = 0.001 \]
\[ \text{SS: } \alpha = 1, \beta = 0 \]

F-16 cockpit noise

\[ \text{MMSE: } \alpha = 1.5, p = 0.75 \]

Speech-like noise

MMSE: \[ \alpha = 1.5, p = 0.75 \]
**Missing Feature Handling Approaches**

**Missing Feature Techniques**

- Learning
  - EM
  - Data Imputation

**Integration**

- Marginal Distribution
  - Bounded Integration
    - Conditional
    - Unconditional

**Data Imputation**

- Wiener filtering dependent on the most probable Gaussian pdf

**Recognition**

- Mean Imputation

**Integrated Speech-Background Model**
Missing Feature Handling Approaches

• Integration Methods

In the presence of incomplete set of features, $p(x|\lambda)$ could not be directly evaluated

Feature

$x = (x^p, x^m)$

$x^p$ present features

$x^m$ missing features

Parameters

$\mu = (\mu^p, \mu^m)$

$\Sigma = \begin{bmatrix} 
\Sigma^{pp} & \Sigma^{pm} \\
\Sigma^{mp} & \Sigma^{mm} 
\end{bmatrix}$

1. Bounded Integration Method

Under the knowledge of the missing feature space:

$p(x | \lambda) \approx p(x^p | \lambda) \int_{x_l}^{x_u} p(x^m | x^p, \lambda)dx^m$

$x_u$ upper bound

$x_l$ lower bound

2. Marginal Distribution

$p(x | \lambda) \approx p(x^p | \lambda)$

$x_u + \infty$

$x_l - \infty$
Marginal Distribution in GMMs

Likelihood function

\[ p(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} p_i \Phi(\mathbf{x}, \mu_i, \Sigma_i) \]

In the case of diagonal covariance matrices:

\[ p(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D} \Phi(x_j, m_{ji}, \sigma_{ji}^2) \]

In the case of missing features:

\[ p(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} p_i \prod_{j}^{\text{present}} \Phi(x_j, m_{ji}, \sigma_{ji}^2) \prod_{j}^{\text{missing}} \Phi(x_j, m_{ji}, \sigma_{ji}^2) \]

Modified likelihood function becomes:

\[ p(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} p_i \prod_{j}^{\text{present}} \Phi(x_j, m_{ji}, \sigma_{ji}^2) \quad \text{Easy to implement} \]
Missing Feature Handling Approaches

• Imputation:

Replacement of non-reliable features by estimated values

• Mean

• Conditional mean

Integrated speech-background model

Novel method  Wiener filtering+MFCC

1. Mean Imputation

\[ \mathbf{x}^m = \mu \]

2. Conditional Mean Imputation

\[
\mathbf{x}^m = E\{\mathbf{x}^m | \mathbf{x}^p, \lambda\} = \mu^m + \Sigma^{mp} (\mathbf{x}^p - \mu^p) \Sigma^{pp,-1} \]

\[
\Sigma^{mm} = \Sigma^{mm} + \Sigma^{mp} \Sigma^{pp,-1} \Sigma^{pm}
\]
3. Integrated speech-background model (Rose et al. 94)

1. ML decoding
   - p(noisy speech | models of noise and clean speech)

2. Parameter estimation
   - Clean speech model

3. Speech enhancement
   - Clean speech estimation
Missing Feature Handling Approaches

Max noise model

\[ Y^l = \log(S + N) \approx \max(S^l, N^l) \]

where \( Y^l = \log(Y) \)

and \( Y \) is the energy in a sub-band

\[
E\{S^l | Y^l, i, \lambda\} = p(S^l = Y^l | i, \lambda) \cdot Y^l + (1 - p(S^l = Y^l | i, \lambda)) \cdot E\{S^l | S^l < Y^l, i, \lambda\}
\]

Probability that the noisy sub-band represents clean speech and not noise

Estimation of speech sub-band masked by noise

Can be used to estimate speech sub-bands masked by noise

Swiss Federal Institute of Technology, Lausanne
Signal Processing Laboratory
4. Proposed method of imputation:
Wiener filtering dependent on the most probable Gaussian pdf

1. Selection of the most probable Gaussian pdf

\[ \Phi = \operatorname{argmax}_i (p(x^p | \mu_i, \Sigma_i)) \]

2. Wiener filtering

\[ x^m = \frac{\mu_{\text{max}}}{\mu_{\text{max}} + |N|^2} |Y|^2 \]

Under the assumption of noise and speech additivity in the power spectrum domain

1. \[ |Y|^2 = |S|^2 + |N|^2 = \mu_{\text{max}} + |N|^2 \]

2. \[ |Y|^2 = |S|^2 + |N|^2 > \mu_{\text{max}} + |N|^2 \]

Correction factor

\[ \frac{|Y|^2}{\mu_{\text{max}} + |N|^2} \]
Experimental Results

• Protocol of Test

Database

**NTIMIT**: telephone speech quality
400 speakers
10 utterances for each speaker

**Speaker models**: each built with 24 seconds (8 utterances)
**Test**: 2 utterances, each consists of 3 seconds

Text-independent speaker verification

Noisy Environment Simulation

Noise added at different *a priori* signal-to-noise (SNR) ratios:
White Gaussian, F16 cockpit and speech-like noises

Score Normalization

**World Model**: 100 other speakers from NTIMIT corpus
Experimental Results

Gaussian Mixture Model of Speaker

Speakers: 32 Gaussian pdfs
World model: 256 Gaussian pdfs

Feature Extraction

FFT-based filter-bank: Log-energies of 14 critical bands
Hanning window: 32 ms, 50% overlap
Mel Frequency Cepstral Coefficients (MFCC): 14 coefficients

Noise Detection
Spectral derivative + adaptive threshold
Speech Enhancement in a Pre-Processing Stage

MFCCs

Clean conditions: Diagonal covariance matrix: EER = 7.5 %

- White Gaussian noise
- F16 cockpit noise
- Speech-like noise

• Noise overestimation can lead to the worst performance
• Spectral subtraction yields the best performance
Speech Enhancement in a Pre-Processing Stage

Critical bands

Clean conditions:
• Diagonal covariance matrix: EER = 12.2 %
• Full covariance matrix: EER = 9.5 %

- For band-limited noises, spectral subtraction yields usually the best performance
Experimental Results

- Marginal Distribution Technique: critical bands

- Spectral subtraction leads to the weakest performance
- MMSE spectral amplitude estimator yields the best performance

MMSE: $\alpha = 1.5$, $p = 0.5$
Experimental Results

• Integration Method: critical bands+MMSE

- White Gaussian noise
- F16 cockpit noise
- Speech-like noise

  • I: Diagonal Covariance Matrix
  • II: Full Covariance Matrix

  • Bounded Integration: from - infinity to log(noise energy)

Marginal technique with full covariance matrix leads to the best performance
Experimental Results

- **Imputation Techniques**: critical bands + MMSE

- **Diagonal covariance matrices**

---

**White Gaussian noise**  
**F16 cockpit noise**  
**Speech-like noise**
Experimental Results

- **Imputation Techniques**

  - White Gaussian noise
  - F16 cockpit noise
  - Speech-like noise

  - I: Diagonal Covariance Matrix
  - II: Full Covariance Matrix

  No improvement by imputation techniques
Experimental Results

- **Imputation Techniques**

  Wiener filtering dependent on the most probable Gaussian pdf

  - Noisy features
  - Missing feature detection
  - Reliable features
  - Max. Gaussian selection
  - Wiener filtering
  - Log to MFCC
  - Complete set of features

  - Noise estimate
  - Imputation

  - Speech enhancement
  - Reliable features
  - MFCC to Log
  - Speaker model (MFCC)
Experimental Results

- A Novel imputation Technique
  Wiener filtering dependent on the most probable Gaussian pdf

- Marginal technique with full covariance matrix requires high computational load
- Proposed approach performs well at high SNRs, requires diagonal covariance matrix (lower computational load) and leads to the same performance as the reference method in clean conditions
## Amelioration

<table>
<thead>
<tr>
<th>Source</th>
<th>EER amelioration in percent comparing to the reference system SS+MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>White Gaussian noise</strong></td>
<td><img src="#" alt="Table of EER amelioration" /></td>
</tr>
<tr>
<td>• Marginal</td>
<td>38% 32% 6% 2%</td>
</tr>
<tr>
<td>• Imputation</td>
<td>14% 18% 10% 13%</td>
</tr>
<tr>
<td></td>
<td>0 6 12 18 dB</td>
</tr>
<tr>
<td><strong>F16 cockpit noise</strong></td>
<td><img src="#" alt="Table of EER amelioration" /></td>
</tr>
<tr>
<td>• Marginal</td>
<td>30% 28% 22% 15%</td>
</tr>
<tr>
<td>• Imputation</td>
<td>13% 15% 20% 13%</td>
</tr>
<tr>
<td></td>
<td>0 6 12 18 dB</td>
</tr>
<tr>
<td><strong>Speech-like noise</strong></td>
<td><img src="#" alt="Table of EER amelioration" /></td>
</tr>
<tr>
<td>• Marginal</td>
<td>25% 20% 10% 1%</td>
</tr>
<tr>
<td>• Imputation</td>
<td>10% 13% 14% 5%</td>
</tr>
<tr>
<td></td>
<td>0 6 12 18 dB</td>
</tr>
</tbody>
</table>
Conclusions

• Non-reliable features should be treated in speaker recognition in noisy conditions
• Combination of speech enhancement, missing feature detection and compensation improves speaker verification
• Ignoring missing features increases verification performance
• Wiener filtering based imputation technique + MFCC increases verification performance and decreases computational load