Probabilistic Feature Transformation for Channel Robust Speaker Verification

Man Wai MAK and Kwok Kwong YIU
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
Outline

• What is Speaker Verification
• Challenge in Speaker Verification and How to Deal With it
• Feature Transformation
  • Blind Methods
  • Non-Blind Methods
• Evaluations on NIST2000 Dataset
• Results and Conclusion
What is Speaker Verification?

• To verify the identity of a claimant based on his/her own voices

I am Mary

Is this Mary’s voice?
Challenges in Speaker Verification

- Handle distortion of speech signal due to background noise
- Handle distortion of speech signal due to handsets / transducers

![Diagram showing clean speech, additive noise, channel noise, and noisy speech]
Effect of Noise/Channel on Speech Features

Problem: Noise and channel effect will make the speech of the true speaker and impostors less discriminative.
How do We Deal with the Challenge?

- **Feature transformation** aims to reduce the effects of channel- and handset-distortion by transforming distorted features to bring them closer to the clean speech models.

\[ \hat{x}_t = f(y_t) \]
Type of Feature Transformation

• Non-blind compensation – based on a priori knowledge of the all possible channels
  ➢ Probabilistic feature mapping (PFM), proposed in this paper
  ➢ Fast probabilistic feature mapping (fPFM), proposed in this paper
  ➢ Stochastic feature transformation (SFT), by Mak & Kung (2002)

• Blind compensation – without a priori knowledge of the channel characteristics
  ➢ Cepstral Mean Subtraction (CMS), by Atal (1974)
  ➢ Blind stochastic feature transformation (BSFT), by Yiu, Mak and Kung (2004, 2005)
  ➢ Fast BSFT (fBSFT), proposed in this paper
Feature Mapping (Non-Blind)

Channel-independent (CI) model

Channel-dependent (CD) model

MAP Adaptation

\[ x - \mu_k^{CI} = (y - \mu_k^{CD}) \frac{\sigma_k^{CI}}{\sigma_k^{CD}} \]

\[ \Rightarrow x = f_{FM}(y) = (y - \mu_k^{CD}) \frac{\sigma_k^{CI}}{\sigma_k^{CD}} + \mu_k^{CI} \]
Probabilistic Feature Mapping (Non-Blind)

Channel-independent (CI) model

Channel-dependent (CD) model

\[
\sum_{j=1}^{M} P(j \mid y)(x - \mu^{\text{CI}}_j) = \sum_{j=1}^{M} P(j \mid y)(y - \mu^{\text{CD}}_j) \frac{\sigma^{\text{CI}}_j}{\sigma^{\text{CD}}_j}
\]

\[
\Rightarrow x = \sum_{j=1}^{M} P(j \mid y) \left( (y - \mu^{\text{CD}}_j) \frac{\sigma^{\text{CI}}_j}{\sigma^{\text{CD}}_j} + \mu^{\text{CI}}_j \right)
\]

\[
P(j \mid y) = \frac{\pi^{\text{CD}}_j p(y \mid \mu^{\text{CD}}_j, \Sigma^{\text{CD}}_j)}{\sum_{l=1}^{M} \pi^{\text{CD}}_l p(y \mid \mu^{\text{CD}}_l, \Sigma^{\text{CD}}_l)}
\]
Fast Probabilistic FM (Non-Blind)

Channel-independent (CI) model

Channel-dependent (CD) model

\[
\sum_{j \in C} P(j \mid y) (x - \mu_{j,CI}) = \sum_{j \in C} P(j \mid y) (y - \mu_{j,CD}) \frac{\sigma_{j,CI}}{\sigma_{j,CD}}
\]

\[
\Rightarrow x = \sum_{j \in 1} P(j \mid y) \left[ (y - \mu_{j,CD}) \frac{\sigma_{j,CI}}{\sigma_{j,CD}} + \mu_{j,CI} \right]
\]

\[
P(j \mid y) = \frac{\pi_{j,CD} \cdot p(y \mid \mu_{j,CD}, \Sigma_{j,CD})}{\sum_{l=1}^{M} \pi_{l,CD} \cdot p(y \mid \mu_{l,CD}, \Sigma_{l,CD})}
\]
Different Type of Feature Mapping

Feature Mapping

\[ x = (y - \mu_{k}^{CD}) \frac{\sigma_{k}^{CI}}{\sigma_{k}^{CD}} + \mu_{k}^{CI} \]

where \( k = \arg \max_{j=1}^{M} \pi_{j}^{CD} p(y | \mu_{j}^{CD}, \Sigma_{j}^{CD}) \)

Probabilistic Feature Mapping

\[ x = \sum_{j=1}^{M} P(j | y) \left[ (y - \mu_{j}^{CD}) \frac{\sigma_{j}^{CI}}{\sigma_{j}^{CD}} + \mu_{j}^{CI} \right] \]

\[ P(j | y) = \frac{\pi_{j}^{CD} p(y | \mu_{j}^{CD}, \Sigma_{j}^{CD})}{\sum_{l=1}^{M} \pi_{l}^{CD} p(y | \mu_{l}^{CD}, \Sigma_{l}^{CD})} \]

Fast Probabilistic Feature Mapping

\[ x = \sum_{j \in C} P(j | y) \left[ (y - \mu_{j}^{CD}) \frac{\sigma_{j}^{CI}}{\sigma_{j}^{CD}} + \mu_{j}^{CI} \right] \]

where \( C \) contains the indexes of top-\( C \) Gaussians
Clustering Effect of Feature Mapping

\[ y: \text{data from channel-dependent source} \]

\[ x: \text{transformed features} \]

\[ \boldsymbol{\mu}_k: \text{Centers of channel-independent root model} \]

\[ \boldsymbol{\mu}_k^{\text{CD}_i}: \text{Centers of channel-dependent model} \]
Stochastic Feature Transformation (Non-Blind)

Aim: To transform distorted features to fit the clean speech models using handset detection

\[ k^* = \arg \max_{k=1}^{H} \sum_{t=1}^{T} \log p(y_t | \Gamma_k) \]

Problem: The handset detector may not work for “unseen” handsets
Blind Stochastic Feature Transformation

Aim: To transform distorted features to fit the clean speech models without handset detection.

\[ x_t = f_{v*}(y_t) \]
Blind Stochastic Feature Transformation

- Distorted speech vectors can be transformed to fit the clean speech model.

\[ x_t = f_v(y_t) = Ay_t + b \]

\[ \{A, b\} = \arg \max_{A, b} p(f_v(Y) \mid \Lambda_X) \]

Distorted speech vectors

Clean model

\( \Lambda_X \)
Fast Blind Stochastic Feature Transformation

Blind Stochastic Feature Transformation

Based on all Gaussians:

\[ Q(v' | v) = \sum_{t=1}^{T} \sum_{j=1}^{M_v} h_j(f_v(y_t)) \log \left\{ \frac{p(f_{v'}(y_t) | \mu_j, \Sigma_j)}{|J_{v'}(y_t)|} \right\} \]

Based on the top-C Gaussians only:

\[ Q(v' | v) = \sum_{t=1}^{T} \sum_{j \in C} h_j(f_v(y_t)) \log \left\{ \frac{p(f_{v'}(y_t) | \mu_j, \Sigma_j)}{|J_{v'}(y_t)|} \right\} \]
Experiments

• 2000 NIST speaker recognition evaluation set.
• 1003 target speakers (457 male and 546 female).
• Enrollment: approximately 1 minutes of speech.
• Verification: 6052 utterances (3026 male and 3026 female).
• Each verification utterance is evaluated against 11 hypothesized gender-matched speakers.
• For each gender, gender-dependent evaluation utterances from NIST99 were used to train a 1024-component gender-dependent universal background models (UBMs).
• Speaker and background models:
  – MFCC + ΔMFCC
  – Speaker models were adapted from the gender-dependent background model using MAP adaptation
Results: DET plots

![Diagram showing DET plots for various methods including CMS and BSFT+PFM.](image)
Results: EER & Transformation Time

<table>
<thead>
<tr>
<th>Method</th>
<th>Blind</th>
<th>Non-blind</th>
<th>Blind+ Non-Blind</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BSFT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fBSFT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fPFM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fBSFT+PFM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Equal Error Rate (%)

Transformation Time (sec.)
Conclusions

• Probabilistic feature map performs slightly better than feature map and that the fast approach can reduce computation time substantially.

• Among the approaches investigated, the fast BSFT (fBSFT) strikes a good balance between computational complexity and error rates.

• We advocate Fast BSFT for robust speaker verification because it achieves good performance without any a priori knowledge of the communication channel.
Biometric Verification

Enrolment Phase (Training)

Verification Phase (Testing)

Claimant
State-of-the-art Speaker Verification Systems

- Speaker verification is a biometric technology that aims to authenticate users via their voice patterns.

- The lack of robustness to channel variability and the acoustic mismatch between enrollment and verification conditions remain a major practical challenge.

- Currently, this problem is addressed by a technique called channel mismatch compensation.
Blind Stochastic Feature Transformation

Training Utterance of Speaker $s$ → MAP Adaptation → Model Fusion → Composite Model → Estimation of SFT Parameters

Speech of All Client Speakers → EM → Background Model → Verification Utterance $y_t$ → SFT $\hat{x}_t = Ay_t + b$ → Computation of Likelihood Ratio → Decision

$\Lambda^M_s$ → $\Lambda^M_b$ → $\Lambda^N_b$ → MAP Adaptation → Speaker Model

$\Lambda^N_s$ → $\Lambda^2M_c$ → $v = \{A,b\}$

Enrollment Phase

Verification Phase
Blind Stochastic Feature Transformation

- In BSFT, the transformed feature vector is
  \[ x = f_v(y) = Ay + b \]
- The auxiliary function is
  \[
  Q(A', b' | A, b) = \sum_{t=1}^{T} \sum_{j=1}^{M_c} h_j(f_v(y_t)) \log \left( \frac{p(f_v'(y_t) | \mu_j, \Sigma_j)}{|J_v'(y_t)|} \right)
  \]
  \[
  \frac{\partial Q(A', b' | A, b)}{\partial a'_i} = 0 \quad \frac{\partial Q(A', b' | A, b)}{\partial b'_i} = 0
  \]
  \[
  b'_i = \frac{p_i - q_i a'_i}{r_i}
  \]
  \[
  \Rightarrow \left\{ \begin{array}{c}
  s_i - \frac{q_i^2}{r_i} a'_i^2 + \left( \frac{q_i p_i}{r_i} - u_i \right) a'_i - T = 0
  \end{array} \right.
  \]
Blind Stochastic Feature Transformation

\[
p_i = \sum_{t=1}^{T} \sum_{j=1}^{M_c} h_j(f_{\nu}(y_t))\mu_{ji}\sigma^{-2}_{ji}
\]

\[
q_i = \sum_{t=1}^{T} \sum_{j=1}^{M_c} h_j(f_{\nu}(y_t))\nu_{ji}\sigma^{-2}_{ji}
\]

\[
r_i = \sum_{t=1}^{T} \sum_{j=1}^{M_c} h_j(f_{\nu}(y_t))\sigma^{-2}_{ji}
\]

\[
s_i = \sum_{t=1}^{T} \sum_{j=1}^{M_c} h_j(f_{\nu}(y_t))\nu^2_{ji}\sigma^{-2}_{ji}
\]

\[
u_i = \sum_{t=1}^{T} \sum_{j=1}^{M_c} h_j(f_{\nu}(y_t))\mu_{ji}\nu_{ti}\sigma^{-2}_{ji}
\]

- In the fast BSFT, the auxiliary function is

\[
Q(A', b' | A, b) = \sum_{t=1}^{T} \sum_{j \in C} h_j(f_{\nu}(y_t)) \log \left\{ \frac{p(f_{\nu}(y_t) | \mu_j, \Sigma_j)}{|J_{\nu}(y_t)|} \right\}
\]

where \( C \) contains the indexes of top-C Gaussians
Procedures for creating target speaker models for FM, PFM, and BSFT

Speaker model creation for BSFT

CD target speaker GMM $\Lambda_{g,k}^{CD}$

CD: channel-dependent
GD: gender-dependent
CID: channel-independent

CD speaker-dependent speech data from NIST00

CID GD speech data from NIST99

MAP

EM

CID GD GMM (Root, $\Lambda_g$)

CID speaker-dependent speech data

CD GD GMM $\Lambda_g^{CD}$

CID target speaker GMM $\Lambda_{g,k}^{CD}$

Procedures for creating target speaker models for FM and PFM

Feature Mapping / Probabilistic Feature Mapping

Speaker model creation for FM and PFM

MLSP'06