DEEP FEATURES FOR SPEAKER RECOGNITION

Ana Mantecón Jené
Speech and Language Processing Seminar 2017
Contents

1. Introduction to speaker recognition

2. Speaker verification deep front-end models
   ▪ Results

3. Speaker identification deep front-end models
   ▪ Results

4. Conclusions

5. Homework
Introduction to speaker recognition
Speaker Recognition

- Distinction between speaker verification and identification
Speaker recognition

- Enrollment stage:
  - Spk utt1
  - Spk utt2
  - Spk utt3

- Recognition stage:
  - Spk utt

Diagram:
- Front-end feature extraction
- Speaker modelling
- Decision
Speaker recognition

- **Text-dependent** speaker recognition
  - The text is the same in the enrollement and the recognition stages
  - Speaker verification (more secure)

- **Text-independent** speaker recognition
  - The text is NOT the same in the enrollement and the recognition stages
  - Speaker indentification
Speaker recognition approaches

- From GMM-UBM to i-vector

GMM-UBM framework → JFA → i-vector

GMMs → Supervector → Low-dimensional vector
GMM-UBM approach

- A speaker utterance is represented by a GMM which is adapted from the UBM via MAP

- UBM training
JFA approach

- Joint Factor Analysis is a model that represents speaker and session variability in GMM’s.

- Each speaker utterance is represented by a supervector $M$.

- Each target speaker is represented by a supervector where session variability is removed to compensate for inter-session variability and channel differences.
JFA approach

- Speaker utterance represented by a supervector $M$.

\[ M = m + V y + U x + D z \]

- $m$ speaker and session-independent vector generated from the UBM.
- $V$ and $D$ define the speaker subspace
- $U$ defines the session subspace
- $y$ represents the speaker factors
- $x$ represents channel factors
- $Dz$ residual to compensate for the speaker information that may not be caught by $Vy$. 
i-vector approach

➢ State of the art

➢ Total variability subspace, has both speaker and session variabilities

➢ Supervector now:

\[ M = m + T w \]

• \( T \) low Rank matrix of speaker and session variability.
• \( w \), identity vector -> i-vector
• \( m \) speaker and session-independent vector generated from the UBM
i-vector approach
Evaluation

- **Speaker verification**
  - DET curve
    - False Rejection Rate = FAP
    - False Acceptance Rate = MP
  - Equal Error rate (EER)
    - FRR = FAR

- **Speaker identification**
  - Recognition rate
Feature extraction

- Speaker recognition systems use short-time spectral features.
  - Originally designed for speech recognition tasks.
  - Not optimized for speaker discrimination.

- Deep features: obtained by passing original spectral features such as MFCCs or PLPs through deep models.
Deep Learning

- Successfully used in Speech recognition

- DNNs possess strong capability to model nonlinear representations. They are believed that they can be good to extract discriminative features.

- Deep features are proposed to improve the speaker recognition systems.
Speaker verification deep front-end models

Yuan Liu et al.
**Speaker verification system**

- *Text-dependent* speaker verification

- Approaches:
  - GMM-UBM framework
  - i-vector framework

- Evaluation: RSR2015 data corpus. Designed for text-dependent speaker recognition
  - 30 Fixed pass-phrases
  - Part 1: 72h of audio. 300 speakers divided into background, development and test sets.
  - 9 sessions, 3 repetitions per session, 30 short phrases per session (3.2s avg)
Feature extraction deep models

- Approaches of extracting and using features from deep learning methods to text-dependent speaker verification

- Models:
  - Deep RBMs
  - Speech-discriminant DNN
  - Speaker-discriminant DNN
  - Multi-task joint-learned DNN
FE: Deep RBMs

- Generative model
- Unsupervised process
  - Very large amount of data can be used
- All speech characteristics may be represented
  - Phone-level
  - Speaker-level
  - Channel-level
- Context information can also be included
FE: Speech-discriminant DNN

- Trained for speech discrimination
- Supervised mode
- Triphone states labels as targets
- RBM pretraining
FE: Speaker-discriminant DNN

- Speaker discriminative ability is enhanced
- Output classes represent individual speakers
- RBM pretraining
FE: Multi-task joint-learned DNN

- Combination of the two previous DNNs
- The target are both speakers and texts. Two models:
  - Speaker+phrase
  - Speaker+triphone
- Total loss function is the sum of the two original loss functions.
Deep models training

- Deep models:
  - 7 hidden layers
  - 1024 nodes per layer
  - Input: 39-dim PLP features with a context window of ±5 frames

- RSR2015 bkg data: 194 speakers, 100 male/94 female, 30 fixed phrases.
- 3001 tied-triphone-states, 50h English task

- Deep RBM: contrastive divergence algorithm
- DNNs: back-propagation
  - Cross-entropy objective function
  - L2-norm weight-decay
Evaluation: GMM-UBM approach

- Deep Features + PCA with dimensionality reduction
- Tandem features = Deep Features + PCA with dimensionality reduction + PLP spectral features
- Mean and variance normalization

- Baseline system
  - 39-dimensional PLP features with mean and variance normalization
  - Energy-based VAD
  - Gender-independent UBM of 1024 components
    - RSR2015 bkg data: 194 speakers 100 male/94 female
Evaluation results

- GMM-UBM framework

<table>
<thead>
<tr>
<th>Layer index</th>
<th>RBM</th>
<th>Speech-DNN</th>
<th>Speaker-DNN</th>
<th>Speaker + phrase DNN</th>
<th>Speaker + phone DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deep fea + PLP</td>
<td>Deep fea + PLP</td>
<td>Deep fea + PLP</td>
<td>Deep fea + PLP</td>
<td>Deep fea + PLP</td>
</tr>
<tr>
<td>Base GMM-UBM</td>
<td>1.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd-layer</td>
<td>1.25</td>
<td>1.45</td>
<td>1.08</td>
<td>1.06</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.89</td>
<td>0.80</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>4th-layer</td>
<td>1.23</td>
<td>1.86</td>
<td>1.48</td>
<td>1.07</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>0.94</td>
<td>1.04</td>
<td>0.97</td>
<td>0.84</td>
<td>0.93</td>
</tr>
<tr>
<td>7th-layer</td>
<td>1.46</td>
<td>1.94</td>
<td>2.15</td>
<td>1.13</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>1.06</td>
<td>1.15</td>
<td>0.96</td>
<td>0.92</td>
<td>0.93</td>
</tr>
</tbody>
</table>

- Tandem features

- Best DNN: multi-task joint-learned
Evaluation results

- GMM-UBM framework

<table>
<thead>
<tr>
<th>Performance of different deep features combination.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PLP</strong></td>
</tr>
<tr>
<td>_______</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
</tr>
</tbody>
</table>

- Multi-Deep features combinations
Evaluation: i-vector approach

- Last hidden layer outputs
- Deep features are the speaker identity vectors (average of all the frames)
- Classifications: Cosine similarity, LDA, PLDA.

Baseline system
- LDA+Cosine similarity classification measure
- Matrix T trained with NIST SRE 2005 and NIST SER 2008
Evaluation results

- i-vector approach

- Baseline LDA + Cosine similarity: 5.01 EER

<table>
<thead>
<tr>
<th>DNN</th>
<th>Classifier</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>r-vector</td>
<td>Cosine sim.</td>
<td>17.61</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>PLDA</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>PLDA + LDA</td>
<td>0.22</td>
</tr>
<tr>
<td>p-vector</td>
<td>Cosine sim.</td>
<td>4.67</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>PLDA + LDA</td>
<td>0.30</td>
</tr>
<tr>
<td>d-vector</td>
<td>Cosine sim.</td>
<td>7.55</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>PLDA</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>PLDA + LDA</td>
<td>1.15</td>
</tr>
<tr>
<td>spkr-phr-vector</td>
<td>Cosine sim.</td>
<td>4.41</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>PLDA</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>PLDA + LDA</td>
<td>0.20</td>
</tr>
<tr>
<td>spkr-pho-vector</td>
<td>Cosine sim.</td>
<td>4.41</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>PLDA</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>PLDA + LDA</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Speaker identification deep front-end models

Zhaofeng Zhang et al.
Reverberation in speaker identification

- Distant talking speech capture useful for many applications

- Microphones records reverberation signals of the direct sounds

- Accuracy of distant-talking speaker identification system is reduced
Bottleneck features

- Used for nonlinear feature transformation and dimensionality reduction

- Obtained by sampling the activation function of one of the DNN hidden layers. Designed to have a low dimension -> bottleneck layer
Feature extraction deep models

- DNN-based bottleneck feature
  - Discriminative feature
- DAE-based dereverberation
  - Surpressed reverberation

- Combination of both models

- Training data: 100 speakers (50/50) x 5 utterances x 3 environments
- Inputs: 25-dimensional MFCCs
Proposed identification system

- GMM-UBM framework
- Speaker with maximum likelihood selected as target speaker
BF-DNN model

- Output classes represent individual speakers
- Pre-training: unsupervised RBM
- Fine-tuning: Back-propagation
  - Cross-entropy
- 9 hidden layers
  - Size: 1024
  - Bottleneck size: 25
DAE model

- Denoising autoencoder
  - Forces reconstruction of the clean speech
  - Noise removal
- Pre-training: RBM
- Fine-tuning: Back-propagation
  - Cross-entropy?
- Context information ±4
- 3 hidden layers
  - Size: 1024
  - Bottleneck size: 25
Evaluation

- Training set: 100 speakers (50/50) x 5 utterances x 3 environments

- Test set: 100 speakers x 20 utterances x 5 environments

- Comparison with three other dereverberation methods
Evaluation results

- DAE retains speaker characteristics and suppresses reverberation

<table>
<thead>
<tr>
<th>Method</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMN</td>
<td>72.5</td>
</tr>
<tr>
<td>MCLMS-SS</td>
<td>87.8</td>
</tr>
<tr>
<td>MSLP-SS</td>
<td>83.8</td>
</tr>
<tr>
<td>BF-DNN</td>
<td>88.0</td>
</tr>
<tr>
<td>DAE</td>
<td>91.0</td>
</tr>
<tr>
<td>DAE + BF-DNN</td>
<td>92.5</td>
</tr>
</tbody>
</table>
Conclusion

- Using Deep models for feature extraction is very promising in speaker recognition

- Deep models extract discriminative features

- DAE retains speaker characteristics and suppresses reverberation

- Substantial improvement in speaker recognition using Deep features.
Homework

- In the speaker verification GMM-UBM approach, what is the impostor model used for and what model is usually used? Is there an impostor model in speaker identification, if not, why?

- Why are deep features useful for speaker recognition? (brief explanation)
References
