Performance Evaluation of Deep Bottleneck Features for Spoken Language Identification

Bing Jiang¹, Yan Song¹, Si Wei²,
Meng-Ge Wang¹, Ian McLoughlin¹, Li-Rong Dai¹

¹, National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China

², iFlytek Research, Anhui USTC iFlytek Co. Ltd.
Outline

• Background
• Our Method
• Experiments
• Conclusions
Outline

• Background
• Our Method
• Experiments
• Conclusions
• Language Identification is a typical problem in Machine learning
There are many language-independent nuisances covered by original acoustical feature.
- Speaker variations
- Channel variations
- Special content variations
- Noise variations

Feature improvement
- MFCC \rightarrow SDC
  - Temporal extension
- Compensation in feature domain
  - Factor analysis

So difficult
Background

• Model
  – Generative model → Discriminative model
    • GMM-UBM
    • SVM
    • MMI
  – **i-vector** is the state of the art
    • Factor analysis
    • With compensation methods :
      – LDA
      – WCCN
      – PLDA

• More suitable features are wanted.....
Recently, DNN is drawing lots of attention
  - Non-linear modeling capability
    • Deep layers structure
    • Non-linear activation function
  - Feature learning capability
    • Extracting information about the target layer by layer

Using neural network to extract the discriminative feature for LID task??
  - PLLR
  - MLP
  - Deep Bottleneck Feature
Outline

• Background
• Our Method
• Experiments
• Conclusions
• What are Deep bottleneck features?

\[ Y = [y_1, y_2, \ldots, y_f] \]
Our Method

• What are Deep bottleneck features?
Our Method

• Why do we use Deep bottleneck features?
  – The target class
    • Phonemes or phoneme states are suitable for language identification task
      – Statistical method
    • A low-dimensional compact representation of the original inputs
  – Non-linear transformation
  – Discriminative features
Our Method

- Why do we use Deep bottleneck features?

(a) \[ s(t) = [m(t)^T, \Delta c(t, 0)^T, \Delta c(t, 1)^T, \ldots, \Delta c(t, k-1)^T]^T \]

(b) SDC PK DBF
Our Method

• How to train the DBF extractor?

- Bernoulli-Bernoulli RBM ($M_5 \times M_5$) $W_5$
- Bernoulli-Bernoulli RBM ($M_4 \times M_4$) $W_4$
- Bernoulli-Bernoulli RBM ($M_3 \times M_3$) $W_3$
- Bernoulli-Bernoulli RBM ($M_2 \times M_2$) $W_2$
- Gaussian-Bernoulli RBM ($D \times M_1$) $W_1$

DBN

Softmax output layer $W_6$

DNN

Bottleneck layer

Input Feature

DBF Extractor
Outline

- Background
- Our Method
- Experiments
- Conclusions
Experiments

• DNN training database
  – 500 hours Mandarin telephone database

• Evaluation database
  – NIST LRE 2009
Experiments

- Exper1: Comparison with SDC
  - DBF: $43 \times 11 - 2048 - 2048 - 43 - 2048 - 2048 - 2048 - 6004$

<table>
<thead>
<tr>
<th></th>
<th>30s</th>
<th></th>
<th>10s</th>
<th></th>
<th>3s</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>$C_{avg}$</td>
<td>EER</td>
<td>$C_{avg}$</td>
<td>EER</td>
<td>$C_{avg}$</td>
</tr>
<tr>
<td>SDC</td>
<td>5.64</td>
<td>5.41</td>
<td>8.93</td>
<td>8.73</td>
<td>17.64</td>
<td>17.44</td>
</tr>
<tr>
<td>SDC+FA</td>
<td>3.49</td>
<td>3.48</td>
<td>5.77</td>
<td>5.71</td>
<td>17.06</td>
<td>16.80</td>
</tr>
<tr>
<td>DBF</td>
<td>6.70</td>
<td>6.21</td>
<td>7.83</td>
<td>7.56</td>
<td>13.45</td>
<td>13.42</td>
</tr>
</tbody>
</table>

- 2048 mixture GMM-UBM
Experiments

• Exper2: Context window size of DNN input
  – Motivation
    • Context window size is sensitive for LID
    • The parameter for SDC (7-1-3-7)
      – Can cover 21 frames
    • For LID, the input window should be more length than speech recognition
      – Speech recognition: 5-1-5
Experiments

- **Exper2**: Context window size of DNN input
  - DBF:43xn-2048-2048-43-2048-2048-6004

<table>
<thead>
<tr>
<th>Window Size</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>C_{avg}</td>
<td>EER</td>
</tr>
<tr>
<td>5-1-5</td>
<td>6.70</td>
<td>6.21</td>
<td>7.83</td>
</tr>
<tr>
<td>10-1-10</td>
<td>5.56</td>
<td>5.44</td>
<td>6.89</td>
</tr>
<tr>
<td>15-1-15</td>
<td>5.91</td>
<td>5.82</td>
<td>7.76</td>
</tr>
<tr>
<td>20-1-20</td>
<td>5.64</td>
<td>5.51</td>
<td>7.41</td>
</tr>
</tbody>
</table>
Experiments

• Exper3: Dimension of DBF
  – Motivation
    • DNN training forces the activation signals in the bottleneck layer to form a **low-dimensional compact representation** of the original inputs
    • Find the relationship of the feature dimension and the performance.
Experiments

- **Exper3: Dimension of DBF**
  - DBF:43x21-2048-2048-\textbf{d}-2048-2048-6004

<table>
<thead>
<tr>
<th>Dim</th>
<th>30s</th>
<th></th>
<th>10s</th>
<th></th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>(C_{\text{avg}})</td>
<td>EER</td>
<td>(C_{\text{avg}})</td>
<td>EER</td>
</tr>
<tr>
<td>20</td>
<td>5.43</td>
<td>5.40</td>
<td>7.43</td>
<td>7.42</td>
<td>14.92</td>
</tr>
<tr>
<td>25</td>
<td>5.32</td>
<td>5.32</td>
<td>7.53</td>
<td>7.48</td>
<td>15.04</td>
</tr>
<tr>
<td>30</td>
<td>6.44</td>
<td>6.40</td>
<td>8.80</td>
<td>8.67</td>
<td>16.11</td>
</tr>
<tr>
<td>35</td>
<td>6.38</td>
<td>6.18</td>
<td>8.07</td>
<td>8.04</td>
<td>14.15</td>
</tr>
<tr>
<td>43</td>
<td>5.56</td>
<td>5.44</td>
<td>6.89</td>
<td>6.84</td>
<td>13.34</td>
</tr>
<tr>
<td>50</td>
<td>6.44</td>
<td>6.34</td>
<td>8.12</td>
<td>8.09</td>
<td>14.08</td>
</tr>
<tr>
<td>55</td>
<td>6.13</td>
<td>5.93</td>
<td>7.52</td>
<td>7.34</td>
<td>12.82</td>
</tr>
<tr>
<td>60</td>
<td>6.77</td>
<td>6.57</td>
<td>8.15</td>
<td>8.08</td>
<td>14.03</td>
</tr>
</tbody>
</table>
Experiments

• Exper4: Generated in different layers
  – Motivation
    • The feature is more discriminative for target, the more suitable for LID??
    • The bottleneck layer is more closer to the output layer, the performance more better??

Friday, September 19, 2014
Experiments

• Exper4: Generated in different layers
  – Layer3: 43x21-2048-2048-43-2048-2048-6004
  – Layer4: 43x21-2048-2048-2048-43-2048-6004
  – Layer5: 43x21-2048-2048-2048-2048-43-6004

<table>
<thead>
<tr>
<th></th>
<th>30s</th>
<th></th>
<th>10s</th>
<th></th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>$C_{avg}$</td>
<td>EER</td>
<td>$C_{avg}$</td>
<td>EER</td>
</tr>
<tr>
<td>Layer3</td>
<td>5.56</td>
<td>5.44</td>
<td>6.89</td>
<td>6.84</td>
<td>13.34</td>
</tr>
<tr>
<td>Layer4</td>
<td>5.10</td>
<td>5.00</td>
<td>6.38</td>
<td>6.25</td>
<td>11.51</td>
</tr>
<tr>
<td>Layer5</td>
<td>4.75</td>
<td>4.65</td>
<td>6.40</td>
<td>6.29</td>
<td>12.02</td>
</tr>
</tbody>
</table>
Experiments

• Exper5: DBF with PCA
  – Motivation
  • Since we use the diagonal covariance matrix to approximate the GMM, each dimension of the input feature need to be de-correlated.
  • For SDC, (Discrete cosine transformation) DCT.
  • For DBF, we use the classical PCA to have a try.
Experiments

- **Exper5: DBF with PCA**
  - DBF:43x21-2048-2048-43-2048-2048-6004

<table>
<thead>
<tr>
<th>Dim</th>
<th>30s</th>
<th></th>
<th>10s</th>
<th></th>
<th>3s</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>$C_{avg}$</td>
<td>EER</td>
<td>$C_{avg}$</td>
<td>EER</td>
<td>$C_{avg}$</td>
</tr>
<tr>
<td>Ori-43</td>
<td>5.56</td>
<td>5.44</td>
<td>6.89</td>
<td>6.84</td>
<td>13.34</td>
<td>13.27</td>
</tr>
<tr>
<td>PCA-43</td>
<td>5.94</td>
<td>5.80</td>
<td>7.68</td>
<td>7.60</td>
<td>13.80</td>
<td>13.75</td>
</tr>
<tr>
<td>PCA-30</td>
<td>5.22</td>
<td>5.20</td>
<td>7.08</td>
<td>7.03</td>
<td>13.33</td>
<td>13.30</td>
</tr>
<tr>
<td>PCA-20</td>
<td>4.50</td>
<td>4.49</td>
<td>6.76</td>
<td>6.68</td>
<td>14.23</td>
<td>14.18</td>
</tr>
<tr>
<td>PCA-10</td>
<td>8.00</td>
<td>7.99</td>
<td>11.78</td>
<td>11.75</td>
<td>22.17</td>
<td>22.09</td>
</tr>
</tbody>
</table>

Eigenvalue $\lambda_1$ vs. Eigenvalue Number $i$
Outline

• Background
• Our Method
• Experiments
• Conclusions
Conclusions

• In this paper, we investigated the use of bottleneck features for LID task.
  – DBF can significantly improve LID performance, especially for short duration utterances.
  – DBF is a new milestone for LID research.
• We believe that using DNN to extract more suitable feature for LID will make a great process in LID community.
• For more information about DBF for LID, you can see the following paper:
THANK YOU!

Q&A

email: bing2010@mail.ustc.edu.cn

Homepage: http://home.ustc.edu.cn/~bing2010