Unsupervised Cross-Corpus Speech Emotion Recognition Using Domain-Adaptive Subspace Learning
Na Liu 1,3, Yuan Zong 2, Baofeng Zhang 1, Li Liu 3, Jie Chen 3, Guoying Zhao 3, Junchao Zhu 1
1 Key laboratory of computer vision and systems, Ministry of education, Tianjin University of Technology, China
2 Research Center for Learning Science, Southeast University, China
3 Center for Machine Vision and Signal Analysis, University of Oulu, Finland

Motivation: In this paper, we investigate unsupervised cross-corpus speech emotion recognition (SER). The training (source) and testing (target) speech signals come from two different corpora which may have different feature distributions and therefore lots of existing SER methods would not work.

Solution: (1) Construct a label space based on the label information provided by the source speech corpora to serve as the predefined common subspace for Domain-Adaptive Subspace Learning (DoSL). (2) Learn a projection matrix which transforms the source and target speech signals from the original feature space to a common subspace.

Method
Our DoSL aims at learning a projection matrix \( U \) to project the source speech feature matrix \( X_s \) from the original feature space to such a common subspace spanned by the columns of \( L_s \), which can be formulated as the following optimization problem:

\[
\min_U \left\| L_s - U^T X_s \right\|_F^2
\]

Enforce the projected source and target speech features share the similar distributions

\[
\min_U \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} U^T x_i - \frac{1}{N_t} \sum_{i=1}^{N_t} U^T x_i \right\|_2^2
\]

Minimizing the combination of the above objective functions in Eqs. (1) and (2)

\[
\min_U \left\| L_s^U - U^T X_s^U \right\|_F^2 + \lambda_1 \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} U^T x_i^s - \frac{1}{N_t} \sum_{i=1}^{N_t} U^T x_i^t \right\|_2^2 + \lambda_2 \left\| U^T \right\|_{2,1}
\]

Optimization

DoSL model is solved by using inexact augmented Lagrange multiplier (IALM) method. More specifically, by introducing a auxiliary variable \( Q \) which satisfies \( U = Q \), we convert the optimization problem of DoSL to a constrained one which can be expressed as:

\[
\min_{U} \left\| L_s^U - Q^T X_s^U \right\|_F^2 + \lambda_1 \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} Q^T x_i^s - \frac{1}{N_t} \sum_{i=1}^{N_t} Q^T x_i^t \right\|_2^2 + \lambda_2 \left\| U^T \right\|_{2,1}
\]

s.t. \( U = Q \)

Subsequently, the Lagrange function of Eq. (4) can be obtained as follows:

\[
L(U, Q, T, \mu) = \left\| L_s^U - U^T X_s^U \right\|_F^2 + \lambda_1 \left\| Q^T X_s^U \right\|_2^2 + \lambda_2 \left\| U^T \right\|_{2,1} + t_r \left( T^T (U - Q) \right) + \frac{\mu}{2} \left\| U - Q \right\|_F^2
\]

where \( X_s^U = \frac{1}{N_s} \sum_{i=1}^{N_s} x_i^s - \frac{1}{N_t} \sum_{i=1}^{N_t} x_i^t \). \( T \) is the Lagrange multiplier, and \( \mu > 0 \) is the regularization parameter.

Iteratively minimize the Lagrange function of Eq. (4) until convergence:

1. Update \( Q \)

\[
\min_{Q} \left\| L_s^U - Q^T X_s^U \right\|_F^2 + \lambda_1 \left\| Q^T X_s^U \right\|_2^2 + \lambda_2 \left\| U^T \right\|_{2,1} + t_r \left( T^T (U - Q) \right) + \frac{\mu}{2} \left\| U - Q \right\|_F^2
\]

2. Update \( U \):

\[
\min_{U} \lambda_2 \left\| U^T \right\|_{2,1} + \frac{1}{2} \left\| U^T - (Q^T - \frac{T}{\mu}) \right\|_F^2
\]

3. Update \( T \) and \( \mu \)

\[
T = T + \mu (U - Q), \mu = \max(\mu_{\text{max}}, \rho \mu)
\]

4. Check convergence: \( \left\| U^T X_s^U \right\|_F^2 < \epsilon \)

\[
\text{emotion_labels} = \arg \min_k \left\{ \left\| U^T X_s^U \right\|_F^2 \right\}
\]

Results and discussion

Results of the cross-corpus SER experiments in terms of UAR and WAR, where the common emotion states (5 classes) are Angry, Disgust, Fear, Happy and Sad.

<table>
<thead>
<tr>
<th>Source Corpus</th>
<th>Source Corpus</th>
<th>SVM</th>
<th>KMM</th>
<th>KLIEP</th>
<th>aLSIF</th>
<th>DALSIF</th>
<th>WAR</th>
<th>UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmoDB</td>
<td>EmoDB</td>
<td>30.08</td>
<td>30.00</td>
<td>23.88</td>
<td>23.74</td>
<td>23.80</td>
<td>23.52</td>
<td>23.08</td>
</tr>
<tr>
<td>INTERFACE</td>
<td>INTERFACE</td>
<td>27.95</td>
<td>24.72</td>
<td>18.03</td>
<td>28.86</td>
<td>25.05</td>
<td>26.62</td>
<td>32.24</td>
</tr>
<tr>
<td>AFew4.0</td>
<td>EmoDB</td>
<td>26.07</td>
<td>23.89</td>
<td>20.36</td>
<td>27.54</td>
<td>25.27</td>
<td>25.55</td>
<td>25.63</td>
</tr>
<tr>
<td>INTERFACE</td>
<td>INTERFACE</td>
<td>20.88</td>
<td>18.89</td>
<td>17.70</td>
<td>18.52</td>
<td>16.88</td>
<td>16.68</td>
<td>21.21</td>
</tr>
<tr>
<td>AFew4.0</td>
<td>INTERFACE</td>
<td>18.88</td>
<td>18.72</td>
<td>17.70</td>
<td>17.48</td>
<td>17.47</td>
<td>17.47</td>
<td>19.75</td>
</tr>
</tbody>
</table>

- It is convincing that the limited label information provided by a small number of samples in source database will lead to low recognition rate.
- The data imbalance between source and target databases is an important factor which will affect the cross-corpus speech emotion recognition tasks.

Acknowledgements

This research was supported by the Natural Science Foundation of China under Grants 61172185 and 61602345, the Application Foundation and Advanced Technology Research Project of Tianjin, the Academy of Finland, Tekes Fidipro program and Infotech Oulu.