Separation of Emotional and Reconstruction Embeddings on Ladder Network to Improve Speech Emotion Recognition Robustness in Noisy Conditions

Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso
Speech emotion recognition (SER) in real-world applications

Applications
- Entertainment
- Healthcare
Speech emotion recognition (SER) in real-world applications

- Needs to be robust against background noise
  - Speech can be acquired from **unconstrained noisy environment**
  - Background noise can degrade the performance of SER system
Semi-supervised learning for noise robust SER

- **Usage scenario**

  - **Background noise**
  - **Record**
  - **Transfer**

  - **Clean speech from public corpus**
  - **Emotion labels**

  - **Noisy speech from target environment**
  - **Emotional label X**

  - **Emotional attribute**
    - Arousal
    - Valence
    - Dominance

  - **How to train?**

  - **SER model**
Ladder network-based SER

- **Strengths**
  - It does not require emotional labels for target domain recordings
  - It can minimize train/test mismatch

- **Training**
  - Prediction task
    - Predict an emotional label by using labeled set
  - Reconstruction task
    - Reconstruct clean representations for each hidden layer

- **Problem**
  - Audio samples contain complex background noises
    - It can disrupt the emotion prediction task

![Ladder network for noise robust SER](image)
Decoupled ladder network (DLN)

- **Solution**
  - Decouple last hidden layer into emotion and reconstruction embedding

- **Reconstruction embedding**
  - Reconstruction task

- **Emotion embedding**
  - Prediction task

- **Lower layers**
  - Prediction + reconstruction task

Decoupled ladder network architecture
Decoupled ladder network (D LN)

- **Loss function**

\[
C_{DLN} = C_p \left( y, h_e^{(L)} \right) + \sum_{l=0}^{L-2} \lambda^l \times C_r^l \left( \hat{z}_{BN}^{(l)}, z^{(l)} \right) + \lambda^{L-1} \times C_r^{L-1} \left( \hat{z}_{BN}^{(L-1)}, z_{r} \right)
\]

- **Prediction loss** (for emotional attributes)

\[1 - CCC \left( y, h_e^{(L)} \right)\]

\[CCC(x, y) = \frac{2 \rho \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}\]

- **Reconstruction loss**

- **Emotion labels = y**

- **Accuracy** ↑

\[
\sigma_x, \sigma_y : \text{standard deviation} \\
\mu_x, \mu_y : \text{mean} \\
\rho : \text{correlation coefficient}
\]
Decoupled ladder network (DLN)

- **Loss function**

\[
C_{DLN} = C_p\left(y, h_e^{(L)}\right) + \sum_{l=0}^{L-2} \lambda^l \times C_l\left(\hat{z}_B^{(l)}, z^{(l)}\right) + \lambda^{L-1} \times C_{l-1}\left(\hat{z}_B^{(L-1)}, z_r\right)
\]

- **Prediction loss**

\[
\lambda^l = \text{weight of } C_r \text{ in layer } l = 1.0
\]

\[
C_r(x, y) = \text{MSE}(x, y) = (x - y)^2
\]

- **Reconstruction loss**

\[
C_r(x, y) = \text{MSE}(x, y) = (x - y)^2
\]

\[
\lambda^l = \text{weight of } C_r \text{ in layer } l = 1.0
\]

\[
C(r(x, y) = \text{MSE}(x, y) = (x - y)^2
\]

\[
h_r
\]

\[
\hat{z}_r^{(2)}
\]

\[
+ \mathcal{N}(0, \sigma^2)
\]

\[
\hat{z}_r^{(0)} + \mathcal{N}(0, \sigma^2)
\]

\[
+ \mathcal{N}(0, \sigma^2)
\]

\[
x
\]

\[
g(\hat{z}_r^{(0)}, u_r^{(0)}, \hat{z}_r^{(0)} - u_r^{(0)})
\]

\[
g(\hat{z}_r^{(2)}, u_r^{(2)}, \hat{z}_r^{(2)} - u_r^{(2)})
\]

\[
g(\hat{z}_r^{(1)}, u_r^{(1)}, \hat{z}_r^{(1)} - u_r^{(1)})
\]

\[
g(\hat{z}_r^{(2)}, u_r^{(2)}, \hat{z}_r^{(2)} - u_r^{(2)})
\]

\[
g(\hat{z}_r^{(0)}, u_r^{(0)}, \hat{z}_r^{(0)} - u_r^{(0)})
\]

\[
C_{r-1}\left(\hat{z}_B^{(r-1)}, z_r\right)
\]

\[
\hat{z}_B^{(r)}
\]

\[
h_r
\]

\[
z_r^{(2)}
\]

\[
+ \mathcal{N}(0, \sigma^2)
\]

\[
\hat{z}_B^{(r)} + \mathcal{N}(0, \sigma^2)
\]

\[
X
\]

\[
X
\]

\[
\text{Difference ↓}
\]
The MSP-Podcast corpus (v1.8)

- **Spontaneous emotional speech dataset**
  - Podcast recordings are collected (> 113 hours)

- **Clean speech dataset**
  - SNR is above 20dB
Noisy version of the MSP-Podcast corpus

- **Noisy speech used in previous studies**
  - Noisy speech had been artificially synthesized in previous works

- **Limitation**
  - Not enough to simulate actual recording conditions

Fixed noise + Repeated noises
Noisy version of the MSP-Podcast corpus

**Solution**
- Simultaneously playing the MSP-Podcast corpus and noise sound
- Recording it with smartphone

**Radio shows without copyright (noise)**
- Simulating non-stational background noise
  - Human voice, musical sound, and sound effect
Settings for each recording conditions

- 10dB, 5dB, 0dB conditions are collected

<table>
<thead>
<tr>
<th>Recording condition</th>
<th>(A) (inch)</th>
<th>(B) (inch)</th>
<th>Estimated SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10dB</td>
<td>5</td>
<td>35</td>
<td>11.06</td>
</tr>
<tr>
<td>5dB</td>
<td>10</td>
<td>30</td>
<td>4.34</td>
</tr>
<tr>
<td>0dB</td>
<td>15</td>
<td>25</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Emotional labels

- Noise sound is not related to the emotion
- Emotional labels can be transferred from the MSP-Podcast corpus
Data preparation

- MSP-Podcast v1.8 (clean speech set)
- Noisy version of the MSP-Podcast corpus (noisy speech set)

- Recording condition between the test set and the unlabeled training set is matched

Acoustic features

- 6,373 dimensions of 2013 ComParE feature set is used
Experiment setting

- **Baseline models**
  - **Dense network**
    - Model cannot use unlabeled set during training
  - **Ladder network**
    - Its last hidden layer is not separated into emotion and reconstruction embedding
  - All hyperparameters for the training and the number of layers, nodes are same as decoupled ladder network
## Result

- **Concordance correlation coefficient (CCC)**
  - Average CCC over 20 trials

<table>
<thead>
<tr>
<th>Task</th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR</td>
<td>10dB</td>
<td>5dB</td>
</tr>
<tr>
<td>Dense network</td>
<td>Clean</td>
<td>0.631</td>
<td>0.248</td>
</tr>
<tr>
<td>Ladder network</td>
<td>Clean</td>
<td>0.627</td>
<td>0.438</td>
</tr>
<tr>
<td>Decoupled ladder network</td>
<td>Clean</td>
<td>0.625</td>
<td>0.488</td>
</tr>
</tbody>
</table>

### Performance

- **Noisier speech shows lower performance than cleaner speech**
  - Background noise evokes detrimental effects on emotion prediction

- **Ladder network shows better performance in noisy conditions than dense network**
  - Semi-supervised learning can improve the robustness against the noise
### Result

#### Decoupled ladder network improves the ladder network

- Arousal: 11.4% (10dB), 8.4% (5dB), 10.2% (0dB) •
- Dominance: 17.1% (10dB), 13.2% (5dB), 7.0% (0dB) •

<table>
<thead>
<tr>
<th>Task</th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clean 10dB 5dB 0dB</td>
<td>clean 10dB 5dB 0dB</td>
<td>clean 10dB 5dB 0dB</td>
</tr>
<tr>
<td>Dense network</td>
<td>0.631 0.248 0.229 0.192</td>
<td><strong>0.296</strong> 0.151 0.120 0.104</td>
<td><strong>0.562</strong> 0.253 0.252 0.215</td>
</tr>
<tr>
<td>Ladder network</td>
<td>0.627 0.438 0.424 0.364</td>
<td>0.280 <strong>0.146</strong> 0.129 0.111</td>
<td><strong>0.545</strong> 0.381 0.385 0.339</td>
</tr>
<tr>
<td>Decoupled ladder network</td>
<td>0.625 <strong>0.488</strong> <strong>0.460</strong> <strong>0.402</strong></td>
<td>0.283 <strong>0.160</strong> 0.126 0.114</td>
<td><strong>0.556</strong> <strong>0.450</strong> <strong>0.436</strong> 0.397</td>
</tr>
</tbody>
</table>

**No clear improvements**
### Reconstruction by using emotion embedding

- Emotion embedding is fed into the highest layer of decoder
- Loss of using emotion embedding $>$ Loss of using reconstruction embedding
Analysis on separating the embedding

- Reconstruction loss

![Graph showing mean square error for different conditions and predictions](image_url)

More information of reconstruction in reconstruction embedding

Mean square error for:
- Arousal:
  - clean
  - 10db
  - 5db
  - 0db
- Valence:
  - clean
  - 10db
  - 5db
  - 0db
- Dominance:
  - clean
  - 10db
  - 5db
  - 0db

Testing condition / Prediction type:
- Reconstruction from $h_e$
- Reconstruction from $h_r$

Values:
- Arousal: 8.8614
- Valence: 8.8614
- Dominance: 8.8614
Analysis on separating the embedding

- **Reconstruction loss**

<table>
<thead>
<tr>
<th>Task</th>
<th>Valence</th>
<th>SNR</th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense network</td>
<td>0.296</td>
<td>0.151</td>
<td>0.120</td>
<td>0.104</td>
<td></td>
</tr>
<tr>
<td>Ladder network</td>
<td>0.280</td>
<td>0.146</td>
<td>0.129</td>
<td>0.111</td>
<td></td>
</tr>
<tr>
<td>Decoupled ladder network</td>
<td>0.283</td>
<td><strong>0.160</strong></td>
<td>0.126</td>
<td>0.114</td>
<td></td>
</tr>
</tbody>
</table>

Less differences in reconstruction loss

Less improvement in DLN
Decouple ladder network
- Decouples the emotional and residual information to improve performance in noisy conditions

Noisy version of the MSP-Podcast corpus
- Simulates noisy, unconstrained recording environment.

Exclusively focus on emotional information

Noisy version of the MSP-Podcast corpus
- Simulates a noisy, unconstrained real-world environment
- Contains 3 recording conditions with different levels of signal-to-noise ratio (SNR) (10dB, 5dB, 0dB)

Emotion prediction

Reconstruction embedding

Decoupled ladder network
- Decouples emotional and residual information in last hidden layer
- Can be robust against non-stationary background noises
Release of the MSP-Podcast corpus

- **Academic license**
  - Federal Demonstration Partnership (FDP)
    - Data Transfer and Use Agreement
  - Free access to corpus

- **Commercial license**
  - Commercial license through UT Dallas

- **Plan to release the noisy version of the MSP-Podcast corpus**

https://msp.utdallas.edu
This study was supported by NIH under grant 1R01MH122367-01.

Questions or Contact: Seong-Gyun Leem

SeongGyun.Leem@UTDallas.edu