Detection of Mood Disorder Using Speech Emotion Profiles and LSTM

Tsung-Hsien Yang, Chung-Hsien Wu, Kun-Yi Huang, and Ming-Hsiang Su

Department of Computer Science and Information Engineering National Cheng Kung University, Taiwan
Outline

- Introduction
- Mood database collection
- System framework
  - Database adaptation
  - Data Reconstruction
  - Emotion Profile Generation
  - Long short-term memory
- Experimental results
- Conclusions
Mood disorder contains **Unipolar Depression (UD)** and **Bipolar Disorder (BD)**, which are mental illness.

BD experiences two opposite and extreme emotional states: **mania (high)** and **depression (low)** through **euthymia**, which are different from UD.
Motivation

- The doctors are likely to misdiagnose the patients in low mood of bipolar disorder as unipolar depression.
  - According to the statistics, around 40% misdiagnosis leads to patients not receiving appropriate treatment.

- Correct diagnosis, using DSM-5 as diagnostic criteria, needs a long-term tracking.

- Developing a system for mood disorder detection based on physiological signals or audio-visual signals can help doctor to correctly diagnose mood disorder.
Goal

- Among these signals, **speech** is the most natural way to express emotion and the simplest way to collect data.
- How to develop a **mood disorder detection** system for **short-term detection** becomes an important issue.
CHI-MEI Mood Database Collection

- In a closed environment room.

- Only when the subject is in a stable mood-state, the evaluation could be conducted.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>DSSS</th>
<th>MDQ-C</th>
<th>YMRS</th>
<th>SAS</th>
<th>BARS</th>
<th>CGI-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td>Scale &lt; 9</td>
<td>Scale &lt; 6</td>
<td>Scale &lt; 3</td>
<td>Scale &lt; 1</td>
<td>Scale &lt; 1</td>
<td>Scale &lt; 4</td>
</tr>
</tbody>
</table>
CHI-MEI Mood Database Collection

- 39 subjects (27 females and 12 males) from 3 different categories (13 BDs, 13 healthy controls (Cs) and 13 UDs) participated in the data collection process.

- All of the subjects watched 6 eliciting video clips with emotions of happiness, fear, surprise, anger, sadness and disgust and answered the following 5 questions.
  1. What do you think about the above video? (happy, sad, angry, disgusting, fearful and surprised)
  2. How intense is it? (ranging from 1 to 5)
  3. Which scene in the movie is impressive? Why?
  4. Do you have any similar experience like that scene?
  5. Are you feeling sick after watching above film

- Totally 1170 responses segments were collected
The whole process takes about 30 to 40 minutes.
CHI-MEI Mood Database Collection

- Each participant provides six responses.
- Each response contains 5 answers with respect to 5 questions.
Because labeling the sentence with emotion tags is difficult and tedious, the CHI-MEI mood database is not labeled.

The eNTERFACE database is selected as the adaptation database of the emotion detector

Because this database contains six emotional expressions the same as CHI-MEI mood database

The eNTERFACE database were provided by 42 subjects (18 females and 24 males) from 14 different nationalities

Each subject was recorded for 6 emotions, and there were 5 different sentences for each emotion
System Framework

- Reconstructed data
- Training of EP Detector
- SVM-based EP Detector
- HSC_DAE Adaptation & Reconstruction
- Speech Feature Extraction
- EP Generation
- LSTM-Based Training
- eNTERFACE Dataset
- CHI-MEI Database
- Feature Extraction + DAE
- EP Generation
- Modulation Spectrum extraction
- Mood Detection
- Result
OpenSMILE is employed to extract the acoustic features of 384 dimensions

Training data: the eNTERFACE databases with emotion labels (source domain)

Test data: the CHI-MEI Mood database (target domain)

A domain adaptation method, Hierarchical Spectral Clustering (HSC), is adopted to adapt the eNTERFACE databases to fit the CHI-MEI mood database
Database Adaptation – Hierarchical Spectral Clustering (HSC)

1. $u_1$ and $v_1$ are the centroids of $U$ and $V$
2. Shifted the $U$ by the deviation vector $p_1 = u_1 - v_1$.
3. Clustering $V$ by $k$-means.
4. Calculating the centroid of each $V_{2i}$.
5. All elements in $U$ belong to its nearest $V_{2i}$.
6. Calculating the centroid of each $U_{2i}$, and deviation vector between $U_{2i}$ and $V_{2i}$.
7. Each cluster element $c_i$ in shifted $U$ is shifted as
8. Repeat from step 3 to step 7
Reconstruction from Biased (Noisy) Data

- The adapted data are used as the input to train a denoising autoencoder (DAE).
- The DAE reconstructs the CHI-MEI-adapted eNTERFACE emotional data, which are regarded as the eNTERFACE data with noises due to different environments, participants, expressions, etc. to the original eNTERFACE emotional data.
HSC-based Denoising Autoencoder

\[ x_1 \quad x_2 \quad \ldots \quad x_i \quad \ldots \quad x_N \]

\[ h_1 \quad h_2 \quad \ldots \quad h_j \quad \ldots \quad h_L \]

\[ y_1 \quad y_2 \quad \ldots \quad y_i \quad \ldots \quad y_N \]

CHI-MEI-adapted eNTERFACE data

Reconstructed eNTERFACE data

\[ \| x - y \|^2 \]

HSC-Based Adaptation

CHI-MEI Mood Database

eNTERFACE data
An SVM-based *Emotion Profile (EP)* detector is adopted to provide a quantitative measure for expressing the degree of the presence or absence of a set of basic emotions within an expression. [Mower et al. 2011]

- $O_n$ is the $n$-th input feature sequence.
- $e$ is the $e$-th emotion.
Long Short-Term Memory (LSTM)

- LSTM-based method considering the temporal evolution is employed to precisely characterize the time-varying signal characteristics.
- $X_t$ is the EP vector at time $t$
Experimental setup

- The proposed method was evaluated using 13-fold cross validation.
- Each fold contains 36 subjects for training and 3 subjects (one of each category, i.e., BDs, Cs, or UD) for testing.
- Linearly scaling each attribute to the range [0, 1] for both training and test data was used.
Experimental results (1)

- For optimizing the parameters used in the HSC-DAE, the number of hidden nodes should be determined first.
  - X-axis is the number of hidden nodes and Y-axis represents the Mean Squared Error (MSE) of the HSC-DAE.
  - We selected the reconstructed data which were trained by 900 hidden nodes.
Experimental results (2)

- We compared forward/backward LSTMs and BLSTM to analyze if the past and future contexts could influence mood disorder detection.

- All networks consisted of one hidden layer and each LSTM memory block contained one memory cell.

- The networks were trained on the training set until the cross entropy error did not improve for at least 10 epochs.

<table>
<thead>
<tr>
<th>No. of Hidden Nodes</th>
<th>LSTM (forward)</th>
<th>LSTM (backward)</th>
<th>BLSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.6667</td>
<td>0.4615</td>
<td>0.6410</td>
</tr>
<tr>
<td>32</td>
<td>0.6667</td>
<td>0.5897</td>
<td>0.6410</td>
</tr>
<tr>
<td>64</td>
<td><strong>0.6923</strong></td>
<td><strong>0.7179</strong></td>
<td><strong>0.7179</strong></td>
</tr>
<tr>
<td>128</td>
<td>0.6667</td>
<td><strong>0.7179</strong></td>
<td>0.6923</td>
</tr>
<tr>
<td>256</td>
<td><strong>0.6923</strong></td>
<td>0.6410</td>
<td>0.6667</td>
</tr>
</tbody>
</table>
Motivated by the success of Deep Neural Network, we stacked two LSTMs and two BLSTMs respectively.
Comparison among different classifiers

- The SVM performed grid search with Radial basis function (RBF) kernel using LibSVM toolkit.
- The MLP was a three-layer topology with 64 hidden nodes which were fine-tuned to achieve the best performance.
- The LSTM-based classifiers outperformed SVM and MLP classifiers.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SVM</th>
<th>MLP</th>
<th>LSTM-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.4983</td>
<td>0.4197</td>
<td><strong>0.7692</strong></td>
</tr>
</tbody>
</table>
Conclusions

- We proposed an LSTM-based approach to modeling the long-range contextual information based on the temporal change of speech responses for mood disorder detection.
  - The HSC-DAE method was employed for domain adaptation and data denoising.
  - The LSTM-based method is applied to model the EP sequence for mood disorder detection.
- In the future, combining other modalities such as facial expression information is helpful to improve system performance.
Questions?