Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds

Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen

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

- Introduction
- Database
- Proposed Methods
- Experimental Results
- Conclusions
Introduction

- **Speech Emotion Recognition (SER)** is a hot research topic in the field of Human Computer Interaction. It has a potentially wide applications, such as chatbots, banking, call centers, car board systems, computer games etc.

- In the past, research on speech emotion recognition mainly focused on **discriminative emotion features** and **recognition models**.

- Only few existing emotion recognition systems focused on **nonverbal part of speech** in speech emotion recognition.
  - In real-life communication, **nonverbal sounds**, such as laughter, cries or **emotion interjections**, within an utterance play an important role for emotion recognition.

- This work adopted the **nonverbal parts** to improve the performance of emotion recognition.
Goal

- Develop a speech emotion recognition mechanism that considers **verbal** and **nonverbal parts** of speech signals.

Issues to be considered

- Emotion database
  - **A spontaneous speech emotion corpus** containing emotional nonverbal sounds in speech

- Recognition unit
  - **Speech/sound segment** useful to characterize emotion information

- Temporal Change of Emotion
  - **A sequential model** (seq2seg) for characterizing the temporal change of emotions in a conversation
- **NNIME**, a spontaneous speech emotion corpus, containing emotional nonverbal sounds in speech, was used for this study.
## Literature Review – Recognition Unit

<table>
<thead>
<tr>
<th>Segment unit</th>
<th>Audio unit</th>
<th>Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame/phoneme/word/utterance</td>
<td>Turn</td>
<td>IEMOCAP, English</td>
<td>Segment based SER using RNN [Tzinis et al., 2018]</td>
</tr>
<tr>
<td>Sentence/Second</td>
<td>Turn</td>
<td>IEMOCAP, English</td>
<td>Attentive CNN based SER with different length, features, type of speech [Neumann et al., 2017]</td>
</tr>
<tr>
<td>Prosodic action unit</td>
<td>Sentence</td>
<td>English</td>
<td>SVM based SER with discrete intonation patterns [Cao et al., 2014]</td>
</tr>
<tr>
<td>Sentence/Word/Syllable</td>
<td>Sentence</td>
<td>IITKGP-SESC, Telugu</td>
<td>SER with local and global prosodic features [Sreenivasa Rao et al., 2012]</td>
</tr>
</tbody>
</table>

- **Discrete prosodic phenomena** can provide complementary information in prediction of emotion. [Cao et al., 2014]
A sequential model (seq2seg) is helpful for characterizing the temporal change of emotions in a conversation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Input feature</th>
<th>Language</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>Prosodic feature</td>
<td>Telugu</td>
<td>[K. S. Rao et al., 2013]</td>
</tr>
<tr>
<td>Split vector quantization + naive Bayes</td>
<td>Bag of Audio Words representation</td>
<td>German</td>
<td>[F. B. Pokorny et al., 2015]</td>
</tr>
<tr>
<td>Bidirectional LSTM</td>
<td>CNN-extracted vector</td>
<td>French</td>
<td>[G. Trigeorgis et al., 2016]</td>
</tr>
<tr>
<td>Attentive CNN</td>
<td>Log-Mels, MFCCs, eGeMAPS</td>
<td>English</td>
<td>[N. T. V. Michael Neumann et al., 2017]</td>
</tr>
<tr>
<td>CLDNN</td>
<td>Log-Mels, MFCCs</td>
<td>English</td>
<td>[C.-W. Huang et al., 2017]</td>
</tr>
</tbody>
</table>
Problem – Recognition Unit

Problem
- Appropriate emotion unit of emotion expression should have various length for recognition. [Tzinis et al., 2018]

Proposed method:
- We segment the raw audio input utterances with prosodic features as basic emotion unit, which is regarded as a prosodic phrase (PPh).
Problem – Nonverbal Interval Extraction

Problem

Non-verbal part of an utterance is helpful for human to recognize emotion.

Proposed method:

- Define sound types, such as shout, breath(sobbing), …
- Segment speech utterance into verbal and nonverbal segments.
- Extract sound type features
**Problem – Emotion Change in a Conversation**

**Problem:**
- There are different degree of emotion expression in different time periods within a speaking turn, so it should be a sequential emotion result to characterize an utterance.

**Proposed method:**
- We extract emotion type and sound type features for each segment of input utterance.
- Use LSTM-based Seq-to-Seq model to obtain sequential emotion recognition result.
NNIME (NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus)

- Audio, video, and ECG data
- Spontaneous emotional speech
- Recorded by 44 speakers
- 6 types of emotion scenario, 101 sessions, 673.02 mins (11.22 hrs)

<table>
<thead>
<tr>
<th>Emotion type</th>
<th>Angry</th>
<th>Frustration</th>
<th>Happy</th>
<th>Neutral</th>
<th>Sad</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sessions</td>
<td>15</td>
<td>19</td>
<td>15</td>
<td>18</td>
<td>18</td>
<td>16</td>
</tr>
</tbody>
</table>

Example of scenario setting

**Emotion: Angry**

Scenario setting: Before going out in the morning, the woman wanted to clean the house while the man was in a hurry. Later, the woman delayed again because she lost some stuff. The man was very angry while the woman was also mad with the man’s temper.
Data Analysis

- **Verbal data**
  - 7 types of emotions

- **Nonverbal data**
  - 3 human sound types + silence

<table>
<thead>
<tr>
<th>Sound Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shout</td>
<td>shout, scream, howl</td>
</tr>
<tr>
<td>Laughter</td>
<td>laugh, giggle</td>
</tr>
<tr>
<td>Breathing</td>
<td>sigh, yawn, sob, respire</td>
</tr>
<tr>
<td>Silence</td>
<td>silence, noise, audience sound</td>
</tr>
<tr>
<td>Verbal</td>
<td>speech</td>
</tr>
</tbody>
</table>

Emotion types: (high) Happy, (nervous, excited) Surprise, (fear, frustration) Anxiety, (low) Boring (tired, relax), Neutral, (+) Angry, (-) Sad
Data Statistics

- We segmented all sessions in NNIME into 4766 single speaker dialogue turns.
- Number of segments: 14636, duration = 4.3 hr (15492.5 secs, $\mu = 3.25$, $\sigma = 5.42$).

### All

<table>
<thead>
<tr>
<th>Emotion type</th>
<th>Anger</th>
<th>Anxiety</th>
<th>Sadness</th>
<th>Surprise</th>
<th>Neutral</th>
<th>Boring</th>
<th>Happy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment number</td>
<td>900</td>
<td>1090</td>
<td>415</td>
<td>1136</td>
<td>5212</td>
<td>537</td>
<td>753</td>
<td>14636</td>
</tr>
</tbody>
</table>

### Verbal segments

<table>
<thead>
<tr>
<th>Emotion type</th>
<th>Anger</th>
<th>Anxiety</th>
<th>Sadness</th>
<th>Surprise</th>
<th>Neutral</th>
<th>Boring</th>
<th>Happy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment number</td>
<td>863</td>
<td>1032</td>
<td>317</td>
<td>1068</td>
<td>5080</td>
<td>491</td>
<td>533</td>
<td>9384</td>
</tr>
</tbody>
</table>

### Nonverbal segments

<table>
<thead>
<tr>
<th>Sound type</th>
<th>Laugh</th>
<th>Breath</th>
<th>Shout</th>
<th>Silence</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment number</td>
<td>183</td>
<td>409</td>
<td>67</td>
<td>4593</td>
<td>5252</td>
</tr>
</tbody>
</table>
System Framework

Data Segmentation
- Silence Detection
- NNIME emotion corpus
- Verbal/Nonverbal Detection Model Training
  - SVM-based Verbal/Nonverbal Model
  - PPh Detection
  - Nonverbal Interval
  - Verbal Interval

Feature Extraction
- Sound Type Model Training
- Emotion Type Model Training
- Sound type CNN Model
  - Emotion Type CNN Model

Emotion Recognition
- Emotion Model Training
- LSTM-based Emotion Model

Training Phase
- Audio File
- Verbal/Nonverbal Segmentation
- PPh Detection

Testing Phase
- Output Emotion Sequence
Prosodic Phrase Annotation

- Annotate Prosodic Phrase based on the following criteria using *Praat*:
  - Pause (silence for more than 0.3 second)
  - Final rising intonation (Rising F0)
  - Lengthening of last word
  - Sharp fall in intensity (Falling intensity)
  - Modified wrong annotation of silence interval
Audio Data Segmentation

- **Silence interval detection**: produced by *Praat*

- **Verbal/ Non-verbal Segmentation**:
  1. Extract frame-based 384-dim audio feature by *openSMILE* [F. Eyben et al.]
  2. Calculate probability sequence of verbal/non-verbal frames by SVM
  3. Smoothing the probability sequence and compute boundary score

\[
\delta(P) = | \sum_{i=1}^{3} (4 - i)^2 \times P[b - i] - \sum_{i=1}^{3} (4 - i)^2 \times P[b + i] |
\]

4. If boundary score > threshold, set it as a boundary.

- **Prosodic Phrase Detection**: PPh detected by *PPh Autotagger* [Domínguez et al., 2016a]
Feature Vector for each Segment

- Using raw waveforms as input of CNN. [Bertero et al., 2017]
- 4 sound types and 7 emotion types
- The last hidden layer output is used as feature vector for recognition.

\[
X_i = S_i \oplus E_i
\]

Emotion feature/Sound type Vector

\[
\begin{align*}
E_i & : \text{the hidden vector of emotion feature extraction CNN} \\
S_i & : \text{the hidden vector of sound type classification CNN}
\end{align*}
\]

\[
i = \begin{cases} 
\text{the } i_{th} \text{ pure verbal segment} & \text{in training phase} \\
\text{the } i_{th} \text{ pure nonverbal segment} & \text{in training phase} \\
\text{the } i_{th} \text{ segment of input audio} & \text{in testing phase}
\end{cases}
\]

Input Vector for Seq2seq Emotion Recognition

Feature vector of audio segment → Seq2seq Emotion Recognition Model
Attentive Bi-LSTM based Seq-to-Seq Model

- The sound features for nonverbal segment and emotion features for each segment were adopted as feature vector $X_i$ to feed to the LSTM based Seq-to-Seq emotion recognition model with attention.

\[
X_i = S_i \oplus E_i \quad i = 1, \ldots, N
\]

$N = \text{number of segments in the utterance}$
Experimental Results - Evaluation on verbal/nonverbal segmentation

- 300 dialog turns from each pre-specified emotion and duration range were manually labeled for evaluation.
  - Features with dimensionalities of 32 and 384 were selected with window sizes of 100ms and 200ms and a shift size of 50ms.
  - A boundary is labeled correctly if the detected label is within 100ms of the manually labeled time.
  - The precision, recall, F1 score was used for evaluation:
    - F is feature dimension
    - W is the window size
    - S is the shift size
    - FM (100ms) is full match
    - PM (200ms) is partial match

<table>
<thead>
<tr>
<th></th>
<th>F = 32</th>
<th></th>
<th>F = 32</th>
<th></th>
<th>F = 384</th>
<th></th>
<th>F = 384</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W = 100</td>
<td>S = 50</td>
<td>W = 200</td>
<td>S = 50</td>
<td>W = 100</td>
<td>S = 50</td>
<td>W = 200</td>
<td>S = 50</td>
</tr>
<tr>
<td></td>
<td>FM</td>
<td>PM</td>
<td>FM</td>
<td>PM</td>
<td>FM</td>
<td>PM</td>
<td>FM</td>
<td>PM</td>
</tr>
<tr>
<td>0.6 Pre</td>
<td>0.24</td>
<td>0.46</td>
<td>0.23</td>
<td>0.47</td>
<td>0.34</td>
<td>0.62</td>
<td>0.31</td>
<td>0.57</td>
</tr>
<tr>
<td>0.6 Rec</td>
<td>0.31</td>
<td>0.60</td>
<td>0.29</td>
<td>0.58</td>
<td>0.37</td>
<td>0.65</td>
<td>0.36</td>
<td>0.66</td>
</tr>
<tr>
<td>0.6 F1</td>
<td>0.27</td>
<td>0.52</td>
<td>0.25</td>
<td>0.51</td>
<td>0.35</td>
<td>0.63</td>
<td>0.33</td>
<td>0.61</td>
</tr>
<tr>
<td>0.8 Pre</td>
<td>0.24</td>
<td>0.46</td>
<td>0.23</td>
<td>0.47</td>
<td>0.37</td>
<td>0.66</td>
<td>0.32</td>
<td>0.53</td>
</tr>
<tr>
<td>0.8 Rec</td>
<td>0.31</td>
<td>0.60</td>
<td>0.28</td>
<td>0.57</td>
<td>0.37</td>
<td>0.64</td>
<td>0.36</td>
<td>0.64</td>
</tr>
<tr>
<td>0.8 F1</td>
<td>0.27</td>
<td>0.52</td>
<td>0.25</td>
<td>0.51</td>
<td>0.37</td>
<td>0.64</td>
<td>0.34</td>
<td>0.61</td>
</tr>
<tr>
<td>1 Pre</td>
<td>0.25</td>
<td>0.48</td>
<td>0.23</td>
<td>0.49</td>
<td>0.38</td>
<td>0.67</td>
<td>0.33</td>
<td>0.59</td>
</tr>
<tr>
<td>1 Rec</td>
<td>0.30</td>
<td>0.59</td>
<td>0.27</td>
<td>0.56</td>
<td>0.35</td>
<td>0.60</td>
<td>0.35</td>
<td>0.62</td>
</tr>
<tr>
<td>1 F1</td>
<td>0.27</td>
<td>0.53</td>
<td>0.25</td>
<td>0.51</td>
<td>0.36</td>
<td>0.63</td>
<td>0.34</td>
<td>0.60</td>
</tr>
<tr>
<td>1.2 Pre</td>
<td>0.26</td>
<td>0.50</td>
<td>0.23</td>
<td>0.50</td>
<td>0.41</td>
<td>0.69</td>
<td>0.35</td>
<td>0.61</td>
</tr>
<tr>
<td>1.2 Rec</td>
<td>0.30</td>
<td>0.58</td>
<td>0.26</td>
<td>0.55</td>
<td>0.32</td>
<td>0.54</td>
<td>0.34</td>
<td>0.58</td>
</tr>
<tr>
<td>1.2 F1</td>
<td>0.28</td>
<td>0.54</td>
<td>0.24</td>
<td>0.52</td>
<td>0.36</td>
<td>0.61</td>
<td>0.34</td>
<td>0.59</td>
</tr>
</tbody>
</table>
Experimental Results - Evaluation on Feature Extraction

- This work selected a number of filters and different sizes in the adaptive pooling layer based on the accuracy of emotion classification.
- The results of comparison between the methods using raw speech signal and extracted acoustic feature sets were obtained.
- Performance of emotion type classification

<table>
<thead>
<tr>
<th>Input</th>
<th>Best parameters</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech signal</td>
<td>Filter number = 100, Kernel size = 512, step = 256, pooling = 2</td>
<td>30.10%</td>
</tr>
<tr>
<td>32-dim LLDs</td>
<td>Filter number = 150, Kernel size = 2, step = 1, pooling = 2</td>
<td>26.10%</td>
</tr>
<tr>
<td>32-dim LLDs with 12 functionals</td>
<td>Filter number = 100, Kernel size = 2, step = 1, pooling = 10</td>
<td>21.20%</td>
</tr>
</tbody>
</table>
Performance of sound type classification

The last hidden layer outputs of the CNN emotion/sound models were concatenated and fed to the LSTM-based sequence-to-sequence model for emotion recognition.

<table>
<thead>
<tr>
<th>Input</th>
<th>Best parameters</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech signal</td>
<td>Filter number = 100, Kernel size = 512, step = 256, pooling = 2</td>
<td>54.90%</td>
</tr>
<tr>
<td>32-dim LLDs</td>
<td>Filter number = 100, Kernel size = 2, step = 1, pooling = 2</td>
<td>53.63%</td>
</tr>
<tr>
<td>32-dim LLDs with 12 functionals</td>
<td>Filter number = 250, Kernel size = 2, step = 1, pooling = 10</td>
<td>47.95%</td>
</tr>
</tbody>
</table>

Experimental Results - Evaluation of Feature Extraction

The last hidden layer outputs of the CNN emotion/sound models were concatenated and fed to the LSTM-based sequence-to-sequence model for emotion recognition.
The hidden layer sizes of the LSTM were selected from 32, 64, 128, 256, and 512 to achieve the highest accuracy of emotion recognition.

- The proposed method achieved 52.00% when the hidden size of the LSTM was set to 128.

This work compared the performance of the proposed method with traditional emotion recognition models with frame-based acoustic features or raw speech signal as input.

<table>
<thead>
<tr>
<th>Input</th>
<th>Best parameters</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method</td>
<td>CNN-based feature extraction</td>
<td>Hidden size = 128</td>
</tr>
<tr>
<td>LSTM</td>
<td>32-dim LLDs</td>
<td>Hidden size = 256</td>
</tr>
<tr>
<td>CNN</td>
<td>Speech signal</td>
<td>Pooling = 2, filter number = 100</td>
</tr>
</tbody>
</table>
Conclusion

- Speech emotion recognition considering nonverbal interval and types of sound achieved a better performance.
- Sequence-to-sequence model can characterize emotional change in a dialogue turn.

Discussion

- Emotion expression in spontaneous speech is very diverse and difficult to be labeled with one specific emotion.
- The other difficulty of spontaneous speech emotion recognition is the background noise. Preprocessing of audio data is an important issue.
- There are still many sound types in our daily conversation. The types of emotional sound event should be better defined.
Result Demo – Inside
Result Demo – Outside

- These audios are from NNIME sessions which are used for training.
Questions?