Exploring the Intersection Between Speaker Verification and Emotion Recognition

Michelle Bancroft, Reza Lotfian, John Hansen, and Carlos Busso
Recognizing emotional speech is an important research problem
- Security and defense
- Quality control in customer service
- Human computer interaction

Goal: Create effective methods for retrieving emotional data from known speakers
- Arousal
- Valence
- Dominance

Relevant problem for forensic analysis

Retrieve angry sentences from “Joe”!
Combining emotional recognition with speaker verification tasks

- Use a speaker verification system to identify new sentences from target speakers
- Use emotion recognition model to predict arousal, valence and dominance for sentences
Infrastructure for the study

- Lack of naturalness
- Limited in size
- Limited number of speakers
- Unbalanced emotional content

<table>
<thead>
<tr>
<th>Database</th>
<th>Corpus Size</th>
<th># Spkr.</th>
<th>Type</th>
<th>Lang.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEMOCAP</td>
<td>12h26m</td>
<td>10</td>
<td>acted</td>
<td>English</td>
</tr>
<tr>
<td>MSP-IMPROV</td>
<td>9h35m</td>
<td>12</td>
<td>acted</td>
<td>English</td>
</tr>
<tr>
<td>CREMA-D</td>
<td>7,442 samples</td>
<td>91</td>
<td>acted</td>
<td>English</td>
</tr>
<tr>
<td>Chen Bimodal</td>
<td>9,900 samples</td>
<td>100</td>
<td>acted</td>
<td>English</td>
</tr>
<tr>
<td>Emo-DB</td>
<td>22m</td>
<td>10</td>
<td>acted</td>
<td>German</td>
</tr>
<tr>
<td>GEMEP</td>
<td>1,260 samples</td>
<td>10</td>
<td>acted</td>
<td>-</td>
</tr>
<tr>
<td>VAM-Audio</td>
<td>48m</td>
<td>47</td>
<td>spont.</td>
<td>German</td>
</tr>
<tr>
<td>TUM AVIC</td>
<td>10h23m</td>
<td>21</td>
<td>spont.</td>
<td>English</td>
</tr>
<tr>
<td>SEMAINE</td>
<td>6h21m</td>
<td>20</td>
<td>spont.</td>
<td>English</td>
</tr>
<tr>
<td>FAU-AIBO</td>
<td>9h12m</td>
<td>51</td>
<td>spont.</td>
<td>German</td>
</tr>
<tr>
<td>RECOLA</td>
<td>2h50m</td>
<td>46</td>
<td>spont.</td>
<td>French</td>
</tr>
</tbody>
</table>

Acted databases
Use existing podcast recordings
Divide into speaker turns
Emotion retrieval to balance the emotional content
Annotate using crowdsourcing framework
Collection of audio recordings (Podcasts)

- Naturalness and the diversity of emotions
- Creative Commons copyright licenses
- Interviews, talk shows, news, discussions, education, storytelling, comedy, science, technology, politics, economics, business, arts, culture, sports
**MSP-Podcast Corpus**

- **Automatic speaker diarization**
  - Single speaker segments
  - High SNR, no music, no phone quality

- **Duration:**
  - Longer than 2.75sec: Long enough for annotators + extract reliable features
  - Shorter than 11sec: Emotion content not changing significantly

---

“The speaker attribution intelligent service,” http://diarizationservice3.cloudapp.net/
MSP-Podcast Corpus

- Retrieve samples that convey desired emotion
  - Developing and optimizing different machine learning framework using existing databases
  - Balance the emotional content

Audio sharing website

1. Podcast Audio
2. 16kHz, 16b PCM, Mono
3. Diarization
4. Duration filter
   - 2.75s < ... < 11s
5. SNR filter
6. High quality audio

- Perceptual Evaluation
- Manual screening
- Emotion retrieval
- Remove telephone quality
- Speech only audio
- Music detection
- High quality audio

§ Retrieve samples that convey desired emotion
§ Developing and optimizing different machine learning framework using existing databases
§ Balance the emotional content
Perceptual evaluation

- Subjective annotation is costly
- screening only retrieved samples before uploading for annotations
Perceptual Evaluation

- Use Amazon Mechanical Turk Crowdsourcing
- Verify if a worker is spamming in real time

Trace performance in real time

MSP-Podcast corpus version 1.0

With emotion labels: 20,032 sentences (38h, 57m)

Segmented turns
152,975 sentences over 1,000 podcasts

Arousal

Valence

Dominance
The MSP-Podcast Corpus

- Manual annotations of data with emotional labels
  - 16,015 out of 20,032 speaking turns
- Target speakers
  - 146 speakers with at least 150s
- Audio repository: segments without emotional labels
  - 132,930 speaking turns
  - They include speech from target speakers

Ideal infrastructure with labeled data with both emotion and speaker information, but also a large unlabeled speech repository for the retrieval task.
Speaker Verification

- Speaker verification toolkit
- i-Vector Modeling

\[ M = m + T \mathbf{x} \]

GMM supervector
After MAP adaptation
Universal mean vector
Total variability matrix

- Mean normalized Probabilistic linear discriminant analysis (PLDA)

\[ \bar{x}_j = \frac{1}{D} \sum_{d=1}^{D} x^{(d)} \]

- The log-likelihood ratio (LLR)
  - The LLR computes the ratio between two alternative hypothesis:
  
  \[ r = \ln \frac{\rho(x_1, x_2|H_1)}{\rho(x_1|H_0) \cdot \rho(x_2|H_0)} \]

  \[ H_1 : x_1 \text{ and } x_2 \text{ same speaker} \]
  \[ H_2 : x_1 \text{ and } x_2 \text{ different speaker} \]
Multitask learning used to jointly predict arousal, valence and dominance [Parthasarathy and Busso, 2017]
- Two layers, training with CCC
- 6,373 acoustic features (OpenSmile)

Target Regions
- Region 1: low v, high a
- Region 2: high v, high a
- Region 3: low v, low a
- Region 4: high v, low a

Experimental Results

- Speaker verification
  - 146 target speakers
  - Train models with 150s per speaker
  - Test on the rest of the data from the 146 speakers
  - Threshold greater than $r_{\text{threshold}} = 12$
    - 33,628 unique segments identified

![Likelihood Threshold](image)

- Audio Repository
- Target Speakers
- Target Speakers + Target Emotion

- Experimental Results
  - 132,930 sentences
  - 33,628 unique segments
  - Score distribution
    - correct
    - incorrect
  - Likelihood Threshold $r_{\text{threshold}} = 12$
Emotion recognition

- We analyze 33,628 segments using our multitask learning framework
- 1,003 unique segments in the target regions
  - Region 1: 294
  - Region 2: 681
  - Region 3: 15
  - Region 4: 13
Analysis of Results

- **Emotion recognition**
  - Annotate segments with crowdsourcing
  - Precision rate

  Target region (45.8%)
  - Region 1: 37.5%
  - Region 2: 50.6%
  - Region 3: 23.1%
  - Region 4: 0%

  Target quadrant (77.4%)
  - Region 1: 73.3%
  - Region 2: 80.2%
  - Region 3: 61.5%
  - Region 4: 36.4%

Few samples in these regions

\[ \text{CCC}_{\text{arousal}} = 0.532 \]
\[ \text{CCC}_{\text{valence}} = 0.364 \]
Analysis of Results

- **Speaker verification**
  - 1,401 speaker verification evaluations satisfy ratio
    - A segments can have more than one speaker
  - We annotate the speaker identity of the 1,003 turns
    - 80.9% accuracy (1,135 evaluations correct)
    - Emotional speech challenges speaker verification systems
  - Speaker verification performance per target region
    - Region 1: 81.6%
    - Region 2: 80.9%
    - Region 3: 80.0%
    - Region 4: 33.3%
    - Few samples in these regions
Speaker verification and emotion recognition systems successfully combined

This study
- built the infrastructure to pursue this research direction
- revealed the limitations of speaker verification tasks in the presence of emotional speech

Future Work
- Further improve emotional recognition and speaker verification systems
- Compensation schemes for speaker verification systems in the presence of emotion
Thank you

- This work was funded by NSF CAREER award IIS-1453781

Interested on our research?
msp.utdallas.edu