Emotion Recognition in Spoken Dialog Systems

Spoken Dialog Systems
Unimodal Emotion Recognition

- HMM for recognizing emotions from speech signals
- Prosodic features extracted with the aid of Praat (http://www.praat.org)
- HMM training and recognizing using HTK (http://htk.eng.cam.ac.uk)
Dialog Systems

- Dialog flow defined by a set of rules and/or restrictions using VoiceXML
  - Only works for relatively small dialogs

Training a statistical model using collected dialog data

- Dialog modeled as a Markov Decision Process, or Partially observable MDPs
  - Adapted to incorporate affective states
  - Could be quite complex

- Derived statistical models using bi-turns, and tri-turns
  - the current state depends on at most two previous states
Speech Corpora: Acted Emotions

The Database of German Emotional Speech (TU Berlin)
- Six emotions: anger (Wut), boredom (Lengeweile), disgust (Ekel), fear (Angst), happiness (Freude), and sadness (Trauer)
- 10 actors, 5 male, 5 female
- Each utterance of 10 given utterances acted in all emotions
- 810 utterances (some doubles)
- Recordings in anechoic chamber

Quality of emotions were rated by 20 persons => quality of performance and ability to recognize emotions by humans were measured
Speech Corpora: Natural Speech

Ulm Database of English Spontaneous and Affective Speech

- Affective classes
  - Anger, boredom, happiness, sadness
  - Neutral, certainty, doubt

- 586 utterances, 12 speakers, 2 female, 10 male
- Interacting with a quiz and a personality test
- Free speech after an introductory question.
- In a secluded room without any background noise
- 5 Annotators => emotional states by majority and quality measures indicating the persuasiveness of the utterance
Emotion Recognition: Feature Extraction

36 Prosodic Features at intervals of 10ms

- Pitch
- Formants (1st, 2nd, and 3rd)
- Intensity
- Jitter (signal irregularity measure)
- Harmonicity
- Statistical values like: min, max, mean, deviation, range
Emotion Recognition: Feature Extraction

Pitch distribution of the female speakers for the different emotions in the acted data set.
Emotion Recognition: Feature Extraction

Pitch distribution of the male speakers for the different emotions in the acted data set.
Emotion Recognition: Language Models

- Sentence-emotion based language model
  
  \[
  ( \text{SENTENCE01-ANGER} \\
  | \text{SENTENCE01-BOREDOM} | \ldots \\
  | \text{SENTENCE10-SADNESS} )
  \]

  where “SENTENCE01-ANGER” is, e.g., “DER-ANGER LAPPEN-ANGER LIEGT-ANGER AUF-ANGER DEM-ANGER EISSCHRANK-ANGER”, etc.

- Word-emotion based language model

  \[
  \langle \text{ABEND-ANGER} | \text{ABEND-BOREDOM} \\
  | \ldots \\
  | \text{WOCHENENDEN-NEUTRAL} \\
  | \text{WOCHENENDEN-SADNESS} \rangle
  \]

- Word-emotion bi-grams

  adjacent words-emotion pairs (bi-grams). I.e., it includes the probabilities of, e.g., “LAPPEN-ANGER” preceding “LIEGT-HAPPINESS” which are summarized in a weighted word-emotion network.
## Emotion Evaluation

### Table 1  Average frequencies of errors

<table>
<thead>
<tr>
<th>emotions</th>
<th>AEF%</th>
<th>emotions</th>
<th>AEF%</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger &amp; happiness</td>
<td>6.66</td>
<td>disgust &amp; neutral</td>
<td>1.37</td>
</tr>
<tr>
<td>fear &amp; happiness</td>
<td>6.04</td>
<td>anger &amp; disgust</td>
<td>1.12</td>
</tr>
<tr>
<td>boredom &amp; neutral</td>
<td>5.03</td>
<td>happiness &amp; neutral</td>
<td>1.06</td>
</tr>
<tr>
<td>boredom &amp; sadness</td>
<td>3.22</td>
<td>disgust &amp; fear</td>
<td>0.87</td>
</tr>
<tr>
<td>fear &amp; neutral</td>
<td>2.56</td>
<td>boredom &amp; fear</td>
<td>0.83</td>
</tr>
<tr>
<td>sadness &amp; neutral</td>
<td>2.08</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>anger &amp; fear</td>
<td>1.60</td>
<td>anger &amp; sadness</td>
<td>0.00</td>
</tr>
<tr>
<td>happiness &amp; disgust</td>
<td>1.60</td>
<td>fear &amp; sadness</td>
<td>0.00</td>
</tr>
</tbody>
</table>
# Emotion Evaluation

<table>
<thead>
<tr>
<th>emotions</th>
<th>AEF%</th>
<th>emotions</th>
<th>AEF%</th>
</tr>
</thead>
<tbody>
<tr>
<td>fear &amp; happiness</td>
<td>8.47</td>
<td>happiness &amp; neutral</td>
<td>1.33</td>
</tr>
<tr>
<td>anger &amp; happiness</td>
<td>7.98</td>
<td>happiness &amp; sadness</td>
<td>0.59</td>
</tr>
<tr>
<td>sadness &amp; neutral</td>
<td>3.23</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>fear &amp; neutral</td>
<td>2.20</td>
<td>anger &amp; sadness</td>
<td>0.00</td>
</tr>
<tr>
<td>anger &amp; fear</td>
<td>1.60</td>
<td>fear &amp; sadness</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Speech-Emotion Recognition

- Previously emotion and speech were recognized at the same time => multiple HMMs per phoneme as we would have an HMM per phoneme per affective class
- 2-Step SE-Recognizer:
Multiple Recognizers

- Several different recognizers are used in a voting system
- The recognized words $W_i$ are output by the Word Transition Network $WTN_i$

The alignment module will align the recognized words

![Diagram showing alignment process]
Recognizer Output Voting Error Reduction (ROVER)
ROVER Results: Speech Recognition

Table 3  Word accuracies of different ROVER systems

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>ROVER-2</th>
<th>ROVER-3</th>
<th>ROVER-4</th>
<th>ROVER-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>88.7</td>
<td>87.8</td>
<td>89.6</td>
<td>89.2</td>
</tr>
<tr>
<td>0.25</td>
<td>88.7</td>
<td>87.8</td>
<td>89.7</td>
<td>88.8</td>
</tr>
<tr>
<td>0.5</td>
<td>88.7</td>
<td>88.0</td>
<td>89.7</td>
<td>89.1</td>
</tr>
<tr>
<td>0.75</td>
<td>88.7</td>
<td>87.9</td>
<td>89.7</td>
<td>89.2</td>
</tr>
<tr>
<td>1.0</td>
<td>89.1</td>
<td>88.3</td>
<td>89.7</td>
<td>89.2</td>
</tr>
</tbody>
</table>
# Rover Results: Emotion Recognition

## Table 4  Emotion accuracies of different ROVER systems

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>ROVER-2</th>
<th>ROVER-3</th>
<th>ROVER-4</th>
<th>ROVER-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>67.0</td>
<td>64.2</td>
<td>65.1</td>
<td>66.0</td>
</tr>
<tr>
<td>0.25</td>
<td>69.8</td>
<td>68.9</td>
<td>71.7</td>
<td>72.6</td>
</tr>
<tr>
<td>0.5</td>
<td>71.7</td>
<td>74.5</td>
<td>75.5</td>
<td>76.4</td>
</tr>
<tr>
<td>0.75</td>
<td>71.7</td>
<td>71.7</td>
<td>74.5</td>
<td>75.5</td>
</tr>
<tr>
<td>1.0</td>
<td>71.7</td>
<td>70.8</td>
<td>70.8</td>
<td>71.7</td>
</tr>
</tbody>
</table>
**Rover Results: Comparison**

**Table 5**  Average frequencies of errors of a stand-alone recognizer and a ROVER-5 system

<table>
<thead>
<tr>
<th>emotions</th>
<th>AEF% (S-A)</th>
<th>AEF% (R-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger &amp; happiness</td>
<td>7.55</td>
<td>5.66</td>
</tr>
<tr>
<td>fear &amp; happiness</td>
<td>6.42</td>
<td>4.72</td>
</tr>
<tr>
<td>boredom &amp; neutral</td>
<td>6.04</td>
<td>4.72</td>
</tr>
<tr>
<td>boredom &amp; sadness</td>
<td>3.96</td>
<td>1.89</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>anger &amp; sadness</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Adaptive Dialogs

Fig. 8  Valence-arousal diagram (André et al. 2004)
Adaptive Dialogs

User emotional state model

Dialog flow model

Combined model describing the dependencies of dialog flow and emotional state
Training the Stochastic Dialogue model

- 0.3 represent happy, 0.1 very happy, etc.

1: travel_from:0.3  travel_to:0.1
2: travel_to  travel_from
3: travel_to  travel_from  travel_date
4: travel_to:0.0  travel_from:0.0  travel_date:0.0
5: !travel_to:0.0  travel_date:0.4  travel_from:0.2
6: :0.0  :0.4  :0.3

...
Adaptive Dialog using Emotional States

S: “Welcome to the University of Ulm’s Virtual Travel Agency. How can I help you?”
U: “I would like to travel from Ulm to Rome.” [neutral]
S: “When do you want to depart?”
U: “Right after work, tomorrow in the evening” [happy]
S: “That’s great. There is a night train departing at 19:00 o’clock, would you like to reserve a bed in the sleeping car or a seat?”
U: “After ten hours of work, a bed would be fantastic.” [happy]

<field name="travel_from">
  <prompt value="'Where do you want to travel to?'"/>
  <prompt-0.5 value="'Great! So where will you go?'"/>
  <prompt-1.0 value="'Where do you want to travel to?'"/>
  <prompt-1.5 value="'I am sorry to bother you, but what is your destination?'"/>
</field>
Conclusions

- 39 MFCCs plus pitch, intensity, 3 formants, and statistics on these features combined with HMM based recognizer give good results.
- Bi-gram di-gram models for Speech-Emotion Dialogs replacing the more rigid (extended) VoiceXML model is a promising new direction.
- Semantics is seen as an important new direction for Adaptive Speech-Emotion Dialogs.