Emotion Recognition and Cognitive Load Measurement from Speech

Julien Epps, Fang Chen and Bo Yin

fang.chen@nicta.com.au

The material is mainly coming from the tutorial of APSIPA Annual Summit and Conference

December 14th, 2010
Overview

- Emotion and Mental State Recognition Systems
- Emotions and Cognitive Load
  - What are they and how can they be measured?
- Feature Extraction
  - Acoustic origins, methods, comparison and robustness
- Normalisation, Modelling and Classification
- Speech Databases and their Design
- Applications of Emotion Recognition and Cognitive Load Measurement
- Summary
Emotion and Mental State Recognition Systems
Emotion Recognition: Some History

• The Measurement of Emotion
  – Whately-Smith, 1922

• Emotion in speech (science)
  – Scherer, 1980s-present

• Emotion in synthetic speech
  – 1990s

• Emotion recognition
  – Dellaert et al., 1996 – prosodic features

• Applications of emotion recognition
  – Petrushin, 1999, Lee et al., 2001 – call centres

• Affective computing
  – Picard, 2000

• 2000: ~10 papers per year
Emotion Recognition: Some History

- Hidden Markov model emotion recognition
- Large review papers begin appearing
  - Cowie et al., 2003, Douglas Cowie et al., 2003
- HUMAINE
- Open source emotion recognition toolkit
- Large databases
  - LDC Emotional Prosody, AIBO Corpus, others
- Emotion recognition with large feature sets
- 2009: INTERSPEECH Emotion Challenge
- 2010: INTERSPEECH Paralinguistic Challenge
- 2010: >100 papers per year
Cognitive Load: Some History

• Working memory
  – Miller et al., 1960
  – Wickens 1980s, Baddeley 1990s onwards

• Cognitive load in learning
  – Sweller (mid 80s)

• Working memory and language
  – Gathercole, 1993

• Cognitive load in speech (science)
  – Berthold and Jameson, 1999

• Cognitive load classification from speech
  – Yin et al., 2008
• Typical acoustic based system

speech → voicing detection → feature extraction → feature normalisation

emotion modelling → classification / regression → recognized emotion
Emotion Recognition: Different Approaches

• Acoustic vs. prosodic vs. linguistic
• Detailed spectral vs. broad spectral features
• Direct vs. adapted models
• Static (utterance) vs. dynamic (frame/multi-frame)
• Categorical vs. ordinal vs. regression-based classification
• Recognition vs. detection
• Basis for emotion, mental states in physiology and hence speech production
  - Shared with expressive speech synthesis
  - Difficult to reverse-engineer spoken emotion
  - Use of wide range of physiological sensors may be helpful
    • EGG, glottal cameras, vocal tract x-ray movies etc
    • EEG, GSR, HRV, respiration
  - Complex: emotion dependency varies between different contexts (e.g. some words more emotive than others)
• Characterizing emotion, mental states using robust features
  - Lots of features in literature
  - Why should a given feature be / not be useful?
  - Are some features more sensitive to some emotions/mental states, and why?
  - How should temporal information best be used?
• Recognition of emotion in naturally occurring speech
  - Much initial work on emotion recognition done on acted speech
  - Initial work on cognitive load classification done on carefully controlled experimental data
  - Real emotions/metal states ≠ acted or experimental data
  - Taking emotion recognition from the lab to the field
• Understanding application areas
  – Lots of vision statements about affective computing
    • Detail still coming
  – Understanding likely limits of automatic approaches
    • Match approaches with suitable applications
  – Very large number of mental and cognitive states of interest in a wide range of different situations
  – New applications still emerging
Emotions and Cognitive Load
Emotions and CL: Psychology

• Psychology has much to say about emotions
  - Origins of emotion are in survival and evolution
    • Also a signalling system between people
    • Important in learning behaviours
  - Results of studies are complex to interpret as an engineer
  - There is no ‘standard’ set of emotions
  - Cultural variation exists
  - Emotions can often be discretely described only as ‘full-blown’ emotions
    • Cowie et al., 2001

• Pragmatics
  - Emotions of interest are dictated by application and availability of labelled data, usually single-culture, often fully blown.
• Emotion representations and structures
  – Affective circumplex
    - Arousal: alertness/responsiveness to stimuli
    - Valence: positive (e.g. happy) and negative (anger, fear) emotions

• Emotions are high-dimensional
  • Reducing to two dimensions can lead to different structures
    – Cowie et al., 2001
• One choice: FeelTrace
  - For real time
  - Cowie et al., 2000
• Measuring emotions
  - Self-reports of subjective experience
    • Single-item (“how happy did you feel”)
    • Check-list, e.g. mood adjective check-list
  - Real time methods
    • e.g. rotating dial / slider for single-item measure
    • e.g. arousal-affect emotion space cursor
  - Cued review
    • Subjects watch video of themselves performing the task, rate emotions experienced post-hoc
  - Observer ratings
  - Behavioural measures (this presentation)
  - Physiological measures (more later)
• Measuring CL
  - Subjective rating scales
    - NASA Task Load Index (Hart and Staveland, 1988)

source: http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf
Emotions and CL: Measuring CL

- Measuring CL
  - Subjective rating scales
  - Subjective workload assessment technique (Reid and Nygren, 1988)
  - Mental Effort Load
    » a. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.
    » b. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.
    » c. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.
• Measuring other mental states (speech literature)
  - Stress
  - Uncertainty
  - Level of interest
  - Deceptive speech
  - Mental disorders
    - Depression, anxiety, autism
  - Behaviour (e.g. drunkenness)
  - Disfluency (indicative of some mental states)

- Note differences in temporal occurrence of mental state
Feature Extraction
• Effect of emotion on speech production
  - Scherer, 1989
Feature Extraction: Origins

- Effect of emotion/mental state on speech production
  - Larsen and Fredrickson, 2003
  - Arousal component of emotion better conveyed using vocal cues than the affective component
  - Listeners can consistently nominate vocal cues for distinguishing emotions
  - Specific acoustic cues not generally linked to any particular emotions
    - Scherer, 1986
• Effect of emotion/mental state on speech production
  - Typical general acoustic feature set (Larsen and Fredrickson, 2003)
    • Fundamental frequency $F_0$
    • Fluctuation in $F_0$ due to jitter ($\Delta F_0$) and shimmer ($\Delta A_{F_0}$)
    • Energy
    • Speech rate
  - Stress related to (Streeter et al., 1983)
    • Increased energy
    • Increased pitch
Feature Extraction: Methods

• Voicing activity detection
  - Emotion, mental state content primarily conveyed during voiced portion of speech
  - Remove unvoiced/silence
    • Energy based methods
    • Pitch-based methods
    • VADs borrowed from other speech processing applications
      - Virtually nil additional attention

• Effect of cognitive load on speech production
  - Speech rate
  - Energy contour
  - $F_0$
  - Spectral parameters
    • Scherer et al., 2002
Feature Extraction: Methods

- Energy

Neutral speech

Log energy

Angry speech

Log energy
• Spectral features
  - Tilt
    • Many systems
    • ‘Broad’ spectral measure
  - Weighted frequency
    • ‘Broad’ spectral measure
  - Mel Frequency Cepstral Coefficients (MFCC)
    • Detailed spectral measure
  - Log frequency power coefficients (LFPC) (Nwe et al., 2003)
    • Detailed spectral measure
  - Group delay from all pole spectrum
    • Captures bandwidth information (Sethu et al., 2007)
  - Formant frequencies
  - Recent: Spectral Centroid features
Feature Extraction: Methods

• Glottal features
  – Glottal formants $F_g$, $F_c$ from glottal flow spectrum
    • Sethu, 2009
  – Primary open quotient (OQ1), normalised amplitude quotient (NAQ), primary speed quotient (SQ1)
    • Airas and Alku, 2006, Yap et al., 2010
  – Voice quality parameters
    • Lugger and Yang, 2007
  – Glottal statistics
    • Iliev and Scordilis, 2008
  – Vocal fold opening
    • Murphy and Laukkanen, 2010
  – Glottal flow spectrum (depressed speech)
    • Ozdas et al., 2004, others
Feature Extraction: Methods

- Glottal features (example)
  - Yap et al., 2010
  - High load ↔ creaky voice quality

![Graphs showing glottal features across different load levels](image)
Feature Extraction: Empirical Comparison

- Individual features (LDC Emotional Prosody, 5-class)

<table>
<thead>
<tr>
<th>Features</th>
<th>Feature Dimension</th>
<th>Speaker Independent</th>
<th>Speaker Dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>13</td>
<td>49.7 %</td>
<td>74.8 %</td>
</tr>
<tr>
<td>Formant Frequencies (FF)</td>
<td>6</td>
<td>43.7 %</td>
<td>58.3 %</td>
</tr>
<tr>
<td>Reflection Coefficients (RC)</td>
<td>24</td>
<td>48.9 %</td>
<td>71.2 %</td>
</tr>
<tr>
<td>Pitch (P)</td>
<td>1</td>
<td>46.6 %</td>
<td>51.8 %</td>
</tr>
<tr>
<td>Intensity/Energy (E)</td>
<td>1</td>
<td>28.8 %</td>
<td>25.2 %</td>
</tr>
<tr>
<td>Energy Slope (S)</td>
<td>1</td>
<td>43.4 %</td>
<td>59.0 %</td>
</tr>
<tr>
<td>Zero Crossing Rate (Z)</td>
<td>1</td>
<td>47.1 %</td>
<td>46.8 %</td>
</tr>
<tr>
<td>Spectral Centroid (SC)</td>
<td>1</td>
<td>40.2 %</td>
<td>44.6 %</td>
</tr>
<tr>
<td>Phoneme Rate (PhR)</td>
<td>1</td>
<td>20.6 %</td>
<td>23.0 %</td>
</tr>
<tr>
<td>GFCC</td>
<td>13</td>
<td>55.6 %</td>
<td>72.6 %</td>
</tr>
<tr>
<td>LP based Group Delay (GD)</td>
<td>10</td>
<td>42.9 %</td>
<td>69.8 %</td>
</tr>
<tr>
<td>Wavelet Scale Feature (WS)</td>
<td>1</td>
<td>41.8 %</td>
<td>54.7 %</td>
</tr>
<tr>
<td>LPRCC</td>
<td>13</td>
<td>50.0 %</td>
<td>68.4 %</td>
</tr>
<tr>
<td>Frequency Modulation (FM)</td>
<td>30</td>
<td>44.4 %</td>
<td>64.0 %</td>
</tr>
<tr>
<td>FM + GFCC</td>
<td>43</td>
<td>47.1 %</td>
<td>73.4 %</td>
</tr>
<tr>
<td>Weighted Frequency (WF)</td>
<td>3</td>
<td>47.6 %</td>
<td>53.2 %</td>
</tr>
<tr>
<td>Fractal Dimension (FD)</td>
<td>1</td>
<td>46.3 %</td>
<td>41.0 %</td>
</tr>
</tbody>
</table>

Sethu, 2009
Feature Extraction: Empirical Comparison

- Top individual features (12 actors, 14 emotions)
  - Scherer, 1996
  - Pitch
    - Mean, standard deviation, 25\textsuperscript{th} and 75\textsuperscript{th} percentile
  - Energy – mean
  - Speech rate
  - Long-term voiced average spectrum
    - 125-200, 200-300, 500-600, 1000-1600, 5000-8000 Hz
  - Hammarberg index
    - Difference in energy maxima in 0-2 and 2-5 kHz bands (voiced)
  - Spectral slope above 1 kHz
  - Voicing energy up to 1 kHz
  - Long-term unvoiced average spectrum
    - 125-200, 5000-8000 Hz
• Individual features (LDC Emotional Prosody, 5-class)
  - Sethu, 2009

<table>
<thead>
<tr>
<th>Features</th>
<th>Speaker Dependent</th>
<th>Speaker Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Without Normalisation</td>
</tr>
<tr>
<td>SSC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitch + Energy (PE)</td>
<td>51.1 %</td>
<td>37.6 %</td>
</tr>
<tr>
<td>DSM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>74.8 %</td>
<td>48.7 %</td>
</tr>
<tr>
<td>Formant Feature (FF)</td>
<td>58.3 %</td>
<td>35.7 %</td>
</tr>
<tr>
<td>Reflection Coefficients (RC)</td>
<td>71.2 %</td>
<td>41.8 %</td>
</tr>
<tr>
<td>LP based Group Delay (GD)</td>
<td>69.8 %</td>
<td>36.0 %</td>
</tr>
<tr>
<td>BSM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy Slope + ZCR (SZ)</td>
<td>57.6 %</td>
<td>40.2 %</td>
</tr>
<tr>
<td>Weighted Frequency (WF)</td>
<td>53.2 %</td>
<td>40.7 %</td>
</tr>
</tbody>
</table>
• Shifted delta coefficients
  - Feature $C_l$ ($l$ = feature dimension), at time $t$:
    \[
    \Delta C^t_l = C^t_{l+iP+d} - C^t_{l+iP-d}
    \]
  - $k$: how many delta features
  - $P$: spacing between features
  - $d$: controls span of each delta

  - Important in cognitive load (Yin et al., 2008)
    - MFCC 52.2%
    - MFCC+Prosodic (Concatenated) 59.3%
    - MFCC+Prosodic, Acceleration 64.4%
    - MFCC+Prosodic, SDC 65.7%
    - Much larger performance difference in other CL research ~20-60% impr.
**Feature Extraction: Temporal Information**

- **Parameter Contours**
  - Pitch: Sethu, 2009
    - 5-class, LDC Emotional Prosody

<table>
<thead>
<tr>
<th>Classification Test</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human – Actual Speech</td>
<td>63.6 %</td>
</tr>
<tr>
<td>Automatic – using pitch contours (slope-bias)</td>
<td>57.1 %</td>
</tr>
<tr>
<td>Automatic – using glottal parameter contours (midpoint)</td>
<td>55.0 %</td>
</tr>
<tr>
<td>Automatic – using vocal tract parameter contours (midpoint)</td>
<td>45.0 %</td>
</tr>
<tr>
<td>Automatic – using all model parameters</td>
<td>62.6 %</td>
</tr>
</tbody>
</table>
### Feature Extraction: Empirical Comparison

- **Comparison of Cognitive Load Features**
  - 3-class, Stroop Test Database

<table>
<thead>
<tr>
<th>Feature (combination)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>52.2%</td>
</tr>
<tr>
<td>MFCC + SDC</td>
<td>60.2%</td>
</tr>
<tr>
<td>MFCC + Pitch + Energy + SDC</td>
<td>79.2%</td>
</tr>
<tr>
<td>MFCC + Pitch + Energy + 3 Glottal features + SDC</td>
<td>84.4%</td>
</tr>
<tr>
<td>F1 + F2 + F3 + SDC</td>
<td>67.7%</td>
</tr>
<tr>
<td>SCF + SCA + SDC</td>
<td>87.2%</td>
</tr>
</tbody>
</table>
Feature Extraction: Prosodic

• Emotion
  - Duration, rate, pause, pitch, energy features
    • Based on ASR alignment
    • Ang et al., 2002
  - Pitch level and range, speech tempo, loudness
    • More details: Pitch contour slope, stressed/unstressed syllables
    • Emotional speech synthesis
    • Schröder, 2001

• Cognitive load
  - Duration
    • Speech is slower under higher cognitive load
    • Pitch, formant trajectories less variable (more monotone)
    • Yap et al., 2009
Feature Extraction: Linguistic

- Emotion
  - Fused linguistic explored in a few papers
  - e.g. 10-30% impr.
    Lee et al., 2002
  - Right: Schüller et al. 2007

<table>
<thead>
<tr>
<th>feature set</th>
<th>full</th>
<th>reduced</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td>#</td>
<td>F_{SVM}</td>
</tr>
<tr>
<td>Low Level Descriptors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voice quality</td>
<td>153</td>
<td>51.5</td>
</tr>
<tr>
<td>F0</td>
<td>333</td>
<td>56.1</td>
</tr>
<tr>
<td>spectral/formants</td>
<td>656</td>
<td>54.4</td>
</tr>
<tr>
<td>cepstral</td>
<td>1699</td>
<td>52.7</td>
</tr>
<tr>
<td>wavelets</td>
<td>216</td>
<td>56.0</td>
</tr>
<tr>
<td>energy</td>
<td>265</td>
<td>58.5</td>
</tr>
<tr>
<td>duration</td>
<td>391</td>
<td>55.1</td>
</tr>
<tr>
<td>all acoustic</td>
<td>3713</td>
<td>57.7</td>
</tr>
<tr>
<td>disfluencies</td>
<td>4</td>
<td>26.8</td>
</tr>
<tr>
<td>non-verbals</td>
<td>8</td>
<td>24.8</td>
</tr>
<tr>
<td>part of speech</td>
<td>31</td>
<td>54.7</td>
</tr>
<tr>
<td>higher semantics</td>
<td>12</td>
<td>57.6</td>
</tr>
<tr>
<td>bag of words</td>
<td>476</td>
<td>62.6</td>
</tr>
<tr>
<td>all linguistic</td>
<td>531</td>
<td>62.6</td>
</tr>
<tr>
<td>all</td>
<td>4244</td>
<td>61.0</td>
</tr>
</tbody>
</table>

functionals (without linguistic features)

<table>
<thead>
<tr>
<th>feature set</th>
<th>full</th>
<th>reduced</th>
</tr>
</thead>
<tbody>
<tr>
<td>percentiles</td>
<td>1196</td>
<td>53.8</td>
</tr>
<tr>
<td>specific</td>
<td>153</td>
<td>54.5</td>
</tr>
<tr>
<td>extremes</td>
<td>1132</td>
<td>53.4</td>
</tr>
<tr>
<td>higher stat. mom.s</td>
<td>547</td>
<td>57.6</td>
</tr>
<tr>
<td>means</td>
<td>427</td>
<td>59.8</td>
</tr>
<tr>
<td>sequential+comb.</td>
<td>218</td>
<td>61.2</td>
</tr>
<tr>
<td>all functional</td>
<td>3673</td>
<td>57.4</td>
</tr>
</tbody>
</table>
• Higher cognitive load
  - Sentence fragments, false starts and errors increase
  - Onset latency, pauses increase
  - Articulation/speech rate decrease
    • Berthold and Jameson, 1999
  - Measures of language complexity – complexity increases for multiple measures
  - More use of plural pronouns, less of singular
    • Khawaja et al., 2009

  - Linguistic not fused with acoustic to date
• Disfluencies
  – Interruption rate
  – Proportion of the effective speech in the whole speech period
  – Keywords for correction or repeating
• Inter-sentential pausing
  – Length and frequency of big pauses
• Fragmented sentences
  – Length and frequency of small pauses
  – Length of intra-sentence segments
• Slower speech rate
  – Syllable rate
• Response Latency
  – Delay of generating speech
  – Particular hybrid prosodic pattern
Normalisation, Modelling and Classification
• Two main sources of variability:
  - Phonetic
  - Speaker identity

• Both stronger sources of variability

• Two approaches to deal with these:
  - Model it
    - e.g. UBM-GMM with many mixtures
  - Remove it
    - e.g. normalisation

• Other variability:
  - Phonetic/linguistic dependence of emotion information
  - Session
  - Age
Normalisation

- Types
  - Per utterance
  - Per speaker

Sethu et al., 2007
Modelling

- Usual considerations
  - Structure of problem and/or database
  - Amount of training data available → number of parameters to use

- Rate of classification decision
  - Often one per utterance

- Type of result needed
  - Categorical e.g. {“happiness”, ”boredom”,...}
  - Continuous e.g. [-1,1]
  - Ordinal e.g. {“low”, ”medium”, ”high”}
• Gaussian mixture model (GMM)

\[ p(x) = \sum_{m=1}^{M} w_m \frac{1}{(2\pi)^{K/2}|C_m|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_m)^T C_m^{-1}(x - \mu_m)\right) \]

- \( w_m \): weight of the \( m \)'th mixture
- \( \mu_m \): mean vector of the \( m \)'th mixture
- \( C_m \): covariance matrix of the \( m \)'th mixture
• Maximum a priori GMM
  - Previously: Train each GMM separately for each target class
  - MAP-GMM begins with a single model
    • Covering wide range of conditions (emotions/CL)
      - Acoustic space of all speakers
      - Trained on very large database
    • “Universal background” model (UBM) – from speaker recognition
  - Advantages
    • Mixtures in each emotion class with same index correspond to same acoustic class
    • Fast scoring

UBM-GMM
Adapted model: happiness
Adapted model: sadness
• Hidden Markov model (HMM)
  - Can capture temporal variation
    • Main issue: what to model?
  - Schüller et al., 2003: pitch, energy contours
  - Nwe et al., 2003: LFPC features, F0, speaking rate
  - Sethu et al., 2009: pitch trajectories in utterance
    • Assumes emotions have same trajectory structure in all utterances

Nwe et al., 2003
Larger utterance variability than Schüller et al., 2003?
• Hidden Markov model (HMM)
  - Lee et al., 2004
  - Phoneme class based states
    • Vowel/glide/nasal/stop/fricative
    • Clear what a state represents, no assumptions about utterance structure
  - Adapt towards each emotion

<table>
<thead>
<tr>
<th>Classification Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVC with prosodic features</td>
<td>55.68</td>
</tr>
<tr>
<td>generic “emotional” HMM</td>
<td>64.77</td>
</tr>
<tr>
<td>every phoneme class</td>
<td>75.57</td>
</tr>
<tr>
<td>vowel only</td>
<td>72.16</td>
</tr>
<tr>
<td>glide only</td>
<td>54.86</td>
</tr>
<tr>
<td>nasal only</td>
<td>47.43</td>
</tr>
<tr>
<td>stop only</td>
<td>44.89</td>
</tr>
<tr>
<td>fricative only</td>
<td>55.11</td>
</tr>
<tr>
<td>Combination of prosody and phoneme-class classifier</td>
<td>76.12</td>
</tr>
</tbody>
</table>
• GMMs describe common patterns as well as class-specific patterns.
• SVM identify the boundary between classes, ignore the common distributions
Classification: Empirical Results

- LDC Emotional Prosody
- Static classifiers
- Sethu et al., 2009

<table>
<thead>
<tr>
<th>Classifier</th>
<th>P + E + WF</th>
<th>MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>43.1%</td>
<td>37.1%</td>
</tr>
<tr>
<td>PNN</td>
<td>44.1%</td>
<td>35.6%</td>
</tr>
<tr>
<td>SVM</td>
<td>40.5%</td>
<td>39.9%</td>
</tr>
</tbody>
</table>

- Berlin Emotional Speech Database (EMO-DB)
- Schüller et al., 2005

<table>
<thead>
<tr>
<th>Accuracy [%]</th>
<th>All 276 features included</th>
<th>Top 75 by SVM SFFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>73.57</td>
<td>73.98</td>
</tr>
<tr>
<td>1NN</td>
<td>63.52</td>
<td>75.82</td>
</tr>
<tr>
<td>SVM</td>
<td>84.84</td>
<td>87.50</td>
</tr>
<tr>
<td>C4.5</td>
<td>61.07</td>
<td>61.48</td>
</tr>
<tr>
<td>Bagged C4.5</td>
<td>70.70</td>
<td>74.80</td>
</tr>
<tr>
<td>AdaBoosted C4.5</td>
<td>72.34</td>
<td>74.59</td>
</tr>
<tr>
<td>MultiBoosted C4.5</td>
<td>72.54</td>
<td>74.59</td>
</tr>
<tr>
<td>StackingC MLR NB 1NN C4.5</td>
<td>75.41</td>
<td>79.92</td>
</tr>
<tr>
<td>StackingC MLR NB 1NN SVM C4.5</td>
<td>76.23</td>
<td>80.53</td>
</tr>
</tbody>
</table>
Classification: Human Performance

• Emotion
  - Dellaert et al., 1996
  - ~17% for 7-class speaker dependent (independent: 35%)
  - Schüller et al., 2005
  - ~60% acc. for 8-class

<table>
<thead>
<tr>
<th>Category</th>
<th>happy</th>
<th>sad</th>
<th>anger</th>
<th>fear</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>happy</td>
<td>44</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>sad</td>
<td>1</td>
<td>40</td>
<td>3</td>
<td>6</td>
<td>5%</td>
</tr>
<tr>
<td>anger</td>
<td>2</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>1%</td>
</tr>
<tr>
<td>fear</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>32</td>
<td>9%</td>
</tr>
</tbody>
</table>

  Total error: 18%

• Cognitive load
  - Le, 2009
  - 62% acc. overall

<table>
<thead>
<tr>
<th>Recognized CL level</th>
<th>Actual CL level</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>76</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Medium</td>
<td>2</td>
<td>60</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>10</td>
<td>23</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>
Emotional Speech Databases and their Design
• Emotion
  - Emotions sufficiently fully blown to be distinguishable
  - Naturalness
  - Listener ratings
  - Matched to context of use

• Cognitive load
  - Multiple discrete load levels
  - Levels distinguishable by subjective ratings
  - Possibly additional sensor data (physiological)
  - No stress involved, subjects motivated

• All
  - Large, multiple speaker
  - Phonetically, linguistically diverse
Databases: Design

• Scope
  - 4/5 ‘basic’ emotions oversimplifies the problem

• Naturalness
  - Read speech vs. uncontrolled phonetic/linguistic content

• Context
  - Allow use of emotive words?

• Descriptors
  - Elicited emotions → well defined descriptors
  - Natural emotions: Use multiple levels of descriptors or Feeltrace + listener ratings
  - Douglas-Cowie et al., 2003
• AIBO Corpus
  - 9 hours, 51 children, spontaneous (WOz) speech, 11 categories
  - [http://www5.informatik.uni-erlangen.de/our-team/steidl-stefan/fau-aibo-emotion-corpus/](http://www5.informatik.uni-erlangen.de/our-team/steidl-stefan/fau-aibo-emotion-corpus/) 50€

• IEMOCAP
  - 12 hours, 10 speakers, acted speech, 9 classes + valence/activation/dominance labelling
  - [http://sail.usc.edu/iemocap/](http://sail.usc.edu/iemocap/)

• LDC Emotional Prosody
  - ½ hour, 7 speakers, acted speech, 15 classes
  - [http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28](http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28) US$2500
Databases: Resources (English)

• Reading-Leeds
  - 4½ hours, unscripted TV interviews

• Cognitive Load
  - DRIVAWORK
    - 14 hours, in-car speech under varying workload, several physiological reference signals
  - SUSAS / SUSC-1
    - contains some psychological stress conditions
    - http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78 US$500
  - Other databases proprietary

• More complete listing
  - See Douglas-Cowie et al., 2003; Ververidis and Kotropoulos, 2006
• Cognitive load task design: Span tests
  – Word span, digit span
    • Interleave memory tasks presentation of target words/digits with demanding secondary task
      – Conway et al., 2005
  – Reading span
    • Ability to coordinate processing and storage resources
    • Subjects must read fluently and answer comprehension questions
      – Daneman and Carpenter, 1980
  – Speaking span
    • Randomly selected words
    • Speaking span = maximum number of words for which a subject can successfully construct a grammatically correct sentence
      – Daneman, 1991
  – Others: counting, operation, reading digit
Databases: CL Examples

• Stroop test
  - Subjects presented colour words in different colours
    • Low: congruent    Medium: incongruent
  - Time pressure added for high load

• Story reading
  - Story reading followed by Q&A
  - 3 different levels of text difficulty (Lexile Framework for Reading, www.lexile.com)
  - 3 stories in each of the 2 sessions (fixed order)
    • 1st session:
      - “Sleep” (900L),
      - “History of Zero” (1350L) &
      - “Milky Way Galaxy” (1400L)
    • 2nd session:
      - “Smoke Detectors” (950L),
      - “Hurricanes” (1250L) &
      - “The Sun” (1300L)
• ‘Naturally’ elicited emotion
• Larger databases
• Telephone speech
• Realistic data (i.e. without controlled lab conditions)
• Later: noise
Applications of Emotion Recognition and Cognitive Load Measurement
• Like other speech processing apps
  – Need microphone
  – Best in clean environments

• Trend towards processing power in the loop
  – VoIP
  – Mobile phone / iPhone / SmartPhone

• Trend towards user-aware interfaces
  – Affective computing

• The work overload problem
  – Mental processes play an increasing role in determining the workload in most jobs (Gaillard et al., 1993)
• Adjusting the human-computer balance for optimised task performance
  – Ideas date back to Yerkes and Dodson (1908)
    • Arousal induced using mild electric shocks
• Care needed: cognitive load $\propto 1/\text{performance}$
  - Task performance says nothing about spare mental capacity
  - Spare or residual capacity turns out to be an important practical measure
    • Parasuraman et al., 2008

  - Correlation between workload and performance is poorer in underload and overload tasks
    • Vidulich and Wickens, 1986; Yeh and Wickens, 1988

  - No studies for physiological or behavioural measures
Applications: Examples

• From the literature . . .
  - Air traffic control, vehicle handling, shipping, military, rail
    • Embrey et al., 2006
  - Learning and vigilance tasks
    • Sweller, 1988; Berka et al., 2007
  - Driving
    • Lamble et al., 1999
  - Team decision efficiency
    • Johnston et al., 2006
  - Collaboration load
    • Dillenbourg and Betrancourt, 2006
  - Augmented cognition
    • Schmorrow, 2005
Applications: Biomedical

• Brain training for ADHD
  - Improve working memory
  - [www.cogmed.com](http://www.cogmed.com)
  - Supported by substantial research base

• Diagnosis/treatment of mental disorders
  - Objective measure
  - e.g. Workman et al., 2007
  - Still to be investigated

• Management of psychological stress
  - Has occupational health and safety implications
  - Still to be investigated
Cognitive Load and Emotion
Emotion Recognition – Speech Pitch Contour

Emotional Speech → Pitch Estimation → VAD → Segmented Pitch Contour

Segmental Linear Approximation

PITCH – LINEAR APPROXIMATION

\[ \text{tan} \theta = \text{slope, } s \]

Segment Length, \( x \) → Initial Offset, \( b \)

Parameters – \( b, s, x \)

HMM BASED EMOTION MODELS
• Ideally, want to compare emotion recognition and CL systems on same data, but:
  – No ground truth
    • Classification becomes clustering
  – What should systems look like?
    • Could try to build a ‘common’ system, but
      – Where and how to compromise?
  – Which data?
    • Almost by definition, there is no database that contains both emotion and cognitive load that is also usefully annotated
• What we did
  – Took two near-state-of-the-art systems
    • Relative to LDC Emotional Prosody and
    • Stroop Test corpora
  – Applied each system to both corpora
    • Use closed-set training/test
  – Produced a set of emotion and cognitive load labels for each corpus
  – Compared clusterings based on labels
• Adjusted Rand Index

\[
Adjusted\ Index = \frac{Index - Expected\ Index}{Max\ Index - Expected\ Index}
\]

\[
ARI(U, V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}
\]

- Adjusts RI by subtracting expected value
- \( ARI = 1 \) means clustering identical
- \( ARI \approx 0 \) means random labeling
Emotion vs. CL: Clustering comparison

- Emotional Prosody Speech Corpus (LDC)
  - 5 emotions
    - Neutral
    - Anger
    - Sadness
    - Happiness
    - Boredom

- Stroop Test Data
  - 3 CL levels

- Results
  - LDC Emotional Prosody
    - $ARI = 0.0087$
  - Stroop Test Corpus
    - $ARI = 0.0253$
Conclusion

• Data
  - Plenty of acted, experimental databases as reference
  - Natural data are current challenge

• Features
  - Broad vs. detailed spectral features: implications for modelling
  - Temporal information needs to be captured
  - Centroid features (together) are promising

• Classification
  - Speaker variability is a very strong effect
  - Select correct classification paradigm for problem

• Applications
  - Emerging, many still to be well defined
Acknowledgements

• 8 Project staff (NICTA)
  – Prof Eliathamby Ambikairajah, Dr Fang Chen, Dr Julien Epps, Dr Bo Yin, Dr Eric Choi, Dr Ronnie Taib, Dr Natalie Ruiz, Dr Yang Wang

• 11 PhD students (UNSW/NICTA)
  – Tet Fei Yap, Phu Ngoc Le, Asif Khawaja, Vidhyasaharan Sethu, Karen Kua, Tharmarajah Thiruvaran
  – Pega Zarjam, Kun Yu, Siyuan Chen, Guanzhong Li, Sazzad Hassan

• Collaborators
  – UNSW Speech Processing Laboratory, Prof Paul Corballis (Georgia Tech, US), Prof John Sweller (UNSW)