CS 424P/ LINGUIST 287
Extracting Social Meaning and Sentiment

Dan Jurafsky

Lecture 6: Emotion
In the last 20 years

- A huge body of research on emotion
- Just one quick pointer: Ekman: basic emotions:
Ekman’s 6 basic emotions
Surprise, happiness, anger, fear, disgust, sadness
Disgust  Anger  Sadness  Happiness

Fear  Surprise
Dimensional approach.  

- **Arousal**
  - **High arousal**
    - Displeasure (e.g., anger)
  - **Low arousal**
    - Displeasure (e.g., sadness)

- **Valence**
  - **High pleasure** (e.g., excitement)
  - **Low pleasure** (e.g., relaxation)

Slide from Julia Braverman
Distinctive vs. Dimensional approach of emotion

**Distinctive**
- Emotions are units.
- Limited number of basic emotions.
- Basic emotions are innate and universal
- Methodology advantage
  - Useful in analyzing traits of personality.

**Dimensional**
- Emotions are dimensions.
- Limited # of labels but unlimited number of emotions.
- Emotions are culturally learned.
- Methodological advantage:
  - Easier to obtain reliable measures.

Slide from Julia Braverman
Four Theoretical Approaches to Emotion:

1. Darwinian (natural selection)

- Darwin (1872) *The Expression of Emotion in Man and Animals*. Ekman, Izard, Plutchik
- Function: Emotions evolve to help humans survive
- Same in everyone and similar in related species
  - Similar display for Big 6+ (happiness, sadness, fear, disgust, anger, surprise) → ‘basic’ emotions
  - Similar understanding of emotion across cultures

The particulars of fear may differ, but "the brain systems involved in mediating the function are the same in different species" (LeDoux, 1996)

extended from Julia Hirschberg’s slides discussing Cornelius 2000
2. Jamesian: Emotion is experience

- William James 1884. What is an emotion?
  - Perception of **bodily changes** → emotion
    - “we feel sorry because we cry... afraid because we tremble”
    - “our feeling of the ... changes as they occur IS the emotion"
  - The body makes **automatic responses** to environment that help us survive
  - Our experience of these responses constitutes emotion.
  - Thus each emotion accompanied by **unique pattern** of bodily responses
    - Stepper and Strack 1993: emotions follow facial expressions or posture.
    - Botox studies:

extended from Julia Hirschberg’s slides discussing Cornelius 2000
Four Theoretical Approaches to Emotion:

3. Cognitive: Appraisal

- An emotion is produced by appraising (extracting) particular elements of the situation. (Scherer)
  - **Fear**: produced by the appraisal of an event or situation as obstructive to one’s central needs and goals, requiring urgent action, being difficult to control through human agency, and lack of sufficient power or coping potential to deal with the situation.
  - **Anger**: difference: entails much higher evaluation of controllability and available coping potential
  - **Guilt**: appraising a situation as unpleasant, as being one's own responsibility, but as requiring little effort.

Adapted from Cornelius 2000
Four Theoretical Approaches to Emotion:

4. Social Constructivism

- Emotions are cultural products (Averill)
- Explains gender and social group differences
- **anger** is elicited by the appraisal that one has been wronged intentionally and unjustifiably by another person. Based on a moral judgment
  - don’t get angry if you yank my arm accidentally
  - or if you are a doctor and do it to reset a bone
  - only if you do it on purpose

Adapted from Cornelius 2000
Scherer’s typology of affective states

- **Emotion**: relatively brief episode of synchronized response of all or most organismic subsytems in response to the evaluation of an external or internal event as being of major significance
  - *angry, sad, joyful, fearful, ashamed, proud, desperate*

- **Mood**: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause
  - *cheerful, gloomy, irritable, listless, depressed, buoyant*

- **Interpersonal stance**: affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation
  - *distant, cold, warm, supportive, contemptuous*

- **Attitudes**: relatively unduring, affectively color beliefs, preference, predisposition towards objects or persons
  - *liking, loving, hating, valueing, desiring*

- **Personality traits**: emotionally laden, stable personality dispositions and behavior tendencies, typical for a person
  - *nervous, anxious, reckless, morose, hostile, envious, jealous*
### Scherer’s typology

<table>
<thead>
<tr>
<th>Type of affective state: brief definition (examples)</th>
<th>Intensity</th>
<th>Duration</th>
<th>Synchronization</th>
<th>Event focus</th>
<th>Appraisal elicitation</th>
<th>Rapidity of change</th>
<th>Behavioral impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotion: relatively brief episode of synchronized response of all or most organismic subsystems in response to the evaluation of an external or internal event as being of major significance (angry, sad, joyful, fearful, ashamed, proud, elated, desperate)</td>
<td>++ + + + + +</td>
<td>+</td>
<td>+ + + + +</td>
<td>+</td>
<td>+ + + +</td>
<td>+ + +</td>
<td>+ + +</td>
</tr>
<tr>
<td>Mood: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (cheerful, gloomy, irritable, listless, depressed, buoyant)</td>
<td>+ + +</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+ +</td>
<td>+</td>
</tr>
<tr>
<td>Interpersonal stances: affective stance taken toward another person in a specific interaction, colouring the interpersonal exchange in that situation (distant, cold, warm, supportive, contemptuous)</td>
<td>+ + +</td>
<td>+ + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+ + +</td>
<td>+ +</td>
</tr>
<tr>
<td>Attitudes: relatively enduring, affectively coloured beliefs, preferences, and predispositions towards objects or persons (liking, loving, hating, valuing, desiring)</td>
<td>0 + +</td>
<td>+ + + + +</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>0 +</td>
<td>+</td>
</tr>
<tr>
<td>Personality traits: emotionally laden, stable personality dispositions and behavior tendencies, typical for a person (nervous, anxious, reckless, morose, hostile, envious, jealous)</td>
<td>0 + +</td>
<td>+ + +</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+</td>
</tr>
</tbody>
</table>

0: low, +: medium, ++: high, +++: very high, −: indicates a range.
Why Emotion Detection from Speech or Text?

- Detecting frustration of callers to a help line
- Detecting stress in drivers or pilots
- Detecting “interest”, “certainty”, “confusion” in on-line tutors
  - Pacing/Positive feedback
- Lie detection
- Hot spots in meeting browsers
- Synthesis/generation:
  - On-line literacy tutors in the children’s storybook domain
  - Computergames
Hard Questions in Emotion Recognition

- How do we know what emotional speech is?
  - Acted speech vs. natural (hand labeled) corpora

- What can we classify?
  - Distinguish among multiple ‘classic’ emotions
  - Distinguish
    - Valence: is it positive or negative?
    - Activation: how strongly is it felt? (sad/despair)

- What features best predict emotions?

- What techniques best to use in classification?

Slide from Julia Hirschberg
Major Problems for Classification:
Different Valence/Different Activation

slide from Julia Hirschberg
But....

Different Valence/ Same Activation

slide from Julia Hirschberg
Accuracy of facial versus vocal cues to emotion (Scherer 2001)

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Anger</th>
<th>Fear</th>
<th>Joy</th>
<th>Sadness</th>
<th>Disgust</th>
<th>Surprise</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facial/Western/20</td>
<td>74</td>
<td>78</td>
<td>77</td>
<td>95</td>
<td>79</td>
<td>80</td>
<td>88</td>
<td>78</td>
</tr>
<tr>
<td>Vocal/Recent Western/11</td>
<td>59</td>
<td>64</td>
<td>38</td>
<td>28</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facial/Non-Western/11</td>
<td>70</td>
<td>64</td>
<td>38</td>
<td>28</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocal/Non-Western/1</td>
<td>70</td>
<td>64</td>
<td>38</td>
<td>28</td>
<td>58</td>
<td></td>
<td></td>
<td>52</td>
</tr>
</tbody>
</table>
Background: The Brunswikian Lens Model

- is used in several fields to study how observers correctly and incorrectly use objective cues to perceive physical or social reality

- cues have a probabilistic (uncertain) relation to the actual objects
- a (same) cue can signal several objects in the environment
- cues are (often) redundant

slide from Tanja Baenziger

[Diagram of a theoretical model involving concepts such as Trait/state, Externalization, Perceptual representation, Inferential utilization, Attribution, Distal indicator cues, Proximal percepts, Functional validity, Criterion value, Indicator values, Perceptual judgments, Attributional judgments, Operational level, Association coefficients, Representation coefficients, Utilization coefficients, and Accuracy coefficient.]
Expressed emotion

Emotional attribution

Important issues:
- To be measured, cues must be identified a priori
- Inconsistencies on both sides (indiv. diff., broad categories)
- Cue utilization could be different on the left and the right side (e.g. important cues not used)

Example:
- Vocal cues: Loud voice, High pitched
- Facial cues: Frown
- Gestures: Clenched fists, Shaking
- Other cues ...

Expressed anger?

Encoder

Decoder

slide from Tanja Baenziger
Implications for HMI

• If matching is low...

  - Generation: Conversational agent developers should focus on the relation of the cues to the perceived emotion
  - Recognition: Automatic recognition system developers should focus on the relation of the cues to expressed emotion

slide from Tanja Baenziger
Extroversion in Brunswikian Lens

- Simulated jury discussions in German and English
  - speakers had detailed personality tests
- Extroversion personality type accurately identified from naïve listeners by voice alone
- But not emotional stability
  - listeners choose: resonant, warm, low-pitched voices
  - but these don’t correlate with actual emotional stability
Data and tasks for Emotion Detection

- Scripted speech
  - Acted emotions, often using 6 emotions
  - Controls for words, focus on acoustic/prosodic differences
  - Features:
    - F0/pitch
    - Energy
    - Speaking rate

- Spontaneous speech
  - More natural, harder to control
  - Dialogue
  - Kinds of emotion focused on:
    - frustration,
    - annoyance,
    - certainty/uncertainty
    - “activation/hot spots”
Four quick case studies

- Acted speech: LDC’s EPSaT
- Annoyance and Frustration in Natural speech
  - Ang et al on Annoyance and Frustration
- Natural speech:
  - AT&T’s How May I Help You?
- Uncertainty in Natural speech:
  - Liscombe et al’s ITSPoke
Example 1: Acted speech; emotional Prosody
Speech and Transcripts Corpus (EPSaT)

- Recordings from LDC
- 8 actors read short dates and numbers in 15 emotional styles
EPSaT Examples

happy
sad
angry
confident
frustrated
friendly
interested

anxious
bored
encouraging

Slide from Jackson Liscombe
Detecting EPSaT Emotions

- Liscombe et al 2003
- Ratings collected by Julia Hirschberg, Jennifer Venditti at Columbia University
Liscombe et al. Features

- Automatic Acoustic-prosodic
  - [Davitz, 1964] [Huttar, 1968]
- Global characterization
  - pitch
  - loudness
  - speaking rate
Global Pitch Statistics
Global Pitch Statistics

- Happy (M=330, SD=109)
- Angry (M=350, SD=84)
Liscombe et al. Features

- **Automatic Acoustic-prosodic**
  [Davitz, 1964] [Huttar, 1968]

- **ToBI Contours**
  [Mozziconacci & Hermes, 1999]

- **Spectral Tilt**
  [Banse & Scherer, 1996] [Ang et al., 2002]
Liscombe et al. Experiment

- RIPPER 90/10 split
- Binary Classification for Each Emotion
- Results
  - 62% average baseline
  - 75% average accuracy
  - Acoustic-prosodic features for activation
  - /H-L%/ for negative; /L-L%/ for positive
  - Spectral tilt for valence?
Example 2 - Ang 2002

- Ang Shriberg Stolcke 2002 “Prosody-based automatic detection of annoyance and frustration in human-computer dialog”
- Prosody-Based detection of annoyance/ frustration in human computer dialog
- DARPA Communicator Project Travel Planning Data
  - NIST June 2000 collection: 392 dialogs, 7515 utts
  - CMU 1/2001-8/2001 data: 205 dialogs, 5619 utts
  - CU 11/1999-6/2001 data: 240 dialogs, 8765 utts
- Considers contributions of prosody, language model, and speaking style
- Questions
  - How frequent is annoyance and frustration in Communicator dialogs?
  - How reliably can humans label it?
  - How well can machines detect it?
  - What prosodic or other features are useful?
Data Annotation

- 5 undergrads with different backgrounds (emotion should be judged by ‘average Joe’).
- Labeling jointly funded by SRI and ICSI.
- Each dialog labeled by 2+ people independently in 1st pass (July-Sept 2001), after calibration.
- 2nd “Consensus” pass for all disagreements, by two of the same labelers (Oct-Nov 2001).
- Used customized Rochester Dialog Annotation Tool (DAT), produces SGML output.
Data Labeling

- **Emotion**: neutral, annoyed, frustrated, tired/disappointed, amused/surprised, no-speech/NA

- **Speaking style**: hyperarticulation, perceived pausing between words or syllables, raised voice

- **Repeats and corrections**: repeat/rephrase, repeat/rephrase with correction, correction only

- **Miscellaneous useful events**: self-talk, noise, non-native speaker, speaker switches, etc.

*Slide from Shriberg, Ang, Stolcke*
Emotion Samples

- **Neutral**
  - July 30
  - Yes

- **Disappointed/tired**
  - No

- **Amused/surprised**
  - No

- **Annoyed**
  - Yes
  - Late morning (HYP)

- **Frustrated**
  - Yes
  - No
  - No, I am ... (HYP)
  - There is no Manila...
## Emotion Class Distribution

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>17994</td>
<td>.831</td>
</tr>
<tr>
<td>Annoyed</td>
<td>1794</td>
<td>.083</td>
</tr>
<tr>
<td>No-speech</td>
<td>1437</td>
<td>.066</td>
</tr>
<tr>
<td>Frustrated</td>
<td>176</td>
<td>.008</td>
</tr>
<tr>
<td>Amused</td>
<td>127</td>
<td>.006</td>
</tr>
<tr>
<td>Tired</td>
<td>125</td>
<td>.006</td>
</tr>
<tr>
<td>TOTAL</td>
<td>21653</td>
<td></td>
</tr>
</tbody>
</table>

To get enough data, we grouped annoyed and frustrated, versus else (with speech)
Prosodic Model

- Used CART-style decision trees as classifiers
- Downsampled to equal class priors (due to low rate of frustration, and to normalize across sites)
- Automatically extracted prosodic features based on recognizer word alignments
- Used automatic feature-subset selection to avoid problem of greedy tree algorithm
- Used 3/4 for train, 1/4th for test, no call overlap
Prosodic Features

- **Duration and speaking rate features**
  - duration of phones, vowels, syllables
  - normalized by phone/vowel means in training data
  - normalized by speaker (all utterances, first 5 only)
  - speaking rate (vowels/time)

- **Pause features**
  - duration and count of utterance-internal pauses at various threshold durations
  - ratio of speech frames to total utt-internal frames
Prosodic Features (cont.)

- **Pitch features**
  - F0-fitting approach developed at SRI (Sönmez)
  - LTM model of F0 estimates speaker’s F0 range

- Many features to capture pitch range, contour shape & size, slopes, locations of interest
- Normalized using LTM parameters by speaker, using all utts in a call, or only first 5 utts
Features (cont.)

- **Spectral tilt features**
  - average of 1st cepstral coefficient
  - average slope of linear fit to magnitude spectrum
  - difference in log energies btw high and low bands
  - extracted from longest normalized vowel region

- **Other (nonprosodic) features**
  - position of utterance in dialog
  - whether utterance is a repeat or correction
  - to check correlations: hand-coded style features including hyperarticulation
Language Model Features

- Train 3-gram LM on data from each class
- LM used word classes (AIRLINE, CITY, etc.) from SRI Communicator recognizer
- Given a test utterance, chose class that has highest LM likelihood (assumes equal priors)
- In prosodic decision tree, use sign of the likelihood difference as input feature
- Finer-grained LM scores cause overtraining
## Results: Human and Machine

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Each Human with</td>
<td>71.7</td>
<td>.38</td>
</tr>
<tr>
<td>Other Human, overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human with Human</td>
<td>84.2</td>
<td>.68</td>
</tr>
<tr>
<td>“Consensus” (biased)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prosodic Decision</td>
<td>75.6</td>
<td>.51</td>
</tr>
<tr>
<td>Tree with Consensus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree with Consensus,</td>
<td>72.9</td>
<td>.46</td>
</tr>
<tr>
<td>no repeat/correction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree with Consensus,</td>
<td>68.7</td>
<td>.37</td>
</tr>
<tr>
<td>repeat/correction only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language Model</td>
<td>63.8</td>
<td>.28</td>
</tr>
<tr>
<td>features only</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results (cont.)

- H-H labels agree 72%, **complex decision task**
  - inherent continuum
  - speaker differences
  - relative vs. absolute judgements?
- H labels agree 84% with “consensus” (biased)
- Tree model agrees 76% with consensus-- *better than original labelers with each other*
- Prosodic model makes use of a dialog state feature, but without it it’s still better than H-H
- Language model features alone are not good predictors (dialog feature alone is better)
Predictors of Annoyed/Frustrated

- **Prosodic: Pitch features:**
  - high maximum fitted F0 in longest normalized vowel
  - high speaker-norm. (1st 5 utts) ratio of F0 rises/falls
  - maximum F0 close to speaker’s estimated F0 “topline”
  - minimum fitted F0 late in utterance (no “?” intonation)

- **Prosodic: Duration and speaking rate features**
  - long maximum phone-normalized phone duration
  - long max phone- & speaker- norm.(1st 5 utts) vowel
  - low syllable-rate (slower speech)

- **Other:**
  - utterance is repeat, rephrase, explicit correction
  - utterance is after 5-7th in dialog
Effect of Class Definition

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>Entropy Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline prosody model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus labels A,F vs. N,else</td>
<td>75.6</td>
<td>21.6</td>
</tr>
<tr>
<td><strong>Tokens on which labelers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>originally agreed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A,F vs. N,else</td>
<td>78.3</td>
<td>26.4</td>
</tr>
<tr>
<td><strong>All tokens</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus labels F vs. A,N,else</td>
<td>82.7</td>
<td>37.0</td>
</tr>
</tbody>
</table>

For less ambiguous tokens, or more extreme tokens performance is significantly better than baseline.
Ang et al ‘02 Conclusions

- Emotion labeling is a complex decision task
- Cases that labelers independently agree on are classified with high accuracy
  - Extreme emotion (e.g. ‘frustration’) is classified even more accurately
- Classifiers rely heavily on prosodic features, particularly duration and stylized pitch
  - Speaker normalizations help
- Two nonprosodic features are important: utterance position and repeat/correction
  - Language model is an imperfect surrogate feature for the underlying important feature repeat/correction
Example 3: “How May I Help You℠” (HMIHY)

- Giuseppe Riccardi, Dilek Hakkani-Tür, AT&T Labs
- Each turn in 20,000 turns (5690 dialogues) annotated for 7 emotions by one person
  - Positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative, very other negative
  - Distribution was so skewed (73.1% labeled positive/neutral)
  - So classes were collapsed to negative/nonnegative
- Task is hard!
  - Subset of 627 turns labeled by 2 people: kappa .32 (full set) and .42 (reduced set)!
User Emotion Distribution

Emotion Label Distribution

Slide from Jackson Liscombe
Emotion Labels

- **Original Set:**
  - Positive/Neutral
  - Somewhat Frustrated
  - Very Frustrated
  - Somewhat Angry
  - Very Angry
  - Other Somewhat Negative
  - Very Negative

- **Reduced Set:**
  - Positive
  - Negative

Slide from Jackson Liscombe
<table>
<thead>
<tr>
<th>Person</th>
<th>Transcription</th>
<th>Emotion</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Agent:</td>
<td>I'm sorry, I didn't understand that. How may I help you?</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>User:</td>
<td>please an operator please [ .hst ] may I speak-</td>
<td>Positive/Neutral</td>
<td>…</td>
</tr>
<tr>
<td>Agent:</td>
<td>I'm sorry we're having trouble. Can we try one more time? Just briefly tell me how I may help you.</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>User:</td>
<td>may I speak to an assistant please</td>
<td>Somewhat Frustrated</td>
<td>…</td>
</tr>
<tr>
<td>Agent:</td>
<td>Are you calling from your home phone?</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>User:</td>
<td>yes</td>
<td>Somewhat Angry</td>
<td>…</td>
</tr>
</tbody>
</table>
Pitch/Energy/Speaking Rate : Change

Median Pitch □ Mean Energy □ Speaking Rate

Z-score

Positive  Frustrated  Angry

Utterance

20020221/0221080552atf1536
Slide from Jackson Liscombe
Person | Transcription | Emotion | Play
---|---|---|---
Agent: | Hello, I am an AT&T automated assistant. You can speak naturally to me. How may I help you? | Positive/Neutral | 
User: | I need to find out about a number that I don't recognize | Positive/Neutral | 
Agent: | Would you like to look up a number you don't recognize on your bill? | Positive/Neutral | 
User: | yes I would | Positive/Neutral | 
Agent: | Are you calling from your home phone? | Positive/Neutral | 
User: | yes I am | Positive/Neutral |
Pitch/Energy/Speaking Rate: No Change

![Bar chart showing the comparison of median pitch, mean energy, and speaking rate for positive utterances.](Slide from Jackson Liscombe 20020317/0317220739atf9051)
HMIHY Features

- Automatic Acoustic-prosodic
  - [Cauldwell, 2000]
- Contextual
  - [Schröder, 2003] [Brennan, 1995]
- Transcriptive
  - [Ang et al., 2002] [Lee & Narayanan, 2005]
- Pragmatic
Lexical Features

- Language Model (ngrams)
- Examples of words significantly correlated with negative user state \((p<0.001)\):
  - 1\textsuperscript{st} person pronouns: ‘I’, ‘me’
  - billing-related words: ‘dollars’, ‘cents’
  - curse words: ...
Prosodic Features

- **Pitch (F0)**
  1. Overall minimum
  2. overall maximum
  3. overall median
  4. overall standard deviation
  5. mean absolute slope
  6. slope of final vowel
  7. longest vowel mean

- **Other**
  8. local jitter over longest vowel

- **Energy**
  9. overall minimum
  10. overall maximum
  11. overall mean
  12. overall standard deviation
  13. longest vowel mean

- **Speaking Rate**
  14. vowels per second
  15. mean vowel length
  16. ratio voiced frames to total frames
  17. percent internal silence
Contextual Features

- **Lexical (2)**
  - edit distance with previous 2 turns

- **Discourse (10)**
  - turn number
  - call type repetition with previous 2 turns
  - dialog act repetition with previous 2 turns

- **Prosodic (34)**
  - $1^{st}$ and $2^{nd}$ order differentials for each feature

- **Other (2)**
  - user state of previous 2 turns
HMIHY Experiment

• Classes: *Negative vs. Non-negative*
  • Training size = 15,013 turns
  • Testing size = 5,000 turns

• Most frequent user state (*positive*) accounts for 73.1% of testing data

• Learning Algorithm Used:
  • BoosTexter
    • (boosting w/ weak learners)
  • continuous/discrete features
  • 2000 iterations

• Results:

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>73%</td>
</tr>
<tr>
<td>Acoustic-prosodic</td>
<td>75%</td>
</tr>
<tr>
<td>+ transcriptive</td>
<td>76%</td>
</tr>
<tr>
<td>+ pragmatic</td>
<td>77%</td>
</tr>
<tr>
<td>+ contextual</td>
<td>79%</td>
</tr>
</tbody>
</table>
Intelligent Tutoring Spoken Dialogue System

- (ITSpoke)
- Diane Litman, Katherine Forbes-Riley, Scott Silliman, Mihai Rotaru, University of Pittsburgh, Julia Hirschberg, Jennifer Venditti, Columbia University
58. Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys?

The keys will rise above the man’s face because the same gravitational force is being applied to both, yet the man’s mass is greater than the mass of the keys so he will fall faster than the keys.
Task 1

- Negative
  - Confused, bored, frustrated, uncertain
- Positive
  - Confident, interested, encouraged
- Neutral
um <sigh> I don’t even think I have an idea here ...... now .. mass isn’t weight ...... mass is ............... the .......... space that an object takes up ........ is that mass?
PROBLEM (TYPED): If a car is able to accelerate at 2 m/s², what acceleration can it attain if it is towing another car of equal mass?

ESSAY (TYPED): The maximum acceleration a car can reach when towing a car behind it of equal mass will be halved. Therefore, the maximum acceleration will be 1m/s².

DIALOGUE (SPOKEN): ... 9.1 min. into session ...

TUTOR₁: Uh let us talk of one car first.

STUDENT₁: ok. (EMOTION = NEUTRAL)

TUTOR₂: If there is a car, what is it that exerts force on the car such that it accelerates forward?

STUDENT₂: The engine (EMOTION = POSITIVE)

TUTOR₃: Uh well engine is part of the car, so how can it exert force on itself?

STUDENT₃: um... (EMOTION = NEGATIVE)
Acoustic-Prosodic Features

- 4 normalized fundamental frequency (f0) features: maximum, minimum, mean, standard deviation
- 4 normalized energy (RMS) features: maximum, minimum, mean, standard deviation
- 4 normalized temporal features: total turn duration, duration of pause prior to turn, speaking rate, amount of silence in turn

Non-Acoustic-Prosodic Features

- lexical items in turn
- 6 automatic features: turn begin time, turn end time, isTemporalBarge-in, isTemporalOverlap, #words in turn, #syllables in turn
- 6 manual features: #false starts in turn, isPriorTutorQuestion, isQuestion, isSemanticBarge-in, #canonical expressions in turn, isGrounding

Identifier Features: subject, subject gender, problem
Liscombe et al: ITSpoke Experiment

- Human-Human Corpus
- AdaBoost(C4.5) 90/10 split in WEKA
- Classes: *Uncertain vs Certain vs Neutral*
- Results:

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>66%</td>
</tr>
<tr>
<td>Acoustic-prosodic</td>
<td>75%</td>
</tr>
<tr>
<td>+ contextual</td>
<td>76%</td>
</tr>
<tr>
<td>+ breath-groups</td>
<td>77%</td>
</tr>
</tbody>
</table>
Some summaries re: Prosodic features

<table>
<thead>
<tr>
<th></th>
<th>Stress</th>
<th>Anger/rage</th>
<th>Fear/panic</th>
<th>Sadness</th>
<th>Joy/elation</th>
<th>Boredom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
</tr>
<tr>
<td>F0 floor/mean</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td></td>
</tr>
<tr>
<td>F0 variability</td>
<td>↗️</td>
<td></td>
<td></td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
</tr>
<tr>
<td>F0 range</td>
<td>↗️</td>
<td>↗️(.timezone)</td>
<td></td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
</tr>
<tr>
<td>Sentence contours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>↗️</td>
<td></td>
</tr>
<tr>
<td>High frequency energy</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td></td>
<td>(↗️)</td>
<td></td>
</tr>
<tr>
<td>Speech and articulation rate</td>
<td>↗️</td>
<td>↗️</td>
<td>↗️</td>
<td></td>
<td>(↗️)</td>
<td>↗️</td>
</tr>
</tbody>
</table>
### Juslin and Laukka metastudy

#### Speech rate

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Fast</th>
<th>Medium</th>
<th>Slow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>(1, 4, 6, 9, 13, 16, 18, 19, 24, 27, 28, 42, 48, 55, 64, 69, 70, 71, 72, 80, 82, 83, 89, 92, 97, 99, 100, 103)</td>
<td>(51, 63, 75)</td>
<td>(6, 10, 47, 53)</td>
</tr>
<tr>
<td>Fear</td>
<td>(3, 4, 9, 10, 13, 16, 18, 28, 33, 47, 48, 51, 63, 64, 71, 72, 80, 82, 83, 89, 95, 96, 97, 103)</td>
<td>(1, 53, 70)</td>
<td>(12, 42)</td>
</tr>
<tr>
<td>Happiness</td>
<td>(4, 6, 9, 10, 16, 19, 20, 24, 33, 37, 42, 53, 55, 71, 72, 75, 80, 82, 83, 97, 99, 100)</td>
<td>(13, 18, 51, 89, 92)</td>
<td>(1, 3, 48, 64, 70, 95)</td>
</tr>
<tr>
<td>Sadness</td>
<td>(95)</td>
<td>(1, 51, 53, 63, 70)</td>
<td>(2, 3, 4, 7, 8, 9, 10, 13, 15, 16, 18, 19, 20, 21, 22, 23, 25, 26, 27, 29, 34, 37, 39)</td>
</tr>
<tr>
<td>Tenderness</td>
<td>(4)</td>
<td>(24, 33, 96)</td>
<td></td>
</tr>
</tbody>
</table>

#### Tempo

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Fast</th>
<th>Medium</th>
<th>Slow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>(2, 4, 7, 8, 9, 10, 13, 14, 16, 18, 19, 20, 22, 23, 25, 26, 29, 34, 37, 41)</td>
<td>(27)</td>
<td>(2, 8, 10, 23, 25, 26, 27, 32, 37)</td>
</tr>
<tr>
<td>Fear</td>
<td>(18, 20, 22)</td>
<td>(7, 19)</td>
<td>(1, 2, 3, 4, 7, 8, 9, 12, 13, 14, 15, 16, 18, 19, 20, 22, 23, 26, 27, 29, 34, 37, 41)</td>
</tr>
<tr>
<td>Happiness</td>
<td>(18, 20, 22)</td>
<td>(7, 19)</td>
<td>(1, 2, 3, 4, 7, 8, 9, 10, 13, 15, 16, 18, 19, 20, 21, 22, 23, 25, 26, 27, 29, 32, 34, 37)</td>
</tr>
<tr>
<td>Sadness</td>
<td>(25, 32)</td>
<td>(25, 32)</td>
<td>(25, 32)</td>
</tr>
<tr>
<td>Tenderness</td>
<td>(1, 2, 7, 8, 10, 13, 14, 16, 19, 27, 41)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Voice intensity (M)

<table>
<thead>
<tr>
<th>Emotion</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>(1, 3, 6, 8, 9, 10, 18, 24, 27, 33, 42, 43, 44, 47, 50, 51, 55, 56, 61, 63, 64, 70, 72, 82, 89, 91, 95, 97, 99, 101)</td>
<td>(4)</td>
<td>(79)</td>
</tr>
<tr>
<td>Fear</td>
<td>(3, 4, 6, 8, 10, 18, 45, 47, 50, 63, 79, 82, 89)</td>
<td>(70, 95, 97)</td>
<td>(9, 10, 33, 42, 43, 51, 56, 64)</td>
</tr>
<tr>
<td>Happiness</td>
<td>(3, 4, 6, 8, 10, 18, 45, 47, 50, 51, 52, 56, 72, 82, 88, 89, 97, 99, 101)</td>
<td>(18, 42, 55, 70, 91, 95)</td>
<td>(95)</td>
</tr>
<tr>
<td>Sadness</td>
<td>(79)</td>
<td>(4, 95)</td>
<td>(4, 95)</td>
</tr>
<tr>
<td>Tenderness</td>
<td>(1, 2, 24, 33, 95)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Sound level (M)

<table>
<thead>
<tr>
<th>Emotion</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>(2, 5, 7, 8, 9, 13, 14, 15, 16, 18, 19, 20, 23, 25, 26, 27, 29, 32, 34, 35, 36, 41)</td>
<td>(27)</td>
<td>(2, 7, 13, 18, 19, 20, 22, 23, 25, 26, 32)</td>
</tr>
<tr>
<td>Fear</td>
<td>(1, 2, 7, 13, 14, 26, 41)</td>
<td>(5, 8, 16, 18, 19, 20, 22, 23, 25, 27, 29, 32, 34)</td>
<td>(9)</td>
</tr>
<tr>
<td>Happiness</td>
<td>(5, 8, 16, 18, 19, 20, 22, 23, 25, 27, 29, 32, 34)</td>
<td>(18, 25)</td>
<td>(5, 7, 8, 9, 13, 15, 16, 19, 20, 21, 22, 23, 26, 27, 29, 32, 34, 36)</td>
</tr>
<tr>
<td>Sadness</td>
<td>(9)</td>
<td>(5, 18, 25)</td>
<td>(1, 7, 8, 12, 13, 14, 16, 19, 27, 41)</td>
</tr>
<tr>
<td>Tenderness</td>
<td>(1, 7, 8, 12, 13, 14, 16, 19, 27, 41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voice intensity variability</td>
<td>Sound level variability</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Anger</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(51, 61, 64, 70, 80, 82, 89, 95, 101)</td>
<td>High (14, 29, 41)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>(3)</td>
<td>Medium (23, 34)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(10, 53)</td>
<td>Low (8, 13, 23, 37)</td>
<td></td>
</tr>
<tr>
<td><strong>Fear</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(10, 64, 68, 80, 82, 83, 95)</td>
<td>High (1, 14, 34)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>(3, 51, 53, 70)</td>
<td>Medium (41)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(89)</td>
<td>Low (8, 23, 29, 37)</td>
<td></td>
</tr>
<tr>
<td><strong>Happiness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(3, 10, 53, 64, 80, 82, 95, 101)</td>
<td>High (23)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>(51, 70, 89)</td>
<td>Medium (29, 34)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(47, 83)</td>
<td>Low (1, 14, 41)</td>
<td></td>
</tr>
<tr>
<td><strong>Sadness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(10, 89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>(53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(3, 51, 61, 64, 70, 82, 95, 101)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tenderness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>High-frequency energy</th>
<th>High-frequency energy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anger</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(4, 6, 16, 18, 24, 33, 35, 38, 42, 47, 50, 51, 54, 56, 63, 65, 73, 82, 83, 91, 97, 103)</td>
<td>High (8, 10, 13, 15, 16, 19, 20, 22, 26, 36, 37)</td>
</tr>
<tr>
<td>Medium</td>
<td>(4, 6, 16, 18, 24, 33, 35, 38, 42, 47, 50, 51, 54, 56, 63, 65, 73, 82, 83, 91, 97, 103)</td>
<td>Medium (10, 37)</td>
</tr>
<tr>
<td>Low</td>
<td>(16, 38, 42, 50, 51, 82)</td>
<td>Low (15, 19, 20, 22, 26, 36)</td>
</tr>
<tr>
<td><strong>Fear</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(18, 33, 54, 56, 63, 83, 97, 103)</td>
<td>High (15)</td>
</tr>
<tr>
<td>Medium</td>
<td>(4, 6)</td>
<td>Medium (8, 10, 12, 13, 16, 19, 20, 22, 26, 36)</td>
</tr>
<tr>
<td>Low</td>
<td>(47, 83)</td>
<td>Low (37)</td>
</tr>
<tr>
<td><strong>Happiness</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(4, 42, 50, 51, 52, 54, 56, 63, 73, 82, 83, 88, 91, 97)</td>
<td>High (8, 10, 12, 13, 16, 19, 20, 22, 26, 36)</td>
</tr>
<tr>
<td>Medium</td>
<td>(6, 18, 24)</td>
<td>Medium (8, 10, 12, 13, 16, 19, 20, 22, 26, 36)</td>
</tr>
<tr>
<td>Low</td>
<td>(83)</td>
<td>Low (37)</td>
</tr>
<tr>
<td><strong>Sadness</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(4, 10, 14, 41)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>(10, 23, 29, 37)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(8, 13, 23, 37)</td>
<td></td>
</tr>
<tr>
<td><strong>Tenderness</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>(4, 24, 33)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>(4, 10, 14, 41)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(8, 13, 23, 37)</td>
<td></td>
</tr>
<tr>
<td>Emotion</td>
<td>FO (M)</td>
<td>FO (M)²</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>Anger</strong></td>
<td><strong>High</strong> (3, 6, 10, 14, 16, 19, 24, 27, 29, 32, 34, 36, 45, 46, 48, 51, 53, 61, 63, 64, 65, 71, 72, 73, 74, 79, 82, 83, 95, 97, 99, 101, 103)</td>
<td><strong>Sharp</strong> (26, 37)</td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (4, 33, 50, 70, 75)</td>
<td><strong>Precise</strong> (32)</td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong> (18, 43, 44, 53, 89)</td>
<td><strong>Flat</strong> (8)</td>
</tr>
<tr>
<td><strong>Fear</strong></td>
<td><strong>High</strong> (1, 4, 6, 12, 16, 17, 18, 29, 43, 45, 47, 48, 50, 53, 60, 63, 65, 70, 71, 72, 74, 82, 83, 87, 93, 97, 102)</td>
<td><strong>Sharp</strong> (32, 37)</td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (3, 10, 32, 47, 51, 95, 96, 103)</td>
<td><strong>Precise</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong> (3, 64, 79)</td>
<td><strong>Flat</strong> (8)</td>
</tr>
<tr>
<td><strong>Happiness</strong></td>
<td><strong>High</strong> (1, 3, 4, 6, 10, 14, 16, 19, 20, 24, 32, 33, 37, 43, 44, 45, 46, 47, 50, 51, 52, 53, 55, 64, 71, 74, 75, 80, 82, 86, 87, 88, 97, 99)</td>
<td><strong>Sharp</strong> (8, 26, 32)</td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (70, 101)</td>
<td><strong>Hevner, 1937; Rigg, 1940; Wedin, 1972</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong> (18, 89)</td>
<td><strong>High</strong></td>
</tr>
<tr>
<td><strong>Sadness</strong></td>
<td><strong>High</strong> (16, 53, 70, 97)</td>
<td><strong>Precise</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (18)</td>
<td><strong>Flat</strong> (4, 13, 16, 26, 32, 37)</td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong></td>
<td><strong>Sharp</strong> (Gundlach, 1935; Hevner, 1937; Rigg, 1940; K.B. Watson, 1942; Wedin, 1972)</td>
</tr>
<tr>
<td><strong>Tenderness</strong></td>
<td><strong>High</strong> (33)</td>
<td><strong>Precise</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (4, 24, 32, 96)</td>
<td><strong>Flat</strong></td>
</tr>
</tbody>
</table>

**F0 variability**

<table>
<thead>
<tr>
<th>Emotion</th>
<th>FO (M)</th>
<th>FO (M)²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anger</strong></td>
<td><strong>High</strong> (1, 4, 6, 9, 13, 14, 16, 18, 24, 29, 33, 36, 42, 47, 51, 55, 56, 63, 70, 71, 72, 74, 89, 95, 97, 99, 103)</td>
<td><strong>High</strong> (32)</td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (3, 64, 75, 82)</td>
<td><strong>Medium</strong> (3)</td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong> (10, 44, 50, 83)</td>
<td><strong>Low</strong> (37)</td>
</tr>
<tr>
<td><strong>Fear</strong></td>
<td><strong>High</strong> (4, 13, 16, 18, 29, 56, 80, 89, 103)</td>
<td><strong>High</strong> (32)</td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (47, 72, 74, 82, 95, 97)</td>
<td><strong>Medium</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong> (1, 3, 6, 10, 32, 33, 42, 47, 50, 51, 63, 64, 70, 71, 80, 83, 96)</td>
<td><strong>Low</strong> (12, 37)</td>
</tr>
<tr>
<td><strong>Happiness</strong></td>
<td><strong>High</strong> (3, 4, 6, 9, 10, 13, 16, 18, 20, 32, 33, 37, 42, 44, 47, 48, 50, 51, 55, 56, 64, 71, 72, 74, 75, 80, 82, 83, 86, 89, 95, 97, 99)</td>
<td><strong>High</strong> (3, 12, 37)</td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (14, 70)</td>
<td><strong>Medium</strong> (32)</td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong> (1)</td>
<td><strong>High</strong></td>
</tr>
<tr>
<td><strong>Sadness</strong></td>
<td><strong>High</strong> (10, 20)</td>
<td><strong>Medium</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (6)</td>
<td><strong>Low</strong> (32)</td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong> (3, 4, 9, 13, 14, 16, 18, 29, 32, 37, 44, 47, 50, 51, 55, 63, 64, 70, 71, 72, 74, 75, 80, 82, 86, 89, 93, 97, 99, 103)</td>
<td><strong>Low</strong> (3, 29, 32)</td>
</tr>
<tr>
<td><strong>Tenderness</strong></td>
<td><strong>High</strong></td>
<td><strong>High</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Medium</strong> (4, 24, 33, 95, 96)</td>
<td><strong>Medium</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Low</strong></td>
<td><strong>Low</strong> (12)</td>
</tr>
</tbody>
</table>
Discussion

- Data Collection
- Theoretical Assumptions
- Prosodic Features
- Lexical Features
- Discourse/Dialogue Features