Prosody-Based Detection of Annoyance and Frustration in Communicator Dialogs

Liz Shriberg* Andreas Stolcke* Jeremy Ang+

* SRI International
International Computer Science Institute
+ UC Berkeley
Introduction

- Prosody = rhythm, melody, “tone” of speech
- Largely unused in current ASU systems
- Prior work: prosody aids many tasks:
  - Automatic punctuation
  - Topic segmentation
  - Word recognition
- Today’s talk: detection of user frustration in DARPA Communicator data
  *(ROAR project suggested by Jim Bass)*
Talk Outline

- Data and labeling
- Prosodic and other features
- Classifier models
- Results
- Conclusions and future directions
Key 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 Sources

- Labeled Communicator data from various sites
  - 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

- Each site used different formats and conventions, so tried to minimize the number of sources, maximize the amount of data.

- Thanks to NIST, CMU, Colorado, Lucent, UW
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.
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><strong>TOTAL</strong></td>
<td><strong>21653</strong></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 (%) (chance = 50%)</th>
<th>Kappa (Acc-C)/(1-C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each Human with Other Human, overall</td>
<td>71.7</td>
<td>.38</td>
</tr>
<tr>
<td>Human with Human “Consensus” (biased)</td>
<td>84.2</td>
<td>.68</td>
</tr>
<tr>
<td>Prosodic Decision Tree with Consensus</td>
<td>75.6</td>
<td>.51</td>
</tr>
<tr>
<td>Tree with Consensus, no repeat/correction</td>
<td>72.9</td>
<td>.46</td>
</tr>
<tr>
<td>Tree with Consensus, repeat/correction only</td>
<td>68.7</td>
<td>.37</td>
</tr>
<tr>
<td>Language Model features only</td>
<td>63.8</td>
<td>.28</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)
Baseline Prosodic Tree

duration feature  pitch feature  other feature

REPCO in ec2,rr1,rr2,rex2,inc,ec1,rex1 : 0.7699 0.2301 AF
  MAXF0_IN_MAXV_N < 126.93: 0.4735 0.5265 N
  MAXF0_IN_MAXV_N >= 126.93: 0.8296 0.1704 AF
    MAXPHDUR_N < 1.6935: 0.6466 0.3534 AF
      UTTPOS < 5.5: 0.1724 0.8276 N
      UTTPOS >= 5.5: 0.7008 0.2992 AF
    MAXPHDUR_N >= 1.6935: 0.8852 0.1148 AF
REPCO in 0 : 0.3966 0.6034 N
  UTTPOS < 7.5: 0.1704 0.8296 N
  UTTPOS >= 7.5: 0.4658 0.5342 N
  VOWELDUR_DNORM_E_5 < 1.2396: 0.3771 0.6229 N
    MINF0TIME < 0.875: 0.2372 0.7628 N
    MINF0TIME >= 0.875: 0.5 0.5 AF
      SYLRATE < 4.7215: 0.562 0.438 AF
        MAXF0_TOPLN < -0.2177: 0.3942 0.6058 N
        MAXF0_TOPLN >= -0.2177: 0.6637 0.3363 AF
      SYLRATE >= 4.7215: 0.2816 0.7184 N
    VOWELDUR_DNORM_E_5 >= 1.2396: 0.5983 0.4017 AF
      MAXPHDUR_N < 1.5395: 0.3841 0.6159 N
        MINF0TIME < 0.435: 0.1 0.9 N
        MINF0TIME >= 0.435: 0.4545 0.5455 N
        RISERATIO_DNORM_E_5 < 0.69872: 0.3284 0.6716 N
        RISERATIO_DNORM_E_5 >= 0.69872: 0.6111 0.3889 AF
      MAXPHDUR_N >= 1.5395: 0.6728 0.3272 AF
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 (%) (chance = 50%)</th>
<th>Entropy Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline prosody model</td>
<td>75.6</td>
<td>21.6</td>
</tr>
<tr>
<td>Consensus labels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A,F vs. N,else</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tokens on which labelers</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>All tokens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus labels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>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 our baseline.
Error tradeoffs (ROC)
Conclusion

- 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, can be online
Conclusions (cont.)

- Two nonprosodic features are important: utterance position and repeat/correction

- Even if repeat/correction not used, prosody still good predictor (better than human-human)

- Language model is an imperfect surrogate feature for the underlying important feature repeat/correction

- Look for other useful dialog features!
Future Directions

- Use realistic data to get more real frustration

- Improve features:
  - use new F0 fitting, capture voice quality
  - base on ASR output (1-best straightforward)
  - optimize online normalizations

- Extend modeling:
  - model frustration sequences, include dialog state
  - exploit speaker ‘habits’

- Produce prosodically ‘tagged’ data, using combinations of current feature primitives

- Extend task to other useful emotions & domains.
Thank You