Detecting Attempts at Humor in Multiparty Meetings

Kornel Laskowski

Carnegie Mellon University
Pittsburgh PA, USA

14 September, 2008
Why bother with humor?

- generally, systems assume uniform truth across utterances
- humans do not make that assumption
  - a speaker may be unconcerned how their utterance is interpreted
  - but a speaker may covertly perform extra work to pass off as true/serious that which is not
    → speaker is not helping us detect their effort (e.g. lying)
  - or a speaker may overtly perform extra work to pass off as untrue/unserious that which may be taken at face value
    → speaker is helping us detect their effort (e.g. joking)
- need to detect grades of truth, at least when speakers are collaborative
Why bother with humor (part II)?

- humor plays a socially cohesive role
- creates vehicle for expressing, maintaining, constructing, dissolving interpersonal relationships
- systems must detect it, or miss important cues underlying variability across participants to conversation
Why bother with humor (part III)?

- humor does not occur uniformly in time
- its occurrence is colocated with segment boundaries at the detection may be helpful to segmentation of conversation at the
  - turn level
  - topic level
  - meta-conversation level
- systems must detect it, or miss important cues underlying variability across time in conversation
Outline of this Talk

1. Introduction
2. Humor in our Data
3. HMM Decoder Framework
   - baseline (oracle) lexical features
4. Modeling Conversational Context
   - speech activity/interaction features
   - laughter activity/interaction features
5. Analysis
6. Conclusions & Recommendations
must determine if current speaker is intending to amuse
  task may be too hard for a computer
  instead, let **humans** do the work

**offline**: wait to see if **others** laugh
  even if attempt to amuse fails, others may laugh to show that they understand the utterance is not meant seriously

**online**: wait to see if **speaker** laughs
  to show that utterance is not meant seriously

SPKR A  \[ \text{JOKE} \]

SPKR B  \[ \text{........................................} \]

SPKR C  \[ \text{........................................} \]
must determine if current speaker is intending to amuse
  task may be too hard for a computer
  instead, let **humans** do the work

**offline**: wait to see if **others** laugh
  even if attempt to amuse fails, others may laugh to show that
  they understand the utterance is not meant seriously

**online**: wait to see if **speaker** laughs
  to show that utterance is not meant seriously

---

Potential Impact of Modeling Laughter

![Diagram showing laughter and joke relationships between speakers A, B, and C.]

SPKR A → JOKE

SPKR B → LAUGH

SPKR C → LAUGH
Potential Impact of Modeling Laughter

- must determine if current speaker is intending to amuse
  - task may be too hard for a computer
  - instead, let **humans** do the work
- **offline**: wait to see if **others** laugh
  - even if attempt to amuse fails, others may laugh to show that they understand the utterance is not meant seriously
- **online**: wait to see if **speaker** laughs
  - to show that utterance is not meant seriously

![Diagram showing interactions between speakers and laughter reactions](attachment:image.png)
Computational Context and Prior Work

**SENTIMENT DETECTION**
Somasundaran et al, 2007

**HUMOR DETECTION**
Clark & Popescu-Belis, 2004

**EMOTIONAL VALENCE DETECTION**
Laskowski & Burger, 2006
Neiberg et al, 2006

**EMOT. INVOLVED SPEECH DETECTION**
Wrede & Shriberg, 2003

**SPEECH RECOGNITION**

**SPEECH ACTIVITY DETECTION**

**PROSODIC MODELING**

**LAUGHTER ACTIVITY DETECTION**
Kennedy & Ellis, 2004
Truong & van Leeuwen, 2005
Knox & Mirghafori, 2007

**AUDIO**
Computational Context and Prior Work

- **SENTIMENT DETECTION**
  Somasundaran et al., 2007

- **HUMOR DETECTION**
  Clark & Popescu-Belis, 2004

- **EMOTIONAL VALENCE DETECTION**
  Laskowski & Burger, 2006
  Neiberg et al., 2006

- **EMOT. INVOLVED SPEECH DETECTION**
  Wrede & Shriberg, 2003

- **SPEECH ACTIVITY DETECTION**

- **PROSODIC MODELING**

- **SPEECH RECOGNITION**

- **LAUGHTER ACTIVITY DETECTION**
  Kennedy & Ellis, 2004
  Truong & van Leeuwen, 2005
  Knox & Mirghafori, 2007

**AUDIO**
Computational Context and Prior Work

- **SENTIMENT DETECTION**
  - Somasundaran et al., 2007

- **HUMOR DETECTION**
  - Clark & Popescu-Belis, 2004

- **EMOTIONAL VALENCE DETECTION**
  - Laskowski & Burger, 2006
  - Neiberg et al., 2006

- **EMOT. INVOLVED SPEECH DETECTION**
  - Wrede & Shriberg, 2003

- **SPEECH ACTIVITY DETECTION**

- **SPEECH RECOGNITION**

- **PROSODIC MODELING**

- **LAUGHTER ACTIVITY DETECTION**
  - Kennedy & Ellis, 2004
  - Truong & van Leeuwen, 2005
  - Knox & Mirghafori, 2007

- **AUDIO**
Computational Context and Prior Work

- **SENTIMENT DETECTION**
  Somasundaran et al., 2007

- **HUMOR DETECTION**
  Clark & Popescu-Belis, 2004

- **EMOTIONAL VALENCE DETECTION**
  Laskowski & Burger, 2006
  Neiberg et al., 2006

- **EMOT. INVOLVED SPEECH DETECTION**
  Wrede & Shriberg, 2003

- **SPEECH RECOGNITION**

- **SPEECH ACTIVITY DETECTION**

- **PROSODIC MODELING**

- **LAUGHTER ACTIVITY DETECTION**
  Kennedy & Ellis, 2004
  Truong & van Leeuwen, 2005
  Knox & Mirghafori, 2007

- **AUDIO**
Computational Context and Prior Work

- **LAUGHTER ACTIVITY DETECTION**
  - Knox & Mirghafori, 2007
- **EMOTIONAL VALENCE DETECTION**
  - Neiberg et al., 2006
- **HUMOR DETECTION**
  - Clark & Popescu-Belis, 2004
- **EMOT. INVOLVED SPEECH DETECTION**
  - Wrede & Shriberg, 2003
- **SPEECH ACTIVITY DETECTION**
- **PROSODIC MODELING**
- **SENTIMENT DETECTION**
  - Somasundaran et al., 2007
- **SPEECH RECOGNITION**

**AUDIO**

K. Laskowski  ICSC 2009, Berkeley CA, USA  7/26
Computational Context and Prior Work

- **SENTIMENT DETECTION**
  - Somasundaran et al., 2007

- **HUMOR DETECTION**
  - Clark & Popescu-Belis, 2004

- **EMOTIONAL VALENCE DETECTION**
  - Laskowski & Burger, 2006
  - Neiberg et al., 2006

- **EMOT. INVOLVED SPEECH DETECTION**
  - Wrede & Shriberg, 2003

- **SPEECH ACTIVITY DETECTION**
  - Clark & Popescu-Belis, 2004

- **PROSODIC MODELING**

- **LAUGHTER ACTIVITY DETECTION**
  - Kennedy & Ellis, 2004
  - Truong & van Leeuwen, 2005
  - Knox & Mirghafori, 2007

---

K. Laskowski  
ICSC 2009, Berkeley CA, USA
Computational Context and Prior Work

- **Sentiment Detection**: Somasundaran et al., 2007
- **Humor Detection**: Clark & Popescu-Belis, 2004
- **Emotional Valence Detection**: Laskowski & Burger, 2006, Neiberg et al., 2006
- **Emotions Involved Speech Detection**: Wrede & Shriberg, 2003
- **Speech Recognition**
- **Speech Activity Detection**
- **Prosodic Modeling**

**Audio**
Computational Context and Prior Work

- **Sentiment Detection**: Somasundaran et al., 2007
- **Humor Detection**: Clark & Popescu-Belis, 2004
- **Emotional Valence Detection**: Neiberg et al., 2006
- **Emotional Involved Speech Detection**: Laskowski, 2008
- **Speech Recognition**
- **Speech Activity Detection**
- **Prosodic Modeling**

Audio signals are the starting point for all these tasks.
Computational Context and Prior Work

- **Sentiment Detection**
  - Somasundaran et al, 2007

- **Humor Detection**
  - Clark & Popescu-Belis, 2004

- **Emotional Valence Detection**
  - Laskowski & Burger, 2006
  - Neiberg et al, 2006

- **Emot. Involved Speech Detection**
  - Wrede & Shriberg, 2003

- **Speech Recognition**

- **Speech Activity Detection**

- **Prosodic Modeling**

- **Laughter Activity Detection**
  - Kennedy & Ellis, 2004
  - Truong & van Leeuwen, 2005
  - Knox & Mirghafori, 2007

### Audio

K. Laskowski  
ICSC 2009, Berkeley CA, USA  
7/26
Computational Context and Prior Work

- **SENTIMENT DETECTION**
  - Somasundaran et al., 2007

- **HUMOR DETECTION**
  - Clark & Popescu-Belis, 2004

- **EMOTIONAL VALENCE DETECTION**
  - Laskowski & Burger, 2006
  - Neiberg et al., 2006

- **EMOT. INVOLVED SPEECH DETECTION**
  - Wrede & Shriberg, 2003

- **SPEECH RECOGNITION**

- **SPEECH ACTIVITY DETECTION**

- **PROSODIC MODELING**

- **LAUGHTER ACTIVITY DETECTION**
  - Kennedy & Ellis, 2004
  - Truong & van Leeuwen, 2005
  - Knox & Mirghafori, 2007

- **AUDIO**
Computational Context and Prior Work

- **Sentiment Detection**: Somasundaran et al., 2007
- **Humor Detection**: Clark & Popescu-Belis, 2004
- **Emotional Valence Detection**: Laskowski & Burger, 2006; Neiberg et al., 2006
- **Emot. Involved Speech Detection**: Wrede & Shriberg, 2003; Laskowski, 2008
- **Speech Recognition**
- **Speech Activity Detection**
- **Prosodic Modeling**
- **Laughter Activity Detection**: Kennedy & Ellis, 2004; Truong & van Leeuwen, 2005; Knox & Mirghafori, 2007
Computational Context and Prior Work

- **Sentiment Detection**: Somasundaran et al., 2007
- **Humor Detection**: Clark & Popescu-Belis, 2004
- **Emotional Valence Detection**: Laskowski & Burger, 2006
  - Neiberg et al., 2006
- **Emotionally Involved Speech Detection**: Wrede & Shriberg, 2003
  - Laskowski, 2008
- **Speech Recognition**
- **Speech Activity Detection**
- **Prosodic Modeling**
- **Laughter Activity Detection**: Kennedy & Ellis, 2004
  - Truong & van Leeuwen, 2005
  - Knox & Mirghafori, 2007

**Audio**

K. Laskowski  ICSC 2009, Berkeley CA, USA  7/26
ICSI Meeting Corpus (Janin et al, 2003; Shriberg et al, 2004)

- naturally occurring meetings
- 75 meetings, 66 hours of meeting time
  - TrainSet: 51 meetings
  - DevSet: 11 meetings
  - EvalSet: 11 meetings
- 3-9 participants per meeting
- different types
  - unstructured discussion among peers
  - round-table reporting among peers
  - “1 professor and N students” meetings
- human-transcribed words (with forced-alignment), dialog acts
Humor Annotation in ICSI Meetings

Based on the 8 DA types studied in


### Propositional Content DA Types

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbol</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>statement</td>
<td>s</td>
<td>85%</td>
</tr>
<tr>
<td>question</td>
<td>q</td>
<td>6.6%</td>
</tr>
</tbody>
</table>

### Feedback DA Types

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbol</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>backchannel</td>
<td>b</td>
<td>2.8%</td>
</tr>
<tr>
<td>acknowledgment</td>
<td>bk</td>
<td>1.4%</td>
</tr>
<tr>
<td>assert</td>
<td>aa</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

### Floor Mechanism DA Types

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbol</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>floor holder</td>
<td>fh</td>
<td>2.5%</td>
</tr>
<tr>
<td>floor grabber</td>
<td>fg</td>
<td>0.6%</td>
</tr>
<tr>
<td>hold</td>
<td>h</td>
<td>0.3%</td>
</tr>
</tbody>
</table>
Humor Annotation in ICSI Meetings

Based on the 8 DA types studied in


**Propositional Content DA Types**

- statement: s, 85%
- question: q, 6.6%

**Humor-Bearing DA Types**

- joke: j, 0.6%

**Feedback DA Types**

- backchannel: b, 2.8%
- acknowledgment: bk, 1.4%
- assert: aa, 1.1%

**Floor Mechanism DA Types**

- floor holder: fh, 2.5%
- floor grabber: fg, 0.6%
- hold: h, 0.3%
Goal of this Work

SPKR A: ...........................................................
SPKR B: ...........................................................
SPKR C: ...........................................................
SPKR D: ...........................................................
Goal of this Work

SPKR A: .......................................................... .......................................................... .......................................................... TALKSPURT
SPKR B: .......................................................... .......................................................... .......................................................... ..........................................................
SPKR C: .......................................................... ..........................................................
SPKR D: .......................................................... ..........................................................
Goal of this Work

SPKR A: .................................................................
SPKR B: .........................................................
SPKR C: .................................................................
SPKR D: .................................................................

LAUGHABOUT
Task: find speech which is **humor-bearing**
Goal of this Work

TASK: find speech which is **humor-bearing**
(DA segmentation and recognition, with focus on a subset of DAs)
Talkspurt (TS) Boundaries ≠ DA Boundaries

- Decoding the state of one participant at a time
- May have 1:1 correspondence between DAs and TSs
- And 1:1 correspondence between DA-gaps and TS-gaps
- But may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  → Explicitly model intra-DA silence
- Opposite (N:1 correspondence) may also occur
  → Entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries $\neq$ DA Boundaries

- decoding the state of one participant at a time
  - may have 1:1 correspondence between DAs and TSs
  - and 1:1 correspondence between DA-gaps and TS-gaps
  - but may also have TS gaps inside DAs
  - 1:N correspondence between DAs and TSs
    $\rightarrow$ explicitly model intra-DA silence
  - opposite (N:1 correspondence) may also occur
    $\rightarrow$ entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries ≠ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  → explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  → entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries ≠ DA Boundaries

- decoding the state of one participant at a time
  - may have 1:1 correspondence between DAs and TSs
  - and 1:1 correspondence between DA-gaps and TS-gaps
  - but may also have TS gaps inside DAs
  - 1:N correspondence between DAs and TSs
    → explicitly model intra-DA silence
  - opposite (N:1 correspondence) may also occur
    → entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries ≠ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  - → explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  - → entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries $\neq$ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  - explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  - entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries $\neq$ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\rightarrow$ explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  $\rightarrow$ entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries $\neq$ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  $\rightarrow$ explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  $\rightarrow$ entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries ≠ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  \[\rightarrow\] explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  \[\rightarrow\] entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries ≠ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  → explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  → entertain possibility that DA boundaries occur anywhere
Talkspurt (TS) Boundaries ≠ DA Boundaries

- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps inside DAs
- 1:N correspondence between DAs and TSs
  \[\rightarrow\] explicitly model intra-DA silence
- opposite (N:1 correspondence) may also occur
  \[\rightarrow\] entertain possibility that DA boundaries occur anywhere
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs

SPKR B: [Audio waveform]

[Diagram of HMM sub-topology with nodes and transitions labeled as "ENTRY", "EGRESS", "NON-DA-TERMINAL TALKSPURT FRAGMENT", "INTRA-DA TALKSPURT GAP", "DA-TERMINAL TALKSPURT FRAGMENT"]
Proposed HMM Sub-Topology for DAs

SPKR B:

ENTRY

NON–DA–TERMINAL TALKSPURT FRAGMENT

INTRA–DA TALKSPURT GAP

DA–TERMINAL TALKSPURT FRAGMENT

EGRESS
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs

ENTRY EGRESS

NON-DA-TERMINAL TALKSPURT FRAGMENT

INTRA-DA TALKSPURT GAP

DA-TERMINAL TALKSPURT FRAGMENT

ENTRY

EGRESS

SPKR B:
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs

Entrance

K. Laskowski
ICSC 2009, Berkeley CA, USA
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs

NON-DA-TERMINAL TALKSPURT FRAGMENT

INTRA-DA TALKSPURT GAP

DA-TERMINAL TALKSPURT FRAGMENT

ENTRY

EGRESS

SPKR B:
Proposed HMM Sub-Topology for DAs

Non-DA-Terminal Talkspurt Fragment

Intra-DA Talkspurt Gap

DA-Terminal Talkspurt Fragment

ENTRY

EGRESS

SPKR B:
Proposed HMM Sub-Topology for DAs

EGRESS

NON–DA–TERMINAL TALKSPURT FRAGMENT

INTRA–DA TALKSPURT GAP

DA–TERMINAL TALKSPURT FRAGMENT

ENTRY

EGRESS

SPKR B:

K. Laskowski

ICSC 2009, Berkeley CA, USA
Proposed HMM Sub-Topology for DAs

EGRESS

NON-DA-TERMINAL TALKSPURT FRAGMENT

INTRA-DA TALKSPURT GAP

DA-TERMINAL TALKSPURT FRAGMENT

ENTRY

SPKR B:
Proposed HMM Sub-Topology for DAs

- **NON-DA-TERMINAL TALKSPURT FRAGMENT**
- **INTRA-DA TALKSPURT GAP**
- **DA-TERMINAL TALKSPURT FRAGMENT**

**SPKR B:**
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs
Proposed HMM Sub-Topology for DAs
the complete topology consists of
- a DA sub-topology for each of 9 DA types
- fully connected via inter-DA GAP subnetworks
Oracle Lexical Features

- each 100 ms frame of speech can be assigned to one word \( w \)
- assign to that frame the emission probability:
  - of the bigram of which \( w \) is the right token, and
  - of the bigram of which \( w \) is the left token
- train a generative model over left and right bigrams for each HMM state
- bigrams whose probability of occurrence for any DA type is < 0.1% are mapped to UNK
Baseline Performance

"w/o T" fully-connected topology, equiprobable transitions

"w/ T0" proposed topology, equiprobable transitions

"w/ T1" proposed topology, transitions trained using TRAINSET (ML)

<table>
<thead>
<tr>
<th>System</th>
<th>DevSet</th>
<th></th>
<th>EvalSet</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>MS</td>
<td>ERR</td>
<td>FA</td>
</tr>
<tr>
<td>T0</td>
<td>8.1</td>
<td>90.6</td>
<td>98.7</td>
<td>8.3</td>
</tr>
<tr>
<td>T1</td>
<td>0.3</td>
<td>96.7</td>
<td>97.0</td>
<td>0.2</td>
</tr>
<tr>
<td>LEX w/o T</td>
<td>53.6</td>
<td>32.8</td>
<td>86.4</td>
<td>53.7</td>
</tr>
<tr>
<td>LEX w/ T0</td>
<td>40.2</td>
<td>42.9</td>
<td>83.1</td>
<td>40.5</td>
</tr>
<tr>
<td>LEX w/ T1</td>
<td>12.7</td>
<td>67.0</td>
<td>79.6</td>
<td>12.8</td>
</tr>
</tbody>
</table>
Speech Activity/Interaction Features, $S$

- decoding one participant (SPKR) at a time
  - at instant $t$, model the *thumbnail image* of context
    - consider a temporal context of width $T$
  - want invariance under participant-index rotation

- want a fixed-size feature vector: consider only $K$ others
- model features using state-specific GMMs (after LDA)
Speech Activity/Interaction Features, $S$

- decoding one participant (SPKR) at a time
- at instant $t$, model the *thumbnail image* of context
  - consider a temporal context of width $T$
- want invariance under participant-index rotation
  - rank "OTH" participants by local speaking time
- want a fixed-size feature vector: consider only $K$ others
- model features using state-specific GMMs (after LDA)
Speech Activity/Interaction Features, $S$

- decoding one participant (SPKR) at a time
- at instant $t$, model the *thumbnail image* of context
  - consider a temporal context of width $T$
- want invariance under participant-index rotation
- rank "OTH" participants by local speaking time
- want a fixed-size feature vector: consider only $K$ others
- model features using state-specific GMMs (after LDA)
Speech Activity/Interaction Features, $S$

- decoding one participant (SPKR) at a time
- at instant $t$, model the *thumbnail image* of context
  - consider a temporal context of width $T$
- want invariance under participant-index rotation
  - rank “OTH” participants by local speaking time
- want a fixed-size feature vector: consider only $K$ others
- model features using state-specific GMMs (after LDA)
Speech Activity/Interaction Features, $S$

- decoding one participant (SPKR) at a time
- at instant $t$, model the *thumbnail image* of context
  - consider a temporal context of width $T$
- want invariance under participant-index rotation
  - rank "OTH" participants by local speaking time
- want a fixed-size feature vector: consider only $K$ others
- model features using state-specific GMMs (after LDA)
Speech Activity/Interaction Features, $S$

- decoding one participant (SPKR) at a time
- at instant $t$, model the *thumbnail image* of context
  - consider a temporal context of width $T$
- want invariance under participant-index rotation
  - rank “OTH” participants by local speaking time
- want a fixed-size feature vector: consider only $K$ others
- model features using state-specific GMMs (after LDA)
decoding one participant (SPKR) at a time
at instant t, model the *thumbnail image* of context
consider a temporal context of width $T$
want invariance under participant-index rotation
rank “OTH” participants by **local** speaking time
want a fixed-size feature vector: consider only $K$ others
model features using state-specific GMMs (after LDA)
decoding one participant (SPKR) at a time
at instant t, model the *thumbnail image* of context
consider a temporal context of width $T$
want invariance under participant-index rotation
rank “OTH” participants by *local* speaking time
want a fixed-size feature vector: consider only $K$ others
model features using state-specific GMMs (after LDA)
Speech Activity/Interaction Features, $S$

- decoding one participant (SPKR) at a time
- at instant $t$, model the *thumbnail image* of context
  - consider a temporal context of width $T$
- want invariance under participant-index rotation
  - rank “OTH” participants by *local* speaking time
- want a fixed-size feature vector: consider only $K$ others
- model features using state-specific GMMs (after LDA)
process same as for speech activity/interaction features:

1. sort others by amount of **laughing time** in $T$-width window
2. extract features from $K$ most-laughing others

may be suboptimal (too complex $\rightarrow$ overfit)

laughter accounts for 9.6% of vocalizing time

in the paper, also consider subsetting all laughter bouts into:

- voiced bouts (approx. $2/3$ of laughter by time)
- unvoiced bouts (approx. $1/3$ of laughter by time)
System Combination

1. model-space combination (\(\circ\))

\[
P ([F_S, F_L] | [M_S, M_L]) \equiv P (F_S | M_S) P (F_L | M_L)
\]

\[
F_S = f (K, \text{rank}(S), S)
\]

\[
F_L = f (K, \text{rank}(L), L)
\]

2. feature-space combination (\(\mathbb{F}\))

\[
P ([F_S, F_L] | [M_S, M_L]) \equiv P ([F_S, F_L] | M_{S \cup L})
\]

\[
F_S = f (K, \text{rank}(S), S)
\]

\[
F_L = f (K, \text{rank}(L), L)
\]

3. feature-computation-space combination (\(\mathbb{C}\))

\[
P ([F_S, F_L] | [M_S, M_L]) \equiv P ([F_S, F_L] | M_{S \cup L})
\]

\[
F_S = f (K, \text{rank}(S \cup L), S)
\]

\[
F_L = f (K, \text{rank}(S \cup L), L)
\]
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>DevSet</th>
<th></th>
<th>EvalSet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>MS</td>
<td>ERR</td>
</tr>
<tr>
<td>LEX</td>
<td>12.7</td>
<td>67.0</td>
<td>79.6</td>
</tr>
<tr>
<td>S</td>
<td>7.5</td>
<td>47.4</td>
<td>54.9</td>
</tr>
<tr>
<td>L</td>
<td>14.0</td>
<td>5.3</td>
<td>19.3</td>
</tr>
<tr>
<td>S ⊙ M L</td>
<td>9.7</td>
<td>6.6</td>
<td>16.3</td>
</tr>
<tr>
<td>S ⊙ F L</td>
<td>6.0</td>
<td>17.8</td>
<td>23.8</td>
</tr>
<tr>
<td>S ⊙ C L</td>
<td>6.0</td>
<td>16.0</td>
<td>22.0</td>
</tr>
<tr>
<td>LEX ⊙ S ⊙ L</td>
<td>7.7</td>
<td>7.2</td>
<td>14.8</td>
</tr>
</tbody>
</table>

- L is the best single source of information for this task
- **model-space** combination with S leads to improvement
- combination with LEX leads to improvement on **DevSet** only
Receiver Operating Characteristics (DevSet)
Interpreting Emission Probability Diagrams

- condition: given an event of type A occurring at time $t$
- what is the likelihood that an event of type B occurs at time $t' \in [t - 5, t + 5]$
- retrain single-Gaussian model on unnormalized features
Interlocutor Laughter Context at DA Termination

\( j \) DAs

\( \neg j \) DAs

Locally 1st most laughing

Locally 2nd most laughing

K. Laskowski

ICSC 2009, Berkeley CA, USA
Interlocutor Laughter Context at DA Termination

- $j$ DAs

- $\neg j$ DAs

Locally 1st most laughing

Locally 2nd most laughing

K. Laskowski

ICSC 2009, Berkeley CA, USA
Interlocutor Laughter Context at DA Termination

\[ j \text{ DAs} \]

\[ \neg j \text{ DAs} \]
Interlocutor Laughter Context at DA Termination

- $j$ DAs
  - locally 1st most laughing
  - locally 2nd most laughing

- $\neg j$ DAs
  - located 1st most laughing
  - located 2nd most laughing

K. Laskowski
ICSC 2009, Berkeley CA, USA
Interlocutor Laughter Context at DA Termination

j DAs

Locally 1st most laughing

−j DAs

Locally 2nd most laughing
Interlocutor Laughter Context at DA Termination

\( j \) DAs

\( \neg j \) DAs

Locally 1st most laughing

Locally 2nd most laughing
Target Speaker Laughter Context

How well we do with laughter only from the target speaker?
Target Speaker Laughter Context

How well we do with laughter only from the target speaker?
Target Speaker Laughter Context

How well we do with laughter only from the target speaker?
Target Speaker Laughter Context

How well we do with laughter only from the target speaker?
Target Speaker Laughter Context

- How well we do with laughter only from the target speaker?
Target Speaker Laughter Context

How well we do with laughter only from the target speaker?

<table>
<thead>
<tr>
<th>System</th>
<th>DevSet</th>
<th>EVALSet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>MS</td>
</tr>
<tr>
<td>$S$</td>
<td>7.5</td>
<td>47.4</td>
</tr>
<tr>
<td>$L$</td>
<td>14.0</td>
<td>5.3</td>
</tr>
<tr>
<td>$L'$</td>
<td>8.7</td>
<td>20.3</td>
</tr>
</tbody>
</table>
Interlocutor j-Speech Context at j-DA Termination

- Target speaker
- Locally 1st most j-talkative interlocutor
- Locally 2nd most j-talkative interlocutor
Interlocutor j-Speech Context at j-DA Termination

**target speaker**

**locally 1st most j-talkative interlocutor**

**locally 2nd most j-talkative interlocutor**
Interlocutor j-Speech Context at j-DA Termination

**target speaker**

**locally 1st most j-talkative interlocutor**

**locally 2nd most j-talkative interlocutor**
GOAL:
- detect humor-bearing speech

APPROACH:
- frame-level HMM decoding
- consider multiparticipant speech & laughter context

RESULTS:
1. at FPRs of $\approx 5\%$ (DevSet):
   - lexical features yield TPRs $4 \times$ higher than random guessing
   - speech context yields TPRs $2 \times$ higher than lexical features
   - laughter context yields TPRs $2 \times$ higher than speech context
2. laughter context features: EER $< 24\%$ (EvalSet)
3. model-space combination improves EERs by $\approx 5\%$ abs
4. locally most laughing interlocutor more likely to laugh than not
5. evidence that jokers themselves laugh, perhaps to signal intent
6. at most 2 participants likely to joke in any 10 second interval
THANK YOU

Special thanks to Liz Shriberg, for:
- access to the ICSI MRDA annotations
- helpful discussion during this work