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
    $\longrightarrow$ 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
    $\longrightarrow$ 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 important cues underlying variability across participants to conversation

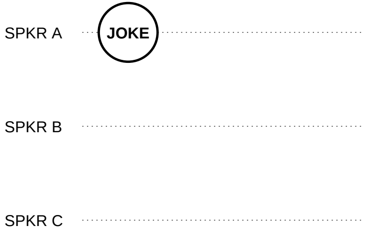
Why bother with humor (part III)?

- 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

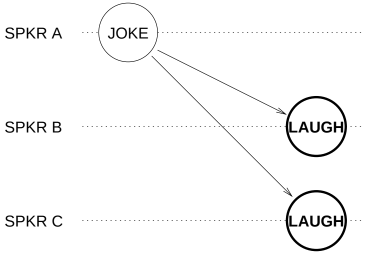
Potential Impact of Modeling Laughter

- 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



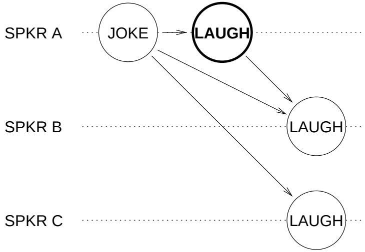
Potential Impact of Modeling Laughter

- 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

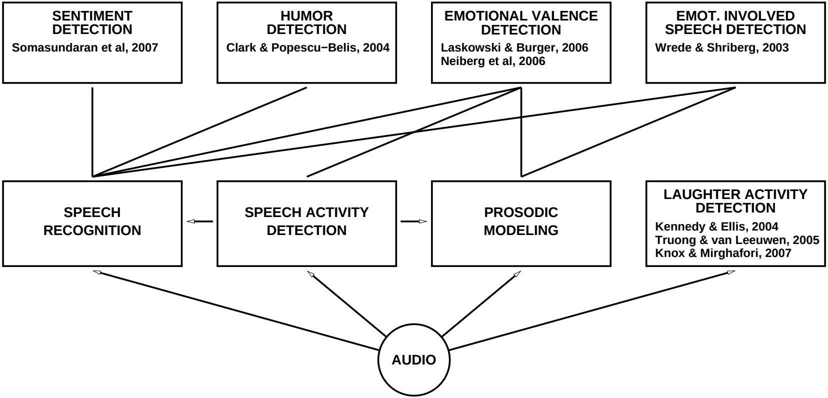


Potential Impact of Modeling Laughter

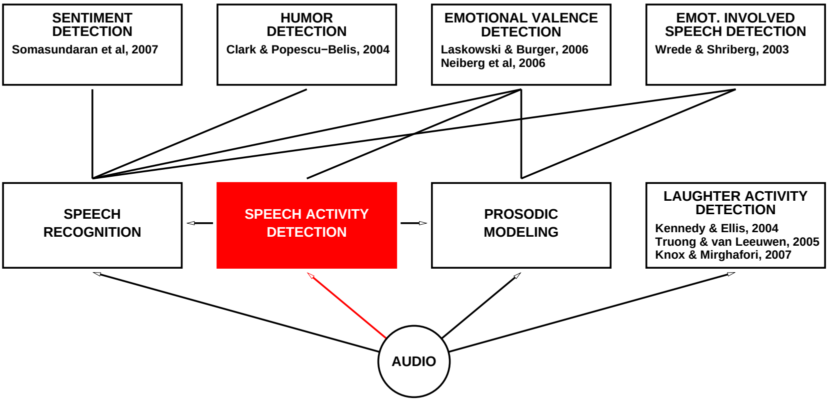
- must determine if current speaker is intending to amuse
  - task may be too hard for a computer
  - instead, let **humans** do the work
- **offline**: wait to see if **others** laugh
  - even if attempt to amuse fails, others may laugh to show that they understand the utterance is not meant seriously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



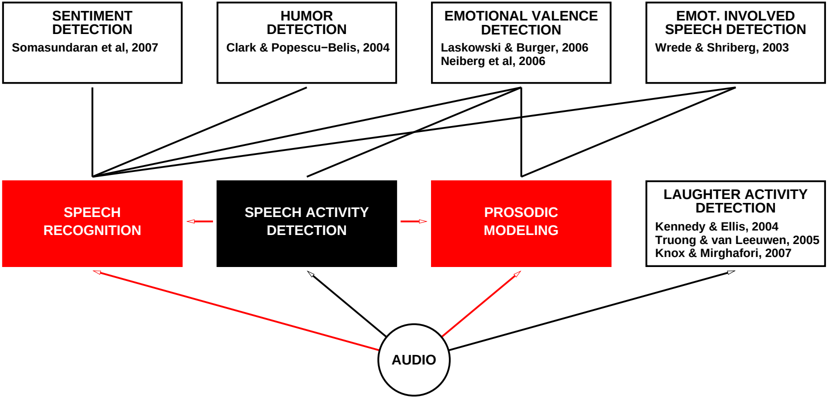
Computational Context and Prior Work



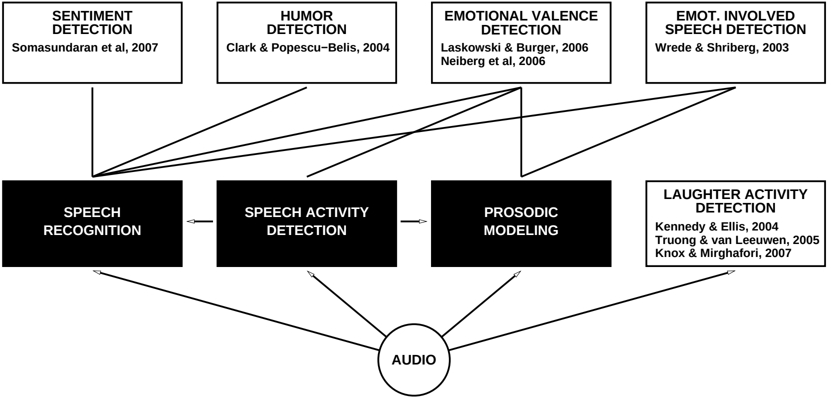
Computational Context and Prior Work



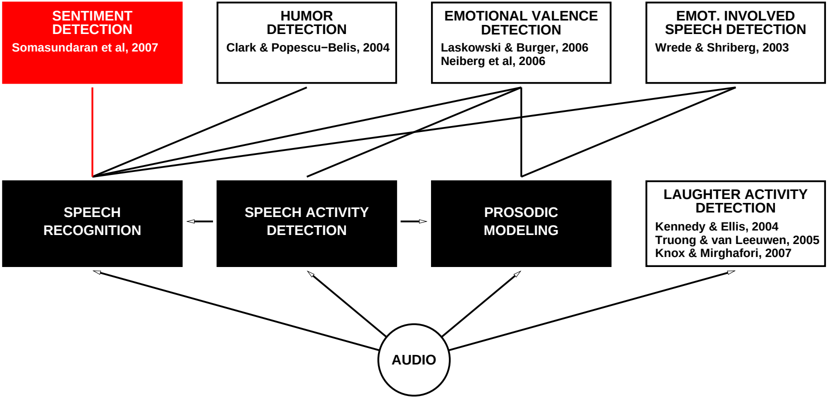
Computational Context and Prior Work



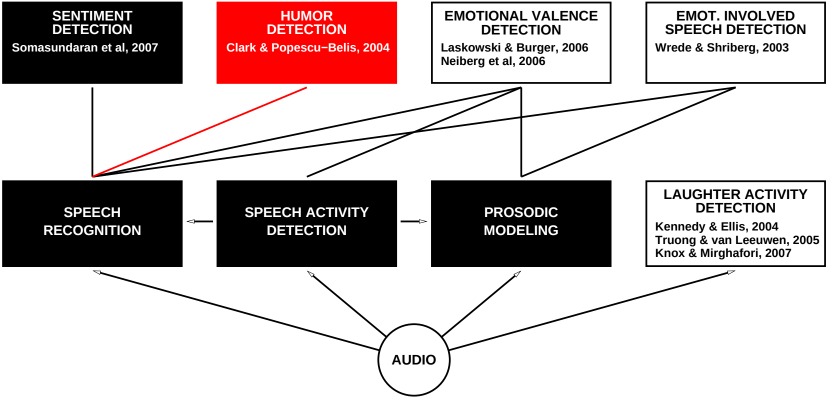
Computational Context and Prior Work



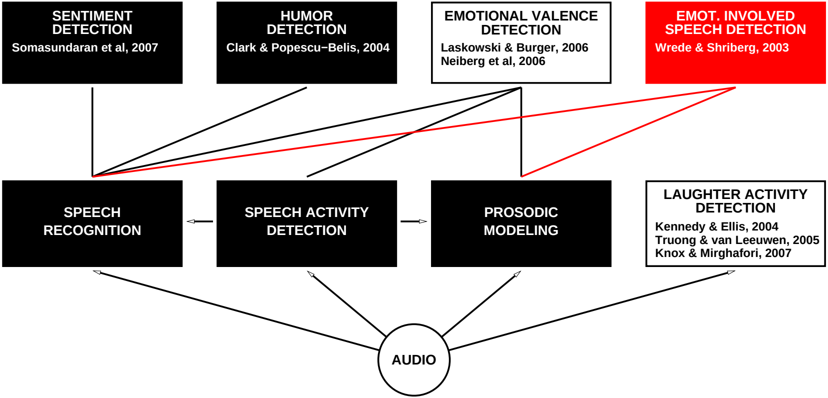
Computational Context and Prior Work



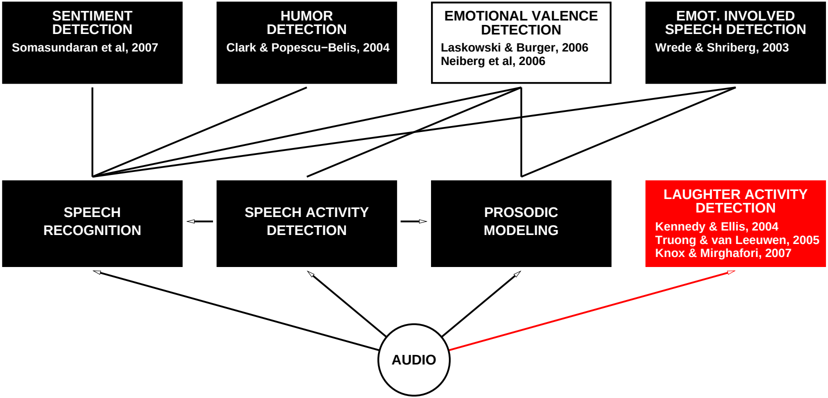
Computational Context and Prior Work



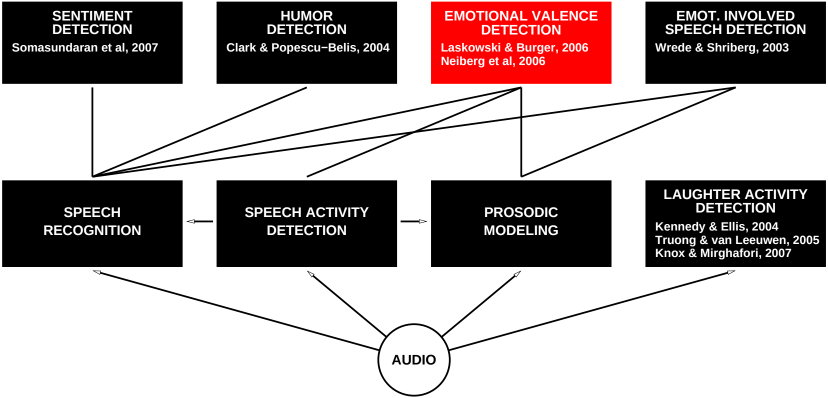
Computational Context and Prior Work



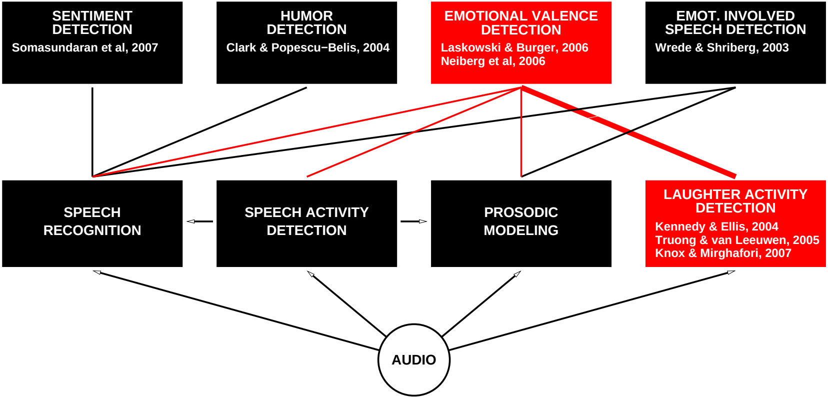
Computational Context and Prior Work



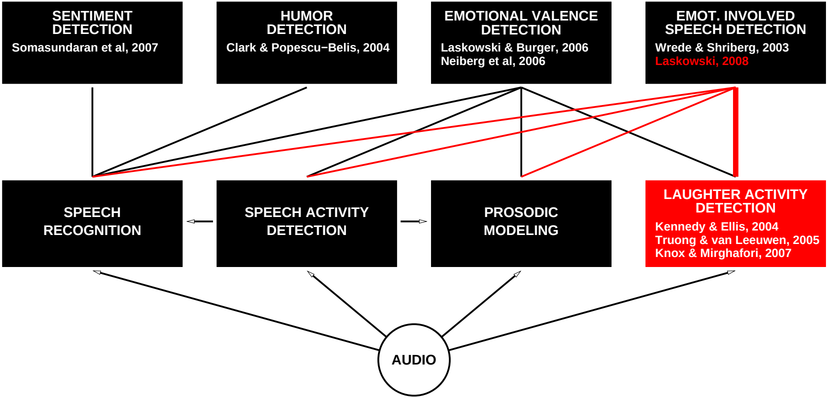
Computational Context and Prior Work



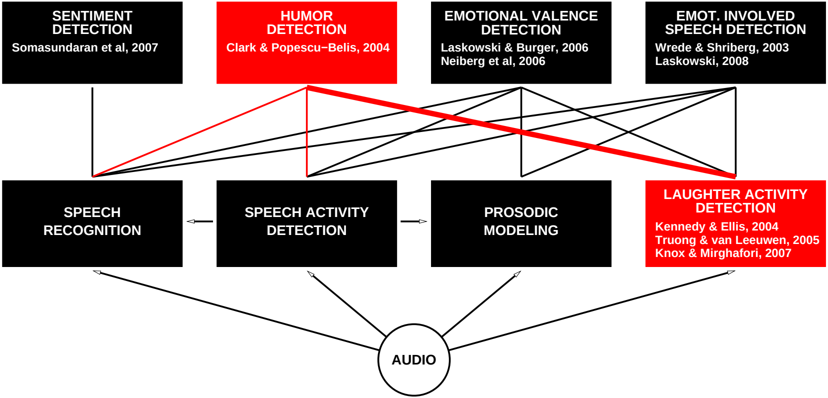
Computational Context and Prior Work



Computational Context and Prior Work



Computational Context and Prior Work



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

- Laskowski & Shriberg, “Modeling Other Talkers for Improved Dialog Act Recognition in Meetings”, INTERSPEECH 2009.

Propositional Content DA Types		
statement	s	85%
question	q	6.6%

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%

Humor Annotation in ICSI Meetings

Based on the 8 DA types studied in

- Laskowski & Shriberg, "Modeling Other Talkers for Improved Dialog Act Recognition in Meetings", INTERSPEECH 2009.

Propositional Content DA Types

statement	s	85%
question	q	6.6%

Humor-Bearing DA Types

joke	j	0.6%

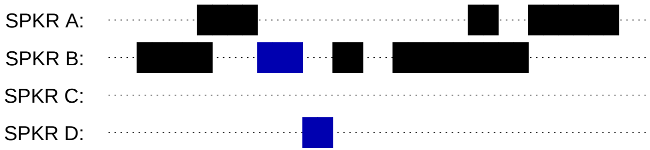
Feedback DA Types

backchannel	b	2.8%
acknowledgment	bk	1.4%
assert	aa	1.1%

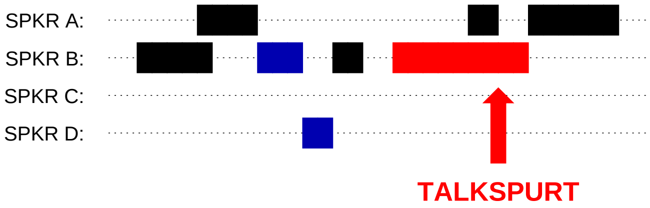
Floor Mechanism DA Types

floor holder	fh	2.5%
floor grabber	fg	0.6%
hold	h	0.3%

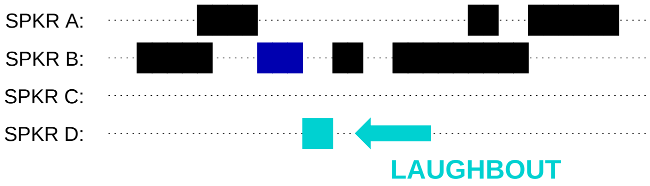
Goal of this Work



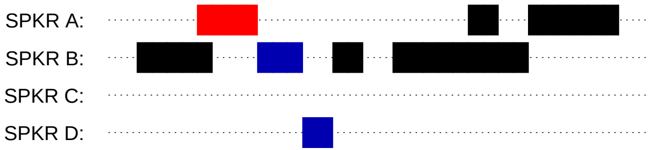
Goal of this Work



Goal of this Work

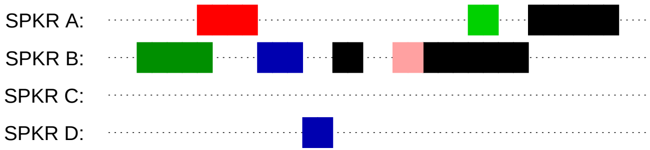


Goal of this Work



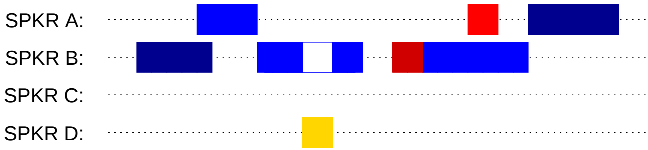
TASK: find speech which is **humor-bearing**

Goal of this Work



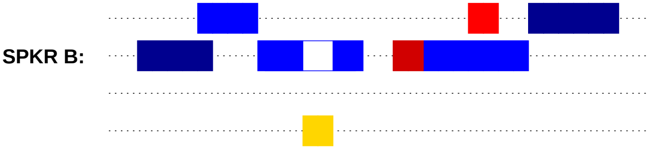
TASK: find speech which is **humor-bearing**
(DA segmentation and recognition, with focus on a subset of DAs)

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



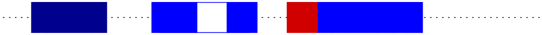
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



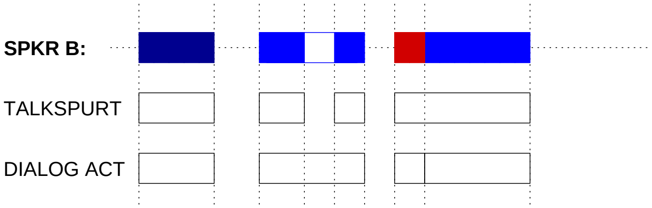
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



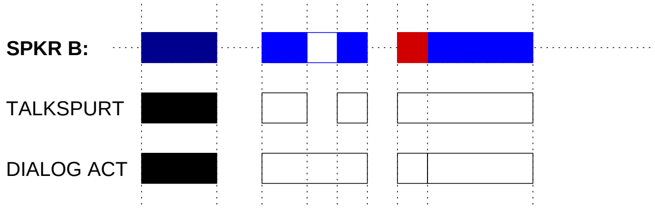
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



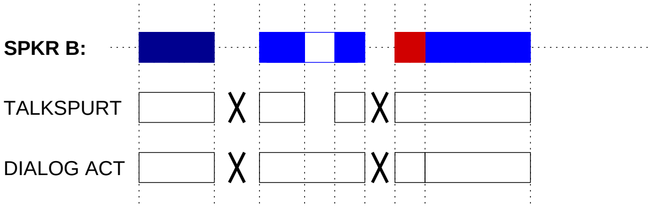
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



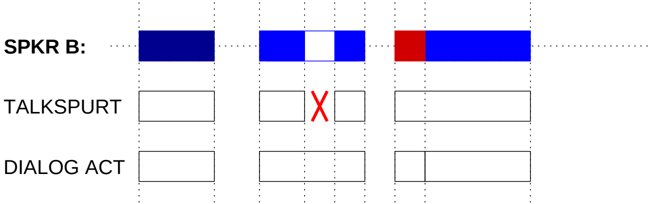
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



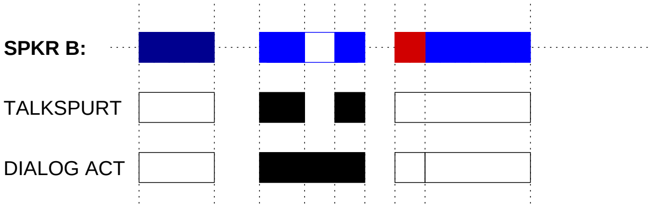
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



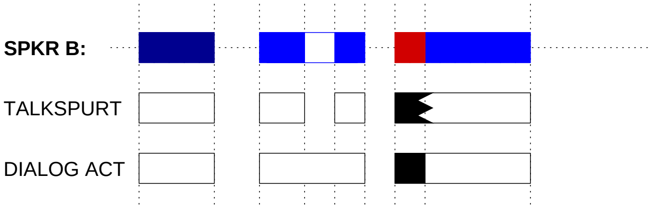
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



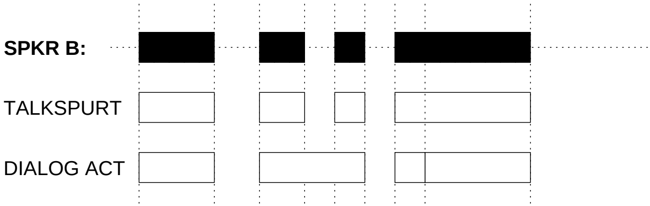
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



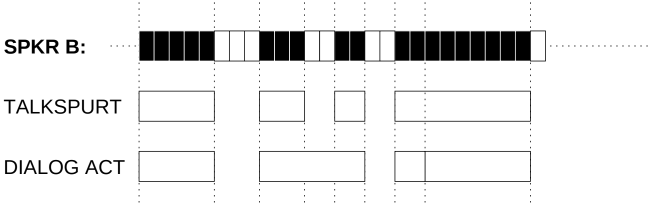
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries ≠ DA Boundaries



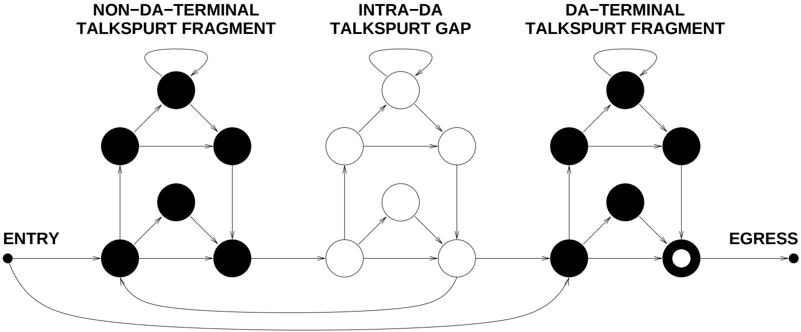
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  ⟶ explicitly model intra-DA silence
- opposite (N:1 correspondence)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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries

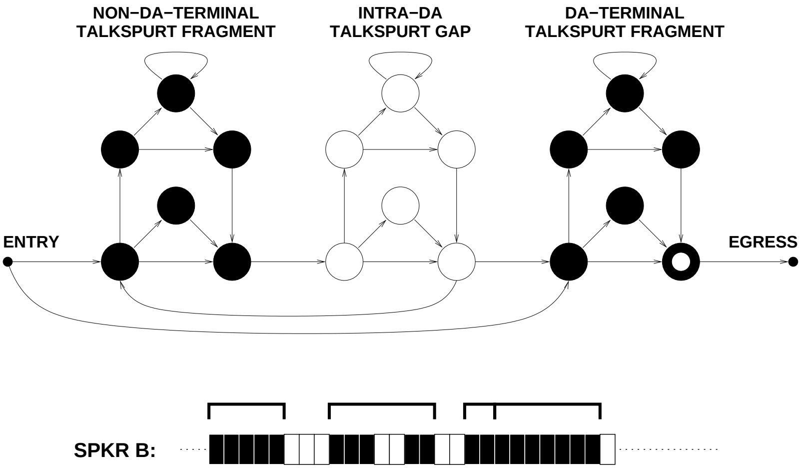


- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

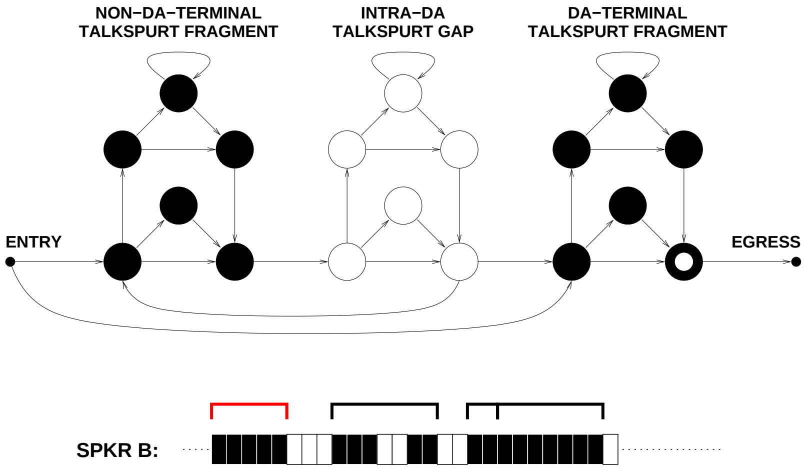
Proposed HMM Sub-Topology for DAs



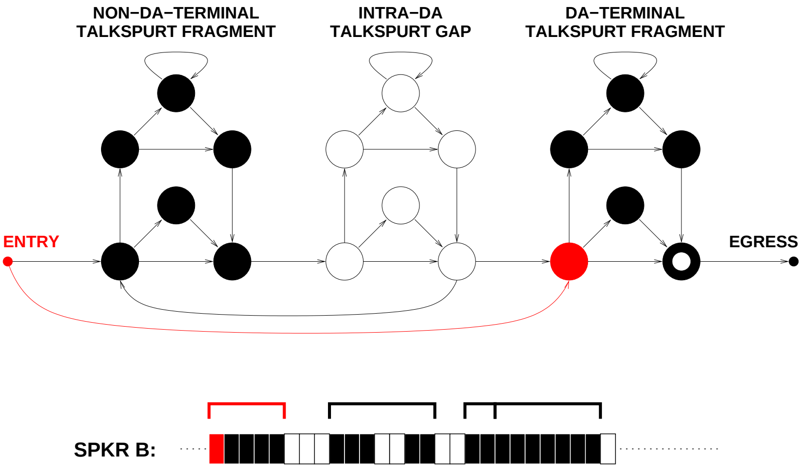
Proposed HMM Sub-Topology for DAs



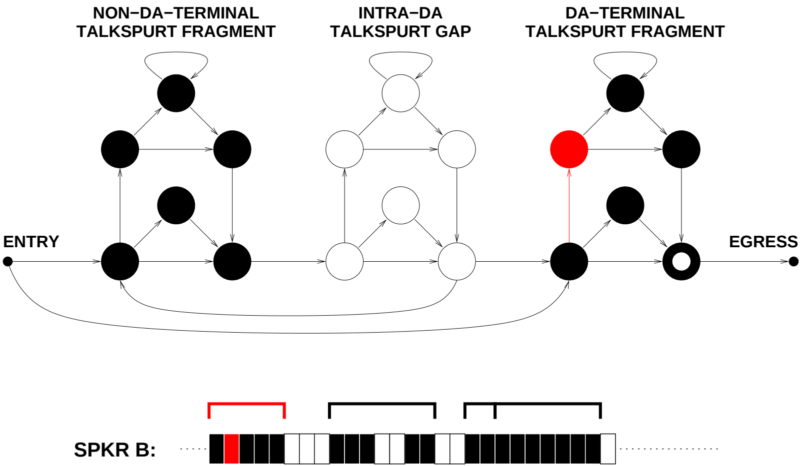
Proposed HMM Sub-Topology for DAs



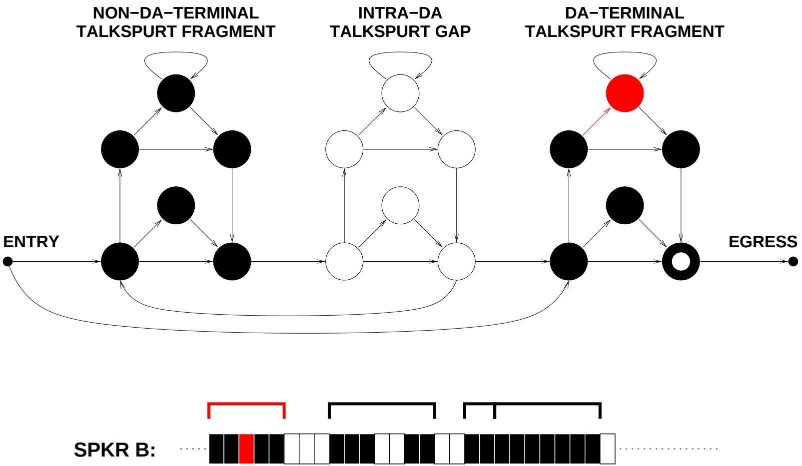
Proposed HMM Sub-Topology for DAs



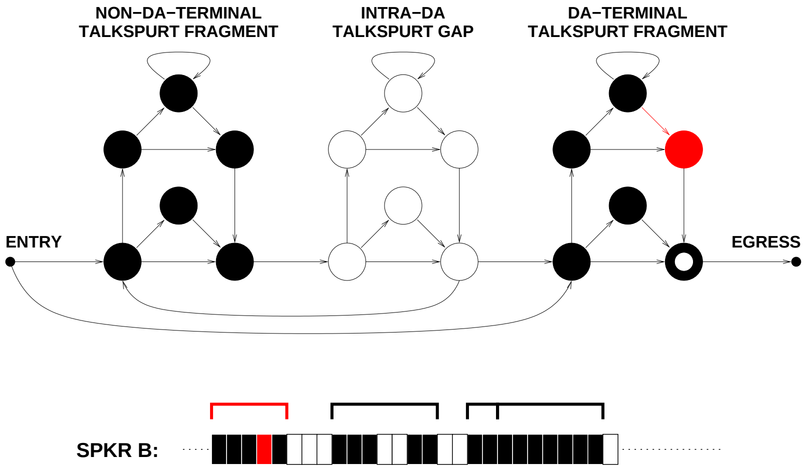
Proposed HMM Sub-Topology for DAs



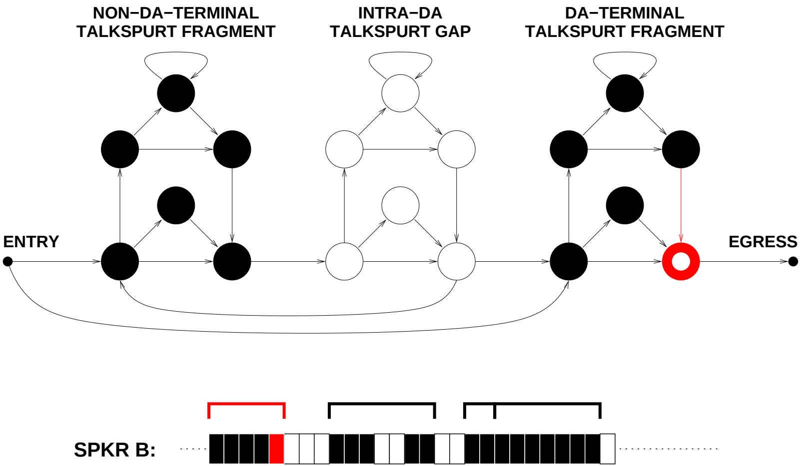
Proposed HMM Sub-Topology for DAs



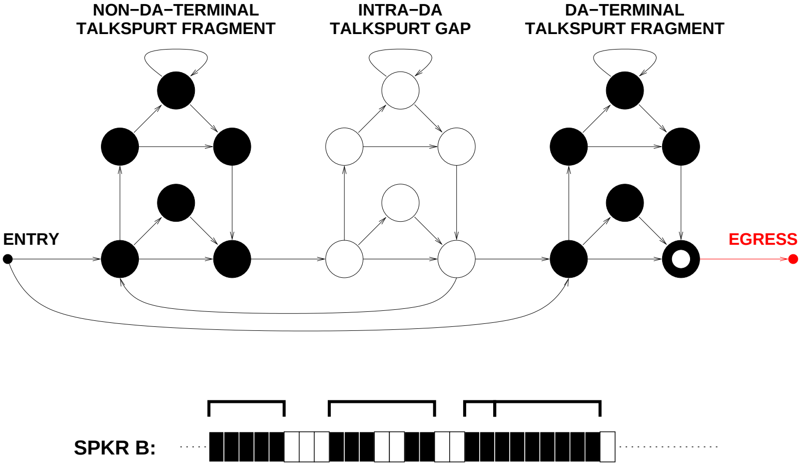
Proposed HMM Sub-Topology for DAs



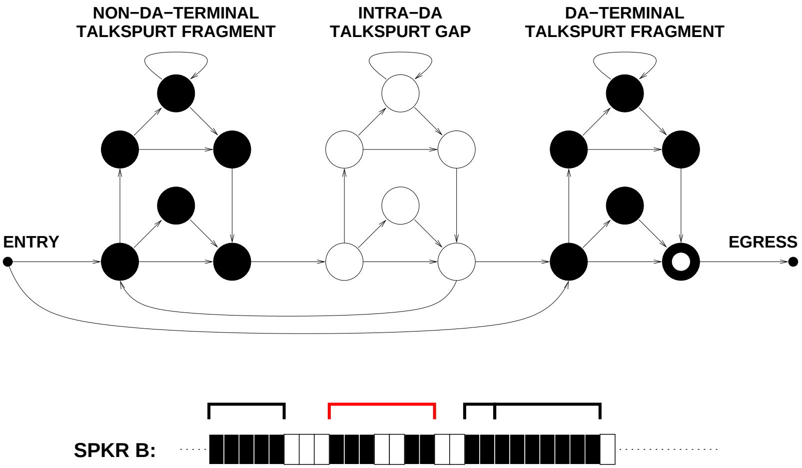
Proposed HMM Sub-Topology for DAs



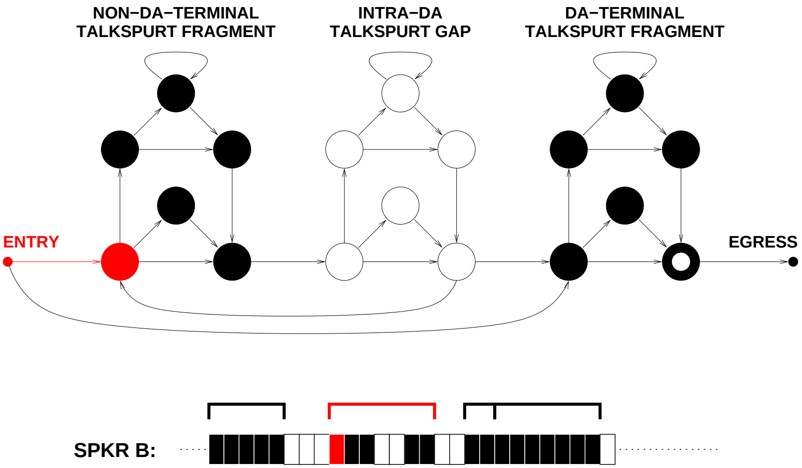
Proposed HMM Sub-Topology for DAs



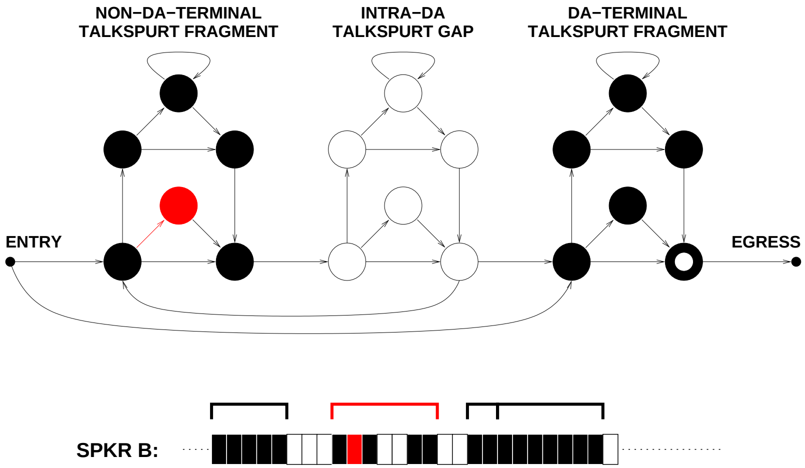
Proposed HMM Sub-Topology for DAs



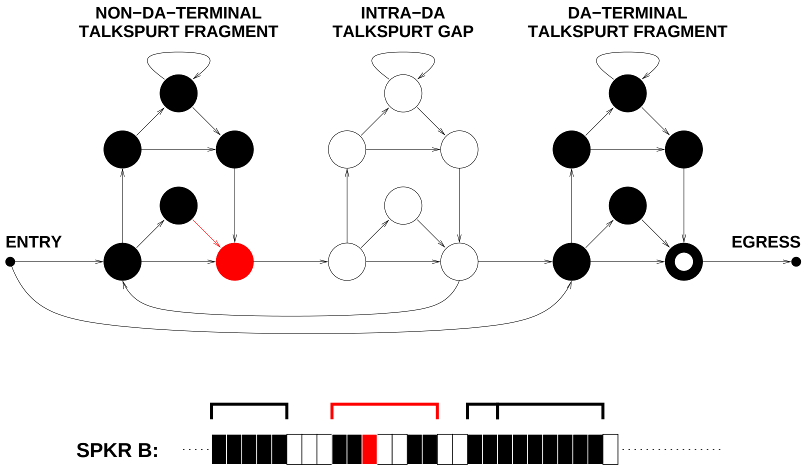
Proposed HMM Sub-Topology for DAs



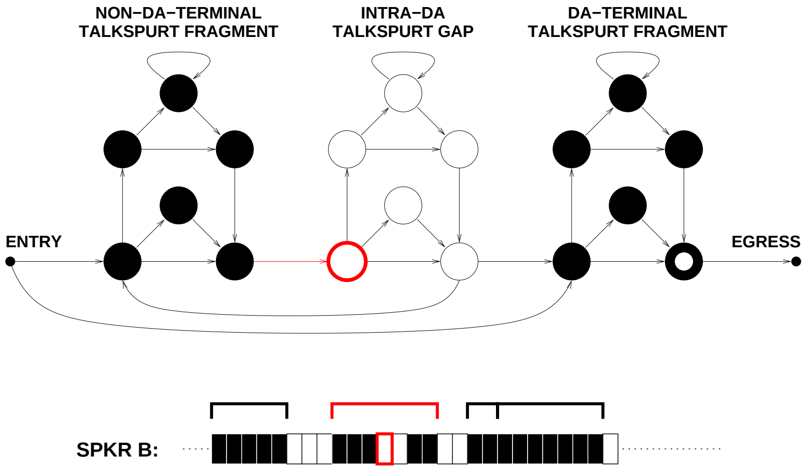
Proposed HMM Sub-Topology for DAs



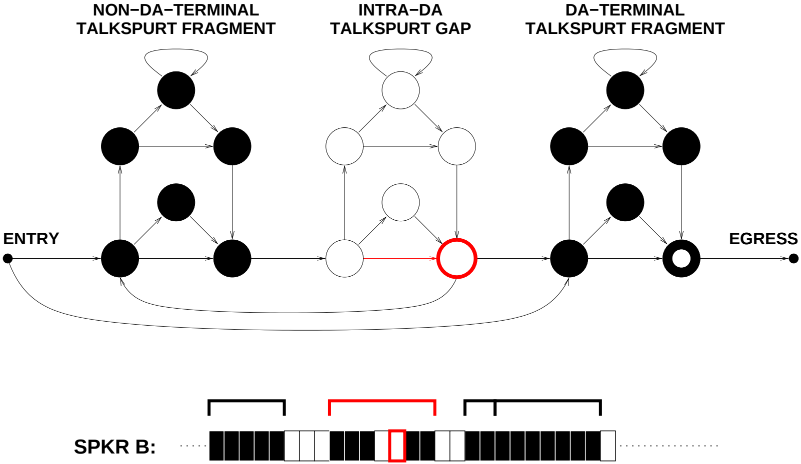
Proposed HMM Sub-Topology for DAs



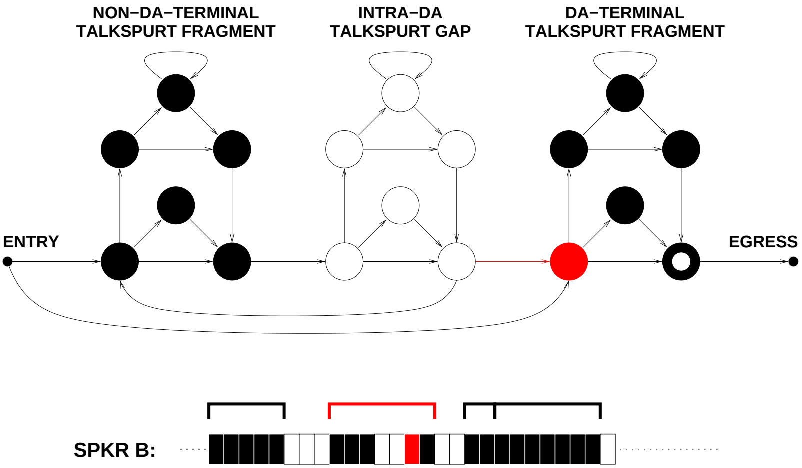
Proposed HMM Sub-Topology for DAs



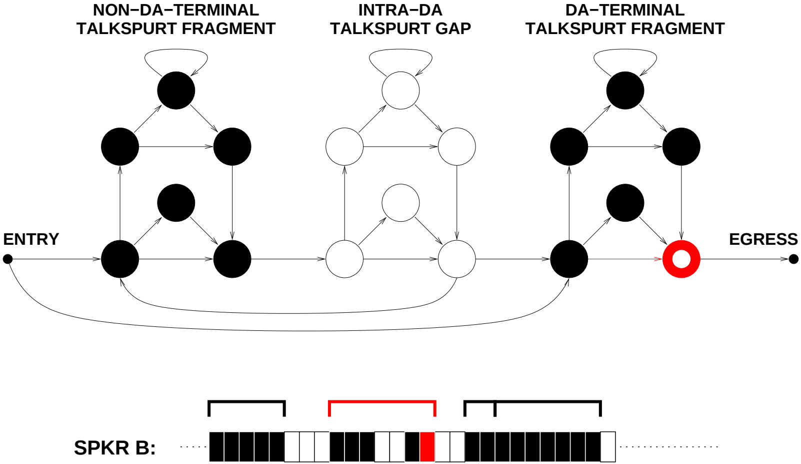
Proposed HMM Sub-Topology for DAs



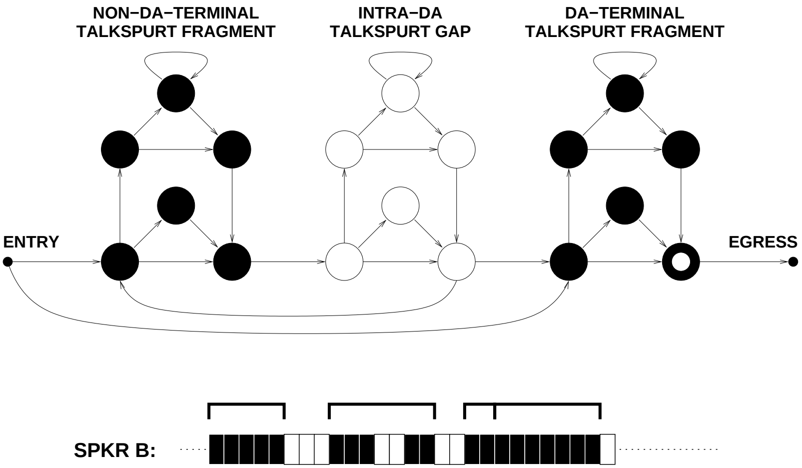
Proposed HMM Sub-Topology for DAs



Proposed HMM Sub-Topology for DAs

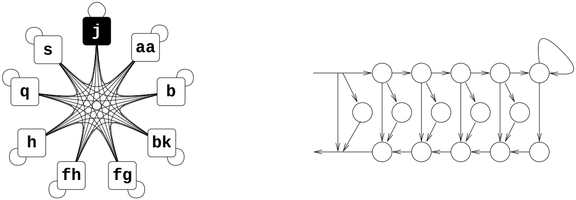


Proposed HMM Sub-Topology for DAs



Proposed HMM Topology for Conversational Speech

- 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

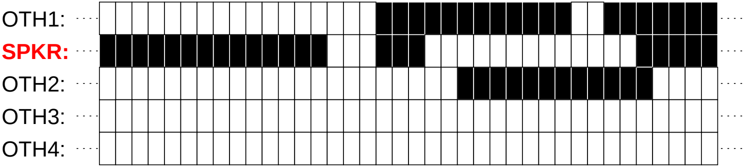
- each 100 ms frame of speech can be assigned to one word w
- assign to that frame the emission probability:
  - of the bigram of which w is the right token, and
  - of the bigram of wihch w is the left token
- train a generative model over left and right bigrams for each HMM state
- bigrams whose probability 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)

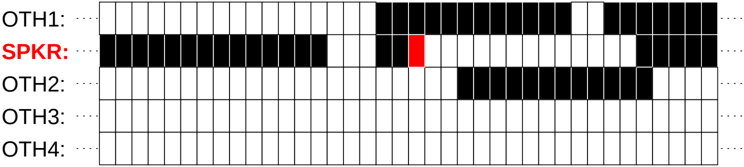
System	DEVSET			EVALSET		
	FA	MS	ERR	FA	MS	ERR
T0	8.1	90.6	98.7	8.3	92.5	100.7
T1	0.3	96.7	97.0	0.2	94.0	94.2
LEX w/o T	53.6	32.8	86.4	53.7	32.9	86.6
LEX w/ T0	40.2	42.9	83.1	40.5	44.2	84.7
LEX w/ T1	12.7	67.0	**79.6**	12.8	70.5	**83.3**

Speech Activity/Interaction Features, $\mathcal{S}$



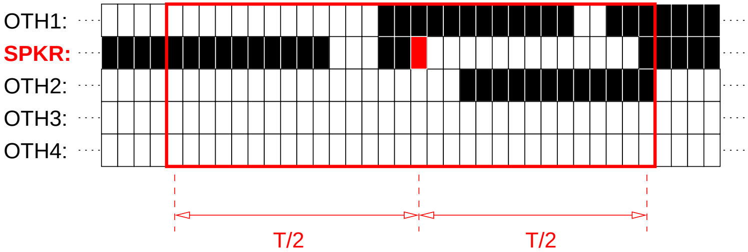
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
- want a fixed-size feature vector: consider only K others
- model features using state-specific GMMs (after LDA)

Speech Activity/Interaction Features, $\mathcal{S}$



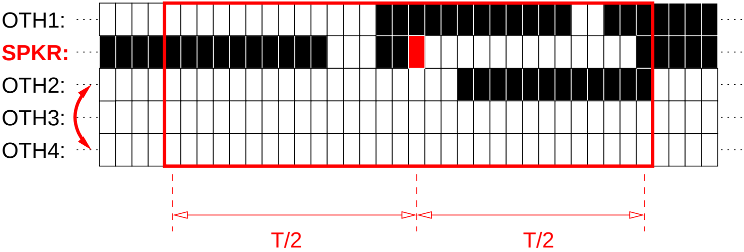
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
- want a fixed-size feature vector: consider only K others
- model features using state-specific GMMs (after LDA)

Speech Activity/Interaction Features, $\mathcal{S}$



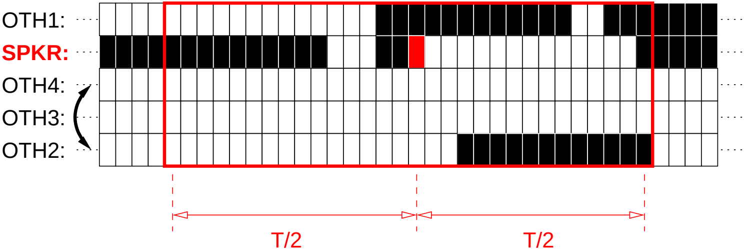
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
- want a fixed-size feature vector: consider only K others
- model features using state-specific GMMs (after LDA)

Speech Activity/Interaction Features, $\mathcal{S}$



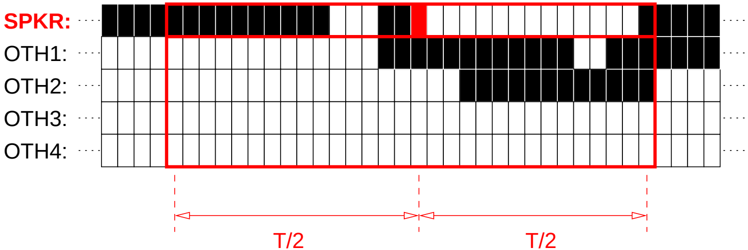
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank “OTH” participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Speech Activity/Interaction Features, $\mathcal{S}$



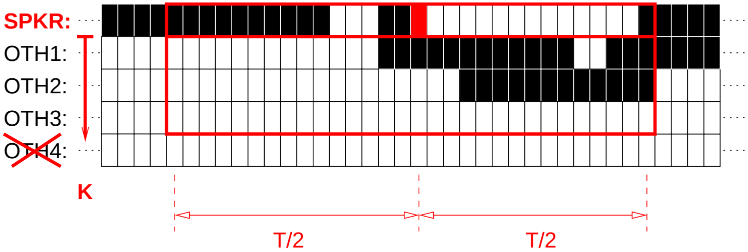
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank "OTH" participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Speech Activity/Interaction Features, $\mathcal{S}$



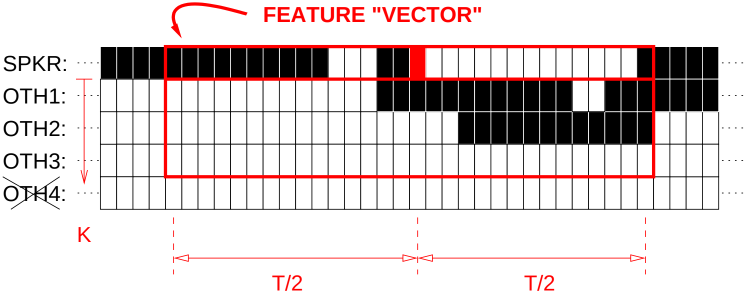
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank "OTH" participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Speech Activity/Interaction Features, $\mathcal{S}$



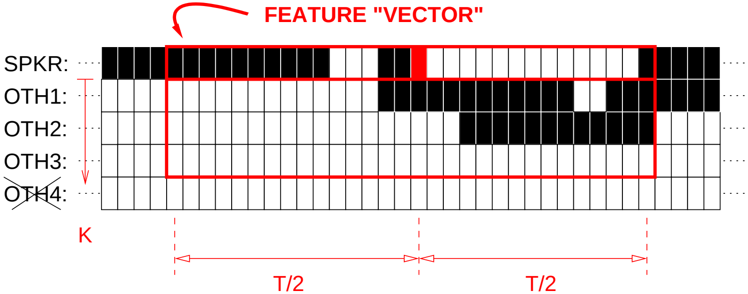
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank “OTH” participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Speech Activity/Interaction Features, $\mathcal{S}$



- 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, $\mathcal{S}$



- 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)

Laughter Activity/Interaction Features, $\mathcal{L}$

- 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 (Ⓜ)

$$
\begin{aligned}
P\left(\left[F_{\mathcal{S}}, F_{\mathcal{L}}\right] \mid \left[\mathcal{M}_{\mathcal{S}}, \mathcal{M}_{\mathcal{L}}\right]\right) &\equiv P\left(F_{\mathcal{S}} | \mathcal{M}_{\mathcal{S}}\right) P\left(F_{\mathcal{L}} | \mathcal{M}_{\mathcal{L}}\right) \\
F_{\mathcal{S}} &= f\left(K, rank\left(\mathcal{S}\right), \mathcal{S}\right) \\
F_{\mathcal{L}} &= f\left(K, rank\left(\mathcal{L}\right), \mathcal{L}\right)
\end{aligned}
$$

2. feature-space combination (Ⓕ)

$$
\begin{aligned}
P\left(\left[F_{\mathcal{S}}, F_{\mathcal{L}}\right] \mid \left[\mathcal{M}_{\mathcal{S}}, \mathcal{M}_{\mathcal{L}}\right]\right) &\equiv P\left(\left[F_{\mathcal{S}}, F_{\mathcal{L}}\right] | \mathcal{M}_{\mathcal{S}\cup\mathcal{L}}\right) \\
F_{\mathcal{S}} &= f\left(K, rank\left(\mathcal{S}\right), \mathcal{S}\right) \\
F_{\mathcal{L}} &= f\left(K, rank\left(\mathcal{L}\right), \mathcal{L}\right)
\end{aligned}
$$

3. feature-computation-space combination (Ⓒ)

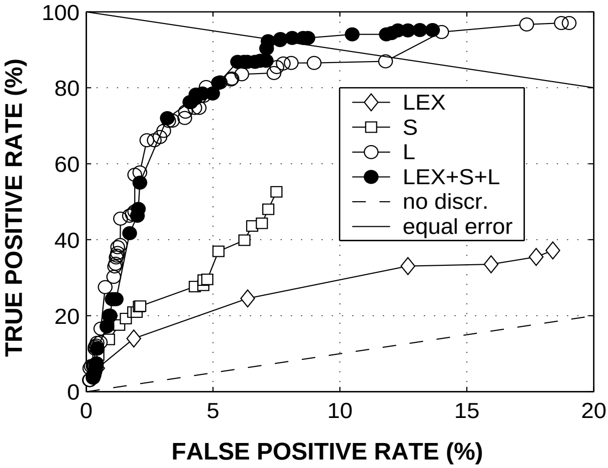
$$
\begin{aligned}
P\left(\left[F_{\mathcal{S}}, F_{\mathcal{L}}\right] \mid \left[\mathcal{M}_{\mathcal{S}}, \mathcal{M}_{\mathcal{L}}\right]\right) &\equiv P\left(\left[F_{\mathcal{S}}, F_{\mathcal{L}}\right] | \mathcal{M}_{\mathcal{S}\cup\mathcal{L}}\right) \\
F_{\mathcal{S}} &= f\left(K, rank\left(\mathcal{S} \cup \mathcal{L}\right), \mathcal{S}\right) \\
F_{\mathcal{L}} &= f\left(K, rank\left(\mathcal{S} \cup \mathcal{L}\right), \mathcal{L}\right)
\end{aligned}
$$

Results

System	DevSet FA	DevSet MS	DevSet ERR	EvalSet FA	EvalSet MS	EvalSet ERR
LEX	12.7	67.0	79.6	12.8	70.5	83.3
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{S}$ Ⓜ $\mathcal{L}$	9.7	6.6	16.3	11.0	8.4	19.4
$\mathcal{S}$ Ⓕ $\mathcal{L}$	6.0	17.8	23.8	6.8	21.6	28.4
$\mathcal{S}$ Ⓒ $\mathcal{L}$	6.0	16.0	22.0	6.4	17.8	24.2
LEX Ⓜ $\mathcal{S}$ Ⓜ $\mathcal{L}$	7.7	7.2	14.8	8.3	11.0	19.4

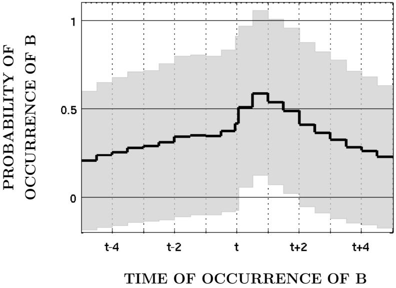
- $\mathcal{L}$ is the best single source of information for this task
- **model-space** combination with $\mathcal{S}$ leads to improvement
- combination with LEX leads to improvement on DevSet only

Receiver Operating Characteristics (DevSet)

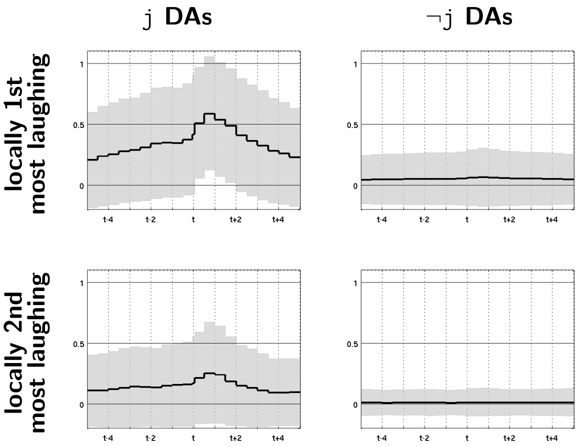


Interpreting Emission Probability Diagrams

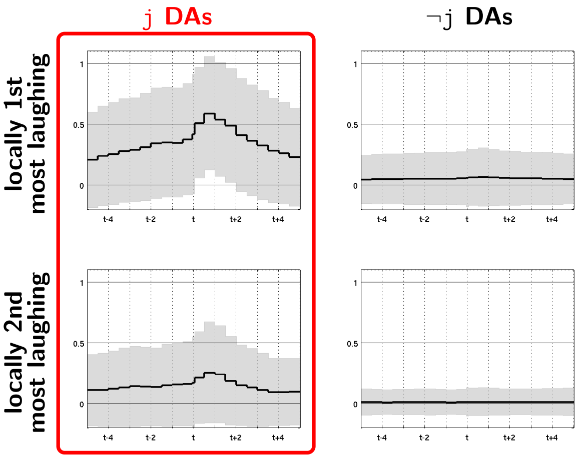
- condition: given an event of type A occurring at time t
- what is the likelihood that an event of type B occurs at time $t' \in [t-5, t+5]$
- retrain *single*-Gaussian model on *unnormalized* features



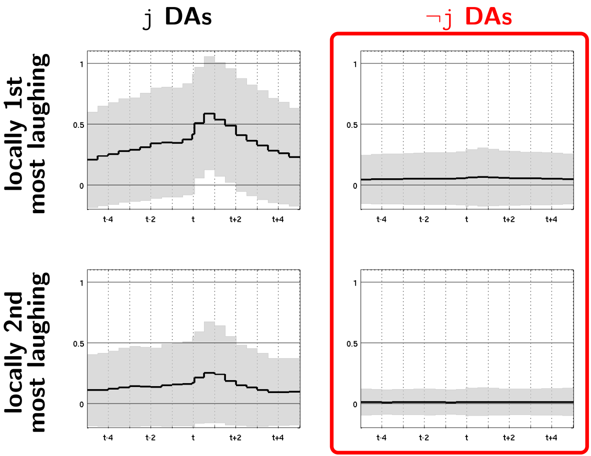
Interlocutor Laughter Context at DA Termination



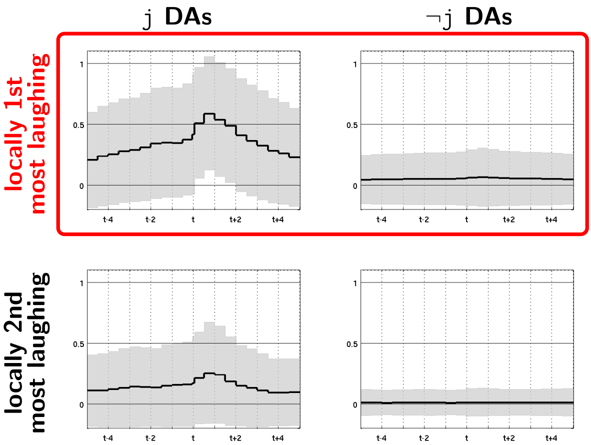
Interlocutor Laughter Context at DA Termination



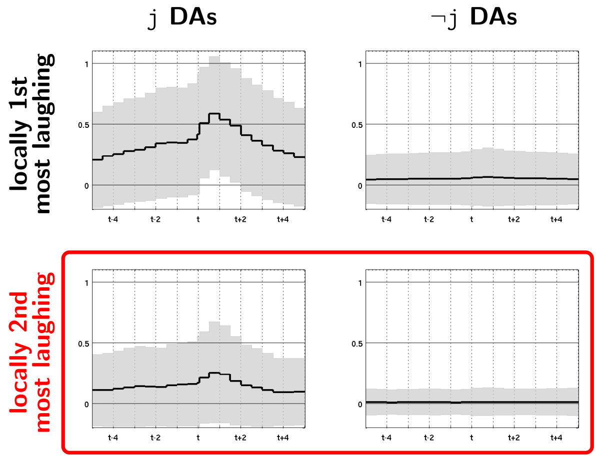
Interlocutor Laughter Context at DA Termination



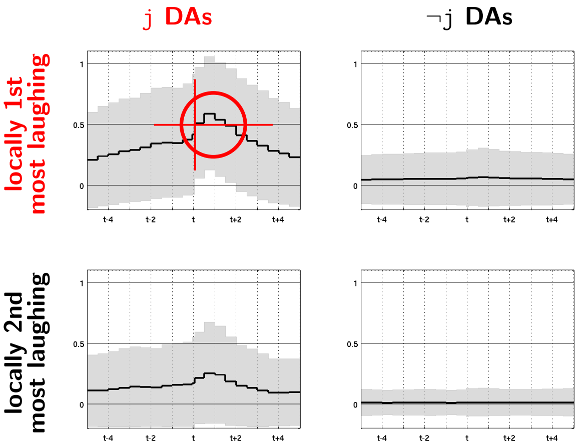
Interlocutor Laughter Context at DA Termination



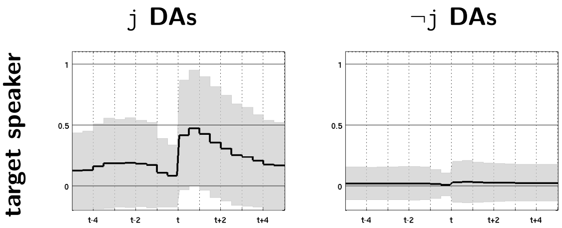
Interlocutor Laughter Context at DA Termination



Interlocutor Laughter Context at DA Termination

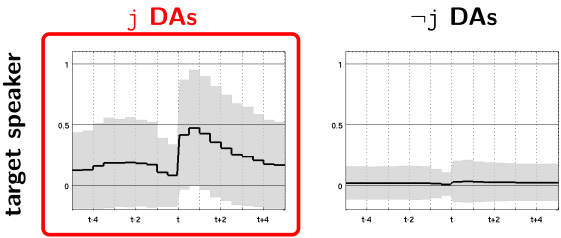


Target Speaker Laughter Context



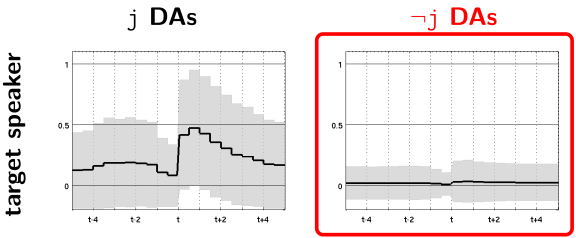
- How well we do with laughter only from the target speaker?

Target Speaker Laughter Context



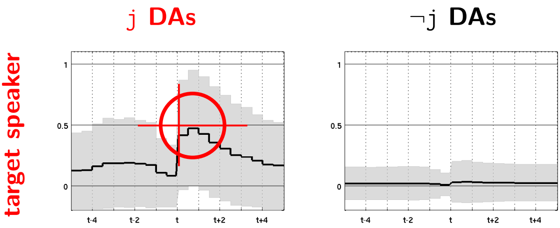
- How well we do with laughter only from the target speaker?

Target Speaker Laughter Context



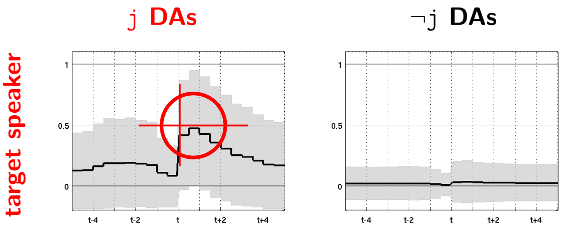
- How well we do with laughter only from the target speaker?

Target Speaker Laughter Context



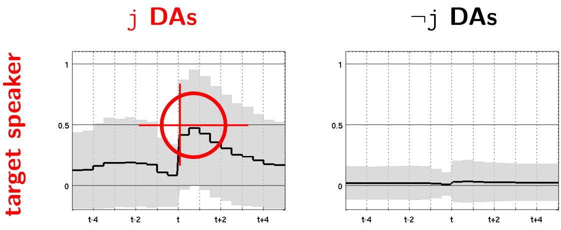
- How well we do with laughter only from the target speaker?

Target Speaker Laughter Context



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

Target Speaker Laughter Context

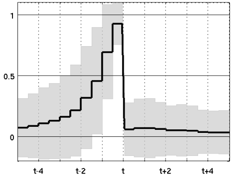


- How well we do with laughter only from the target speaker?

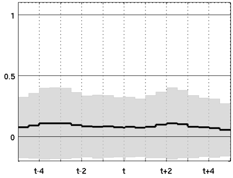
System	DevSet			EvalSet		
	FA	MS	ERR	FA	MS	ERR
$\mathcal{S}$	7.5	47.4	54.9	8.6	62.8	71.4
$\mathcal{L}$	14.0	5.3	19.3	15.6	8.1	23.7
$\mathcal{L}'$	8.7	20.3	**28.9**	8.5	22.4	**31.0**

Interlocutor j-Speech Context at j-DA Termination

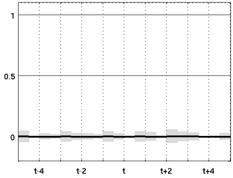
target speaker



locally 1st most j-talkative interlocutor

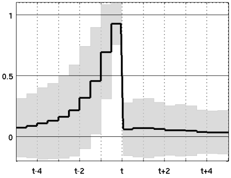


locally 2nd most j-talkative interlocutor

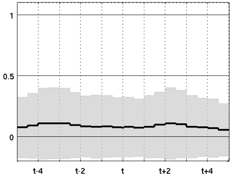


Interlocutor j-Speech Context at j-DA Termination

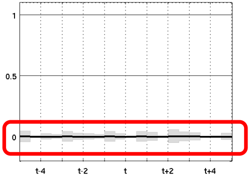
target speaker



locally 1st most j-talkative interlocutor

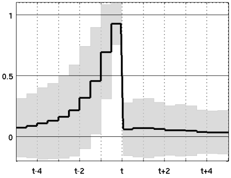


locally 2nd most j-talkative interlocutor

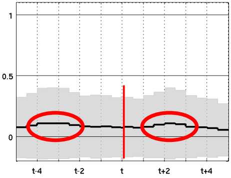


Interlocutor j-Speech Context at j-DA Termination

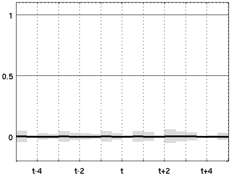
target speaker



locally 1st most j-talkative interlocutor



locally 2nd most j-talkative interlocutor



Summary

- **GOAL**:
  - detect humor-bearing speech
- **APPROACH**:
  - frame-level HMM decoding
  - consider multiparticipant speech & **laughter** context
- **RESULTS**:
  1. at FPRs of ≈5% (DEVSET):
     - lexical features yield TPRs 4× higher than random guessing
     - speech context yields TPRs 2× higher than lexical features
     - laughter context yields TPRs 2× higher than speech context
  2. laughter context features: EER $<$ 24% (EVALSET)
  3. model-space combination improves EERs by ≈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