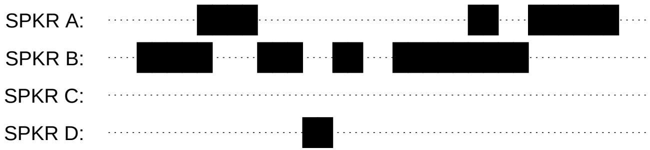
Modeling Other Talkers for Improved Dialog Act Recognition in Meetings

Kornel Laskowski[1] & Elizabeth Shriberg[2,3]

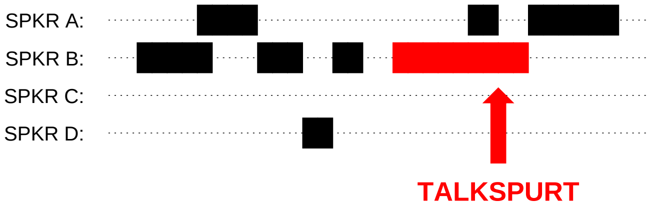
[1]Carnegie Mellon University, Pittsburgh PA, USA
[2]SRI International, Menlo Park CA, USA
[3]International Computer Science Institute, Berkeley CA, USA

10 September, 2008

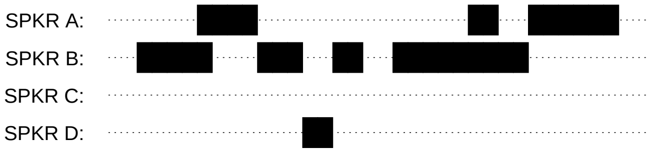
Suppose you're given ...



Suppose you're given ...

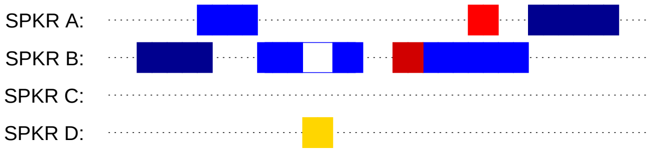


Suppose you're given ...



TASK: segment into dialog acts and classify into dialog act types

Suppose you're given ...



TASK: segment into dialog acts and classify into dialog act types

Why use only speech/non-speech information?

- **sensitive data** in which word information must be masked for privacy reasons
  - Wyatt et al, "Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort", 2007.
- **noisy data** where word recognition performs poorly
- **image-only data** in which speech activity has to be inferred from video only
- **resource-poor languages** in which ASR and/or lexical DA recognizers may be unavailable
- **contexts requiring speed**: SAD is faster than ASR

Why do we care about DAs?

Because sometimes, we want

- to discard specific DA types
  Example 1: summarization systems
  - retain only speech implementing propositional content
- to detect the absence of specific DA types
  Example 2: spoken dialogue systems
  - change strategy when active listening cues not offered
- to detect the presence of specific DA types
  Example 3: discourse analysis systems
  - atypical flooring behavior may indicate grounding problems
- DA segmentation important even when DA classification is not

DA Types in ICSI Meetings

Propositional Content DA Types

- **statement**, `s` (85%)
- **question**, `q` (6.6%)

"Short" DA Types

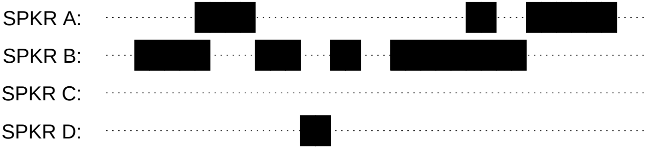
Feedback Types (5.4%)

- **backchannel**, `b` (2.8%)
- **acknowledgment**, `bk` (1.5%)
- **assert**, `aa` (1.1%)

Floor Mechanism Types (3.6%)

- **floor holder**, `fh` (2.7%)
- **floor grabber**, `fg` (0.6%)
- **hold**, `h` (0.3%)

Goal of This Work



Use only speech activity patterns to segment and classify DAs.

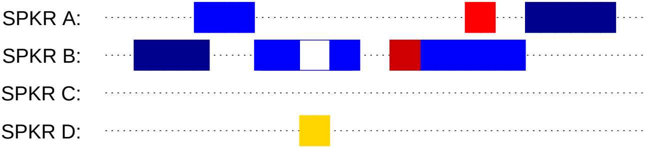
Previous Research on DA Recognition in Meetings

- lots of work, e.g.
  - Ang, Liu & Shriberg, *ICASSP 2005*.
  - Ji & Bilmes, *ICASSP 2005*.
  - Zimmermann, Stolcke & Shriberg, *ICASSP 2006*.
  - Dielmann & Renals, *MLMI 2007*.
- relying on one or more of
  - true DA boundaries (i.e., DA classification only)
  - word identities (true or ASR)
  - word boundaries (true or ASR)
- work in which DA boundaries, word boundaries, and word identities are not assumed has not been done

Previous Research on Talkspurt Modeling in Meetings

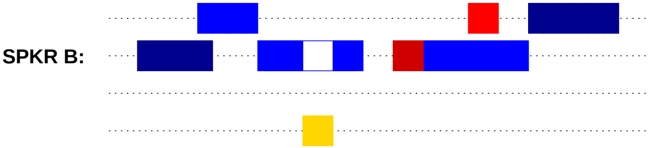
- also lots of work, e.g.
  - Brdiczka, Maisonnasse & Reignier, *ICMI 2005*.
  - Rienks, Zhang, Gatica-Perez & Post, *ICMI 2005*.
  - Laskowski, Ostendorf & Schultz, *SIGdial 2007*.
  - Favre, Salamin, Dines & Vinciarelli, *ICMI 2008*.
- collect and model statistics over long observation intervals
- explicit modeling of speech activity for segmenting and classifying talk in individual talkspurts (and from other participants) has not been done

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

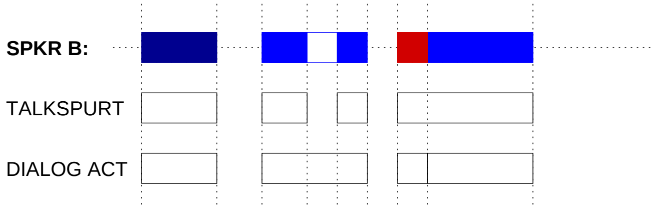


- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries

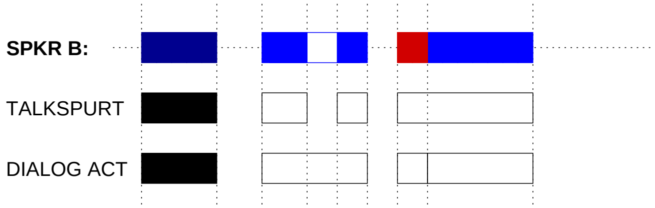
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



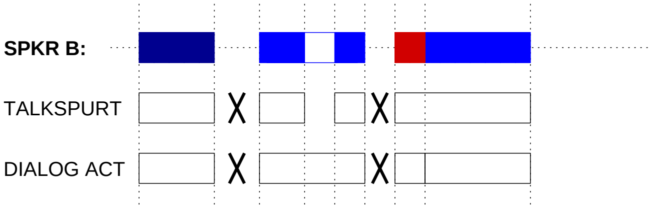
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



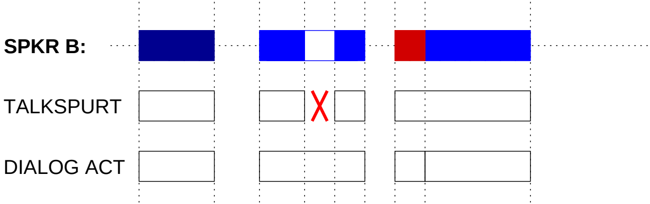
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



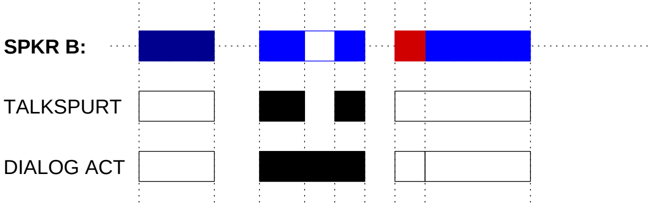
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



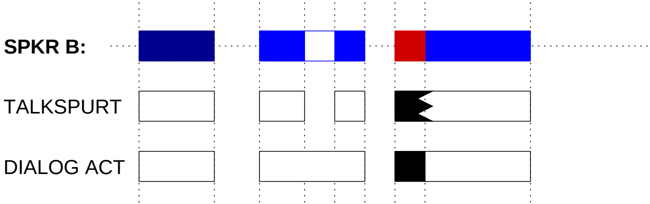
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



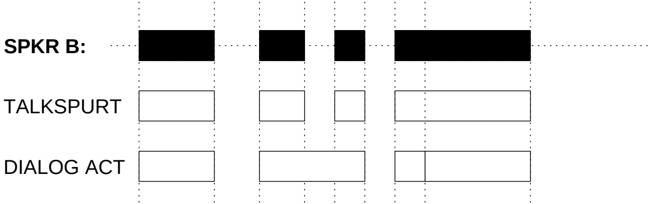
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



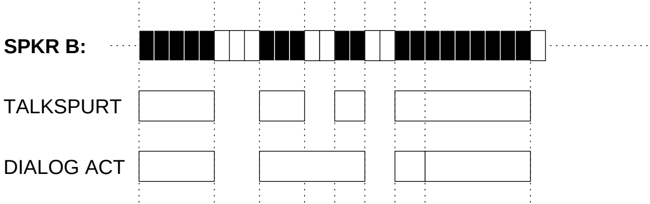
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries



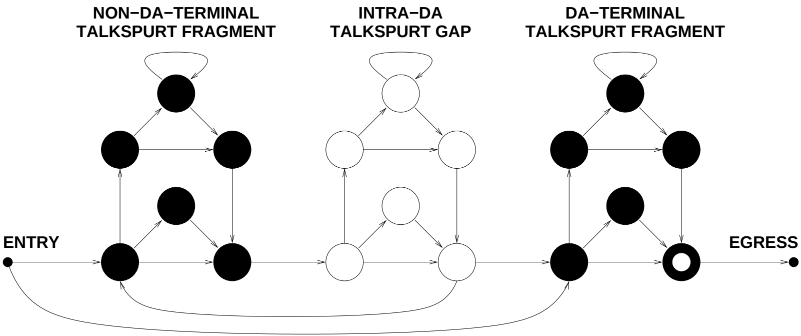
- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

Talkspurt (TS) Boundaries $\neq$ DA Boundaries

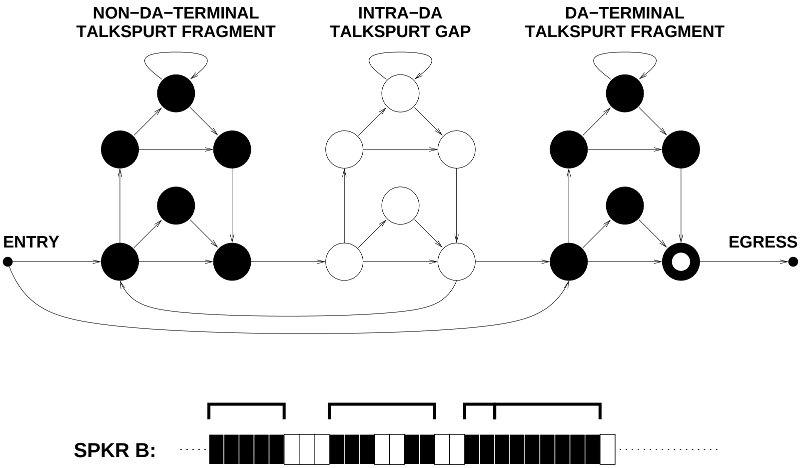


- decoding the state of one participant at a time
- may have 1:1 correspondence between DAs and TSs
- and 1:1 correspondence between DA-gaps and TS-gaps
- but may also have TS gaps **inside** DAs
- 1:N correspondence between DAs and TSs
  $\longrightarrow$ explicitly model intra-DA silence
- opposite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

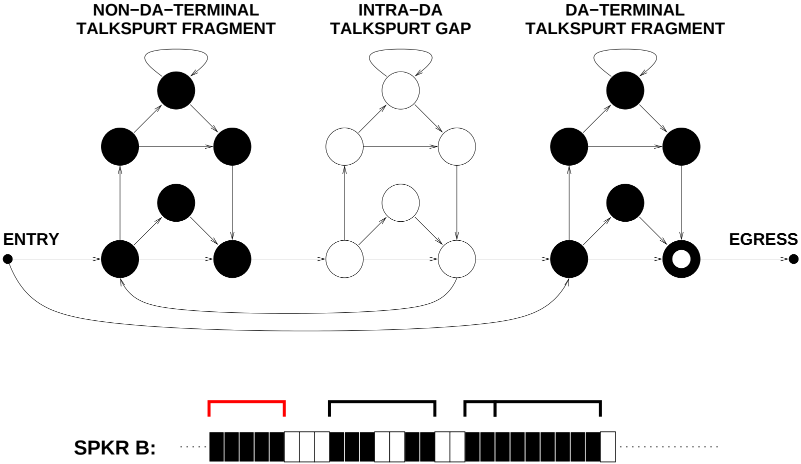
Proposed HMM Sub-Topology for DAs



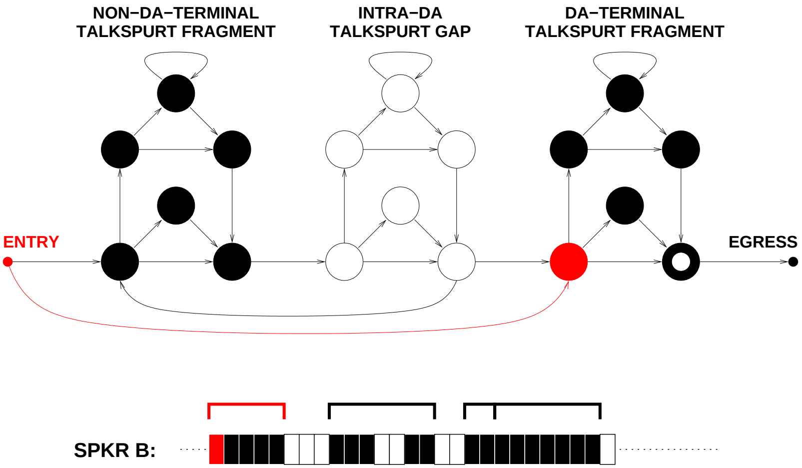
Proposed HMM Sub-Topology for DAs



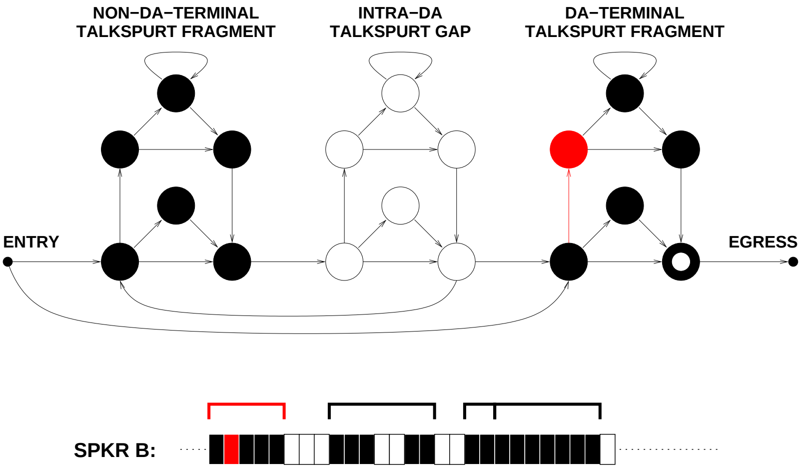
Proposed HMM Sub-Topology for DAs



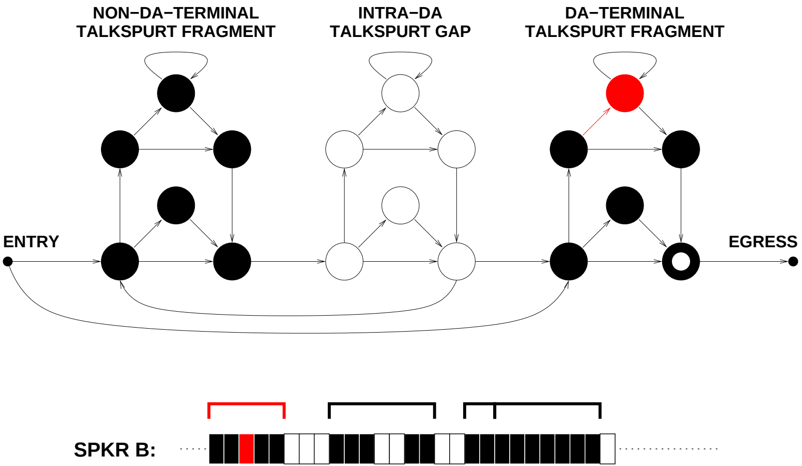
Proposed HMM Sub-Topology for DAs



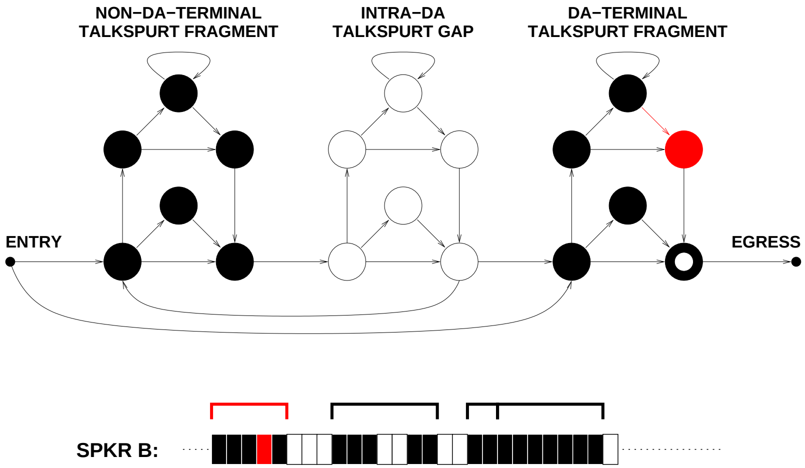
Proposed HMM Sub-Topology for DAs



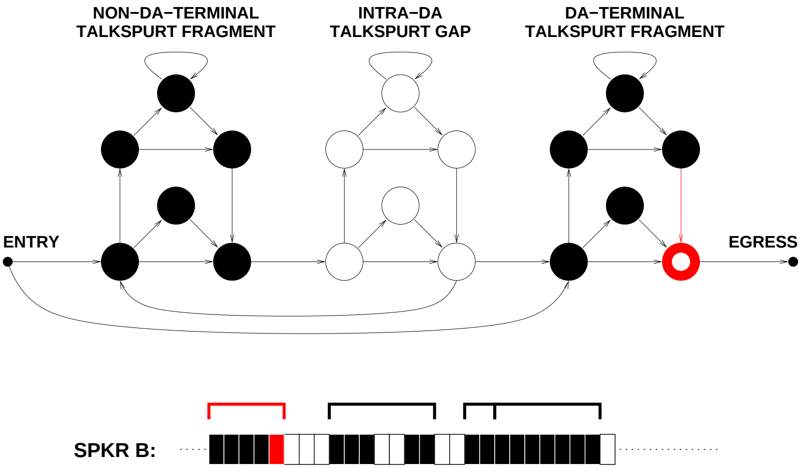
Proposed HMM Sub-Topology for DAs



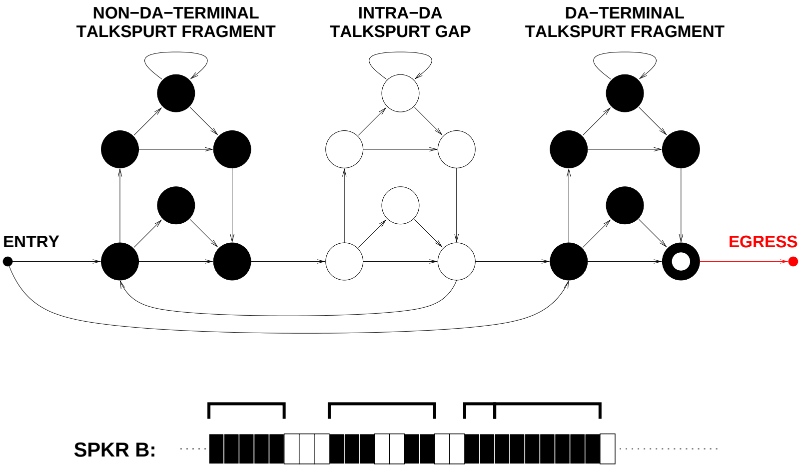
Proposed HMM Sub-Topology for DAs



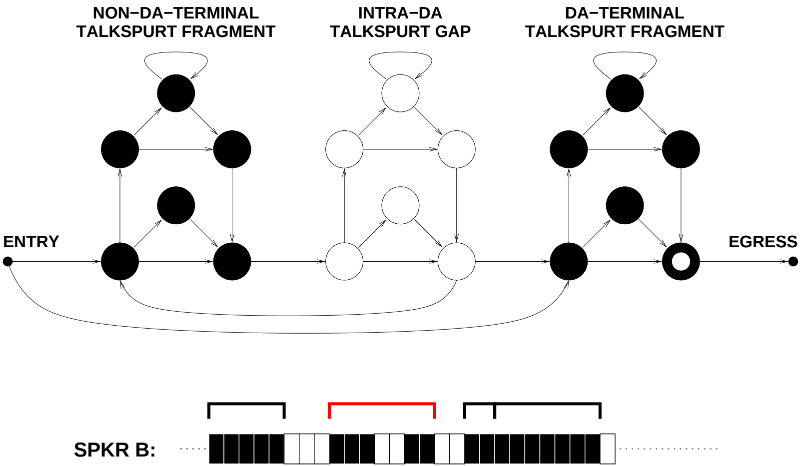
Proposed HMM Sub-Topology for DAs



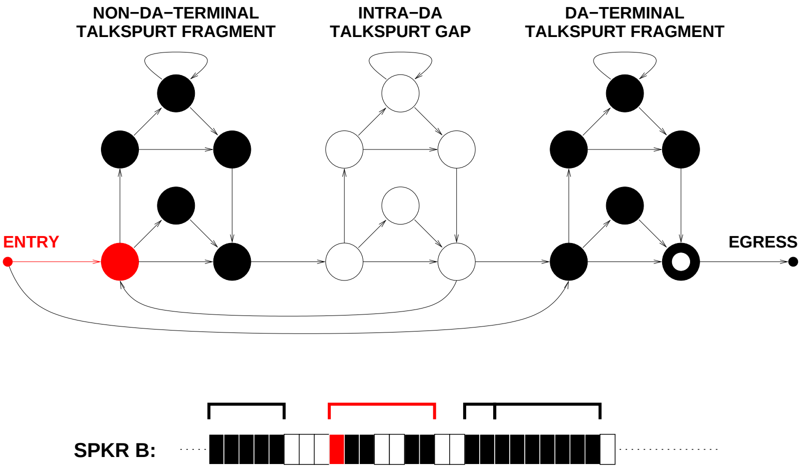
Proposed HMM Sub-Topology for DAs



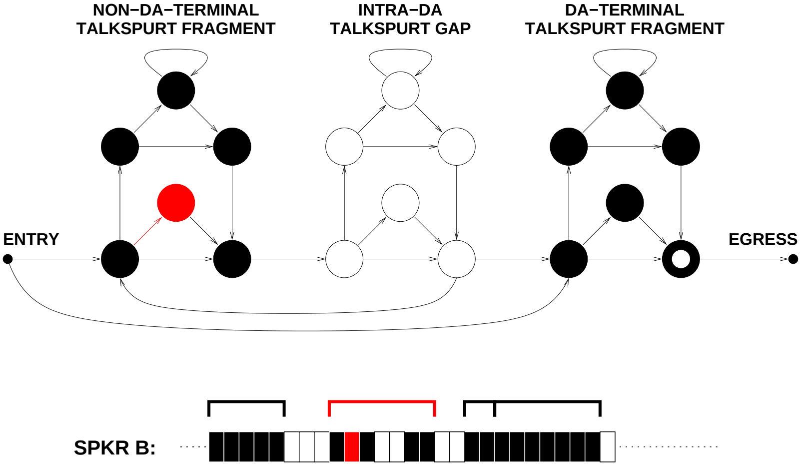
Proposed HMM Sub-Topology for DAs



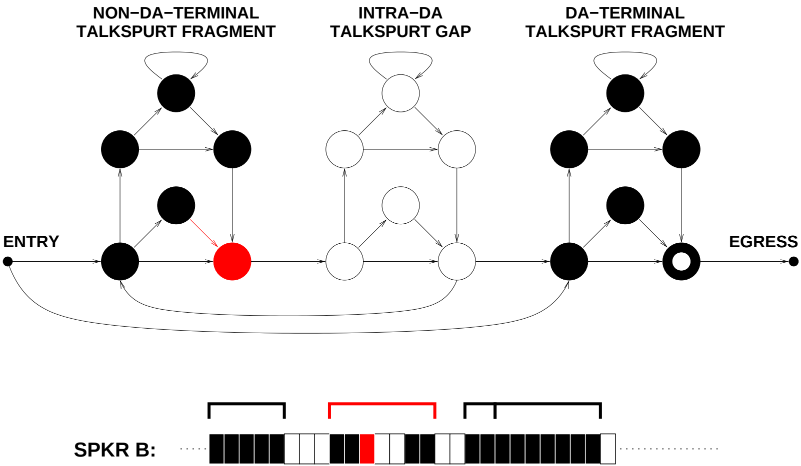
Proposed HMM Sub-Topology for DAs



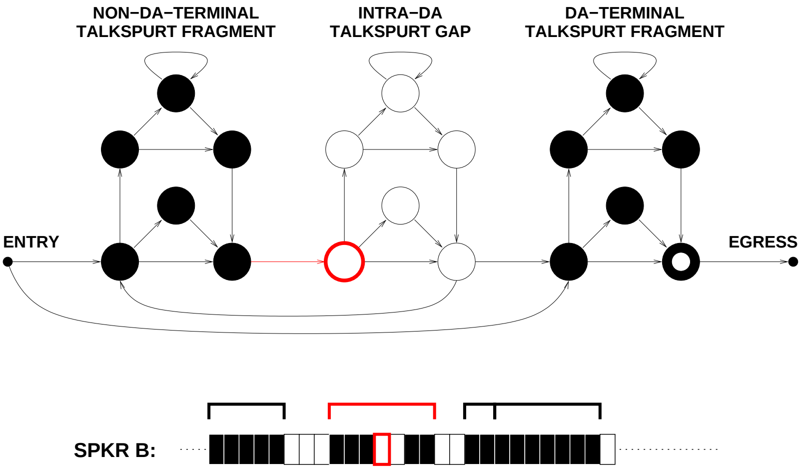
Proposed HMM Sub-Topology for DAs



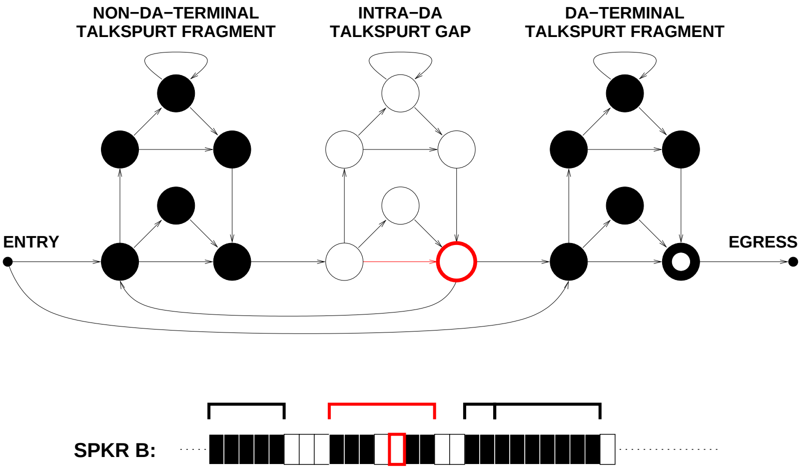
Proposed HMM Sub-Topology for DAs



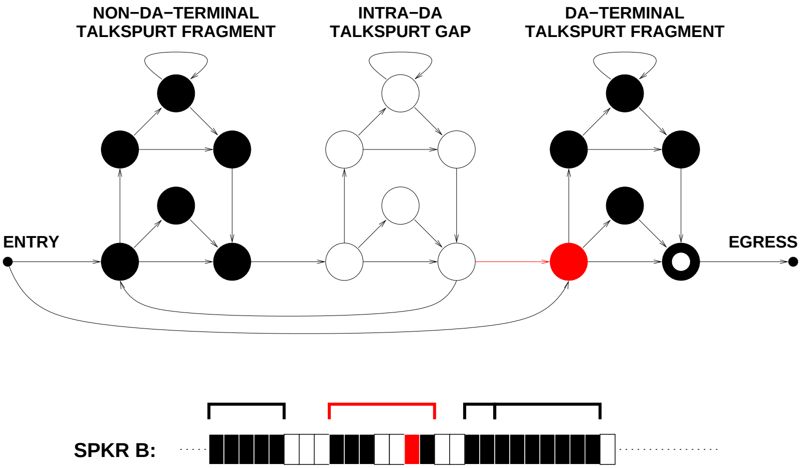
Proposed HMM Sub-Topology for DAs



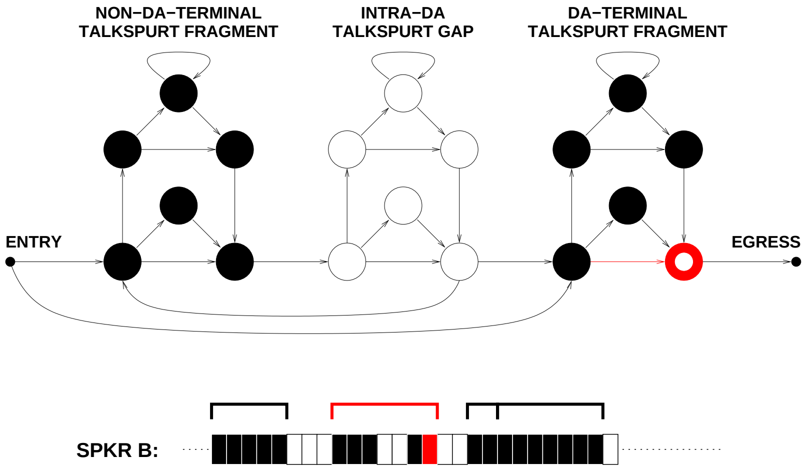
Proposed HMM Sub-Topology for DAs



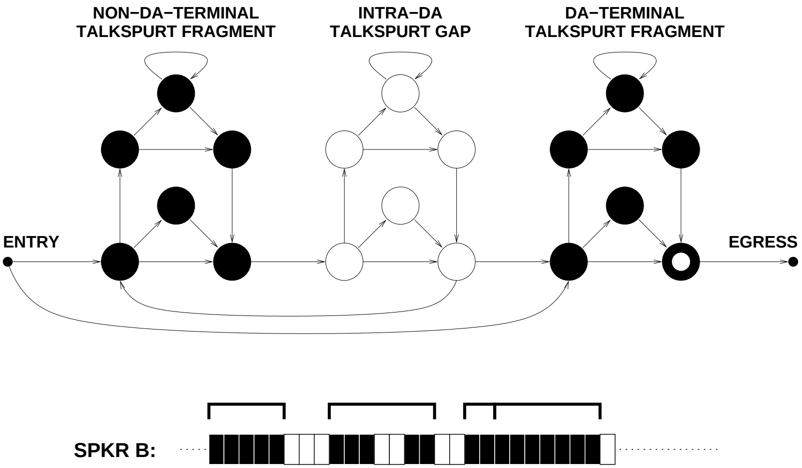
Proposed HMM Sub-Topology for DAs



Proposed HMM Sub-Topology for DAs

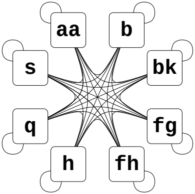


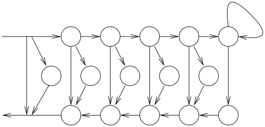
Proposed HMM Sub-Topology for DAs



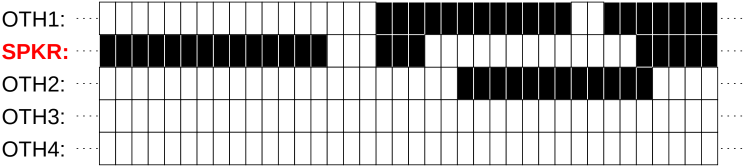
Proposed HMM Topology for Conversational Speech

- the complete topology consists of
  - a DA sub-topology for each of 8 DA types
  - fully connected via **inter-DA GAP subnetworks**



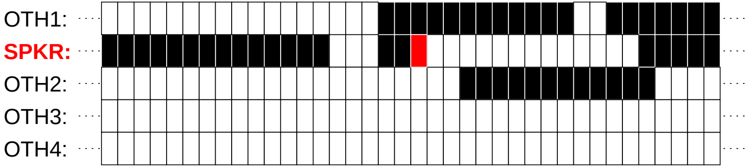


Our HMM Observations



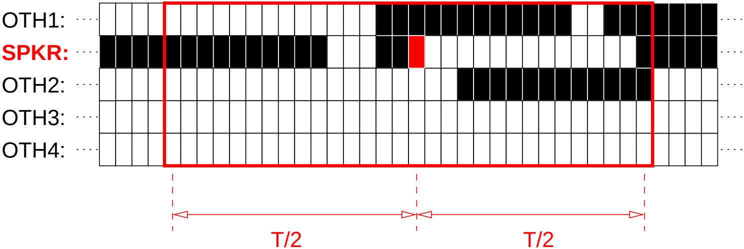
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
- want a fixed-size feature vector: consider only *K* others
- model features using state-specific GMMs (after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

Our HMM Observations



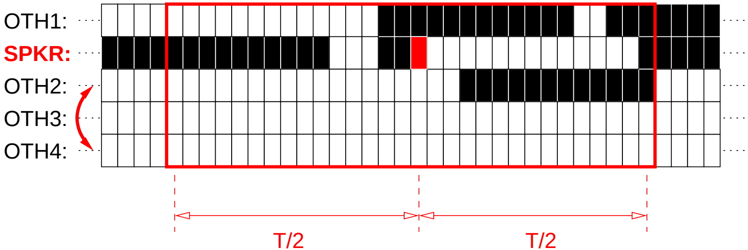
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank "OTH" participants by local speaking time
- want a fixed-size feature vector: consider only K others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

Our HMM Observations



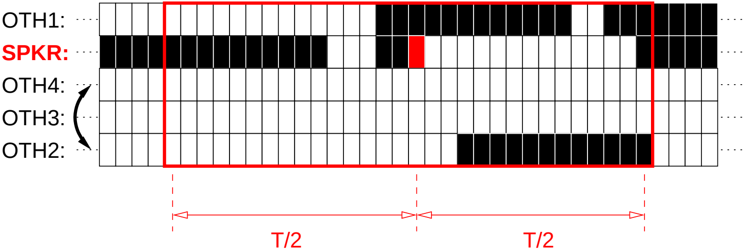
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
- want a fixed-size feature vector: consider only K others
- model features using state-specific GMMs (after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

Our HMM Observations



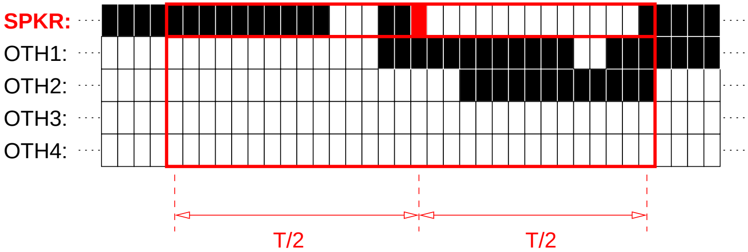
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank "OTH" participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Our HMM Observations



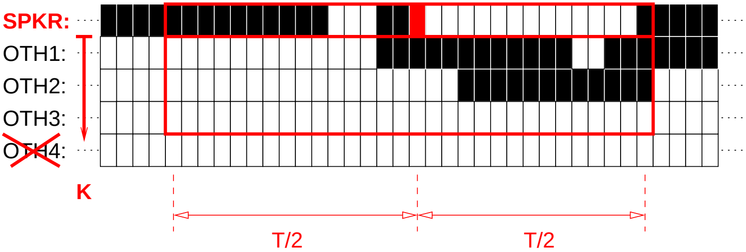
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank "OTH" participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Our HMM Observations



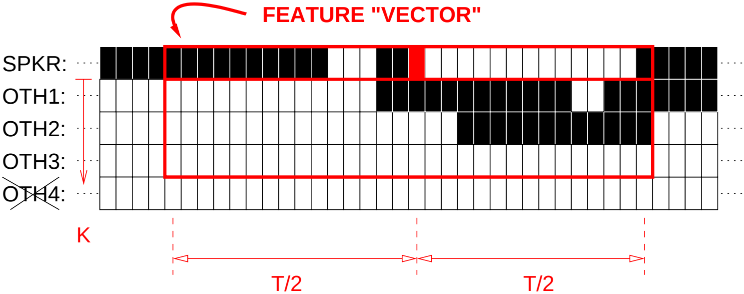
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank "OTH" participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Our HMM Observations



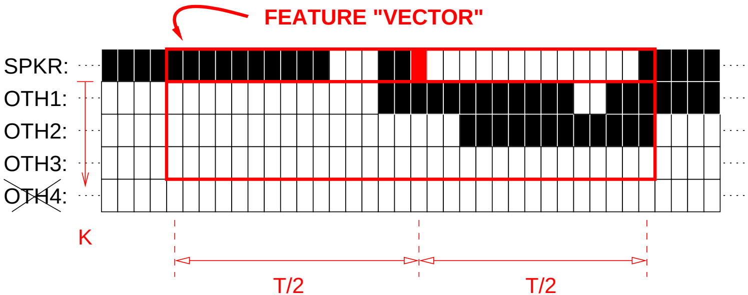
- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank "OTH" participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Our HMM Observations



- 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)

Our HMM Observations

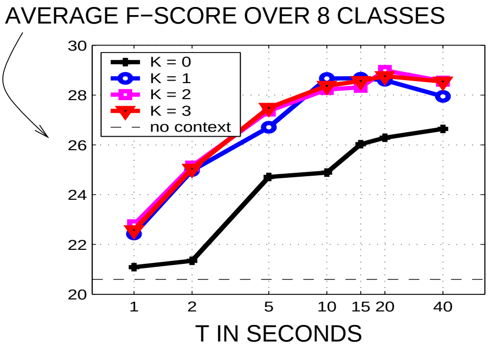


- decoding one participant (SPKR) at a time
- at instant t, model the *thumbnail image* of context
  - consider a temporal context of width T
- want invariance under participant-index rotation
  - rank “OTH” participants by **local** speaking time
- want a fixed-size feature vector: consider only K others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

Experiments

- How well can SAD predict DA boundaries and types?
  - in this work, we decided to use oracle speech activity
  - want to know the inherent information
- three specific questions
  1. Do other talkers matter?
  2. How many others (K) should be considered?
  3. What width (T) of temporal context is needed?
- K and T have a conversation analysis interpretation
  - talk is predominantly one-at-a-time $\longrightarrow$ K is small
  - turns are locally managed $\longrightarrow$ T is small

Effect of Context Size (T) and Number (K) of Interlocutors



- considering $K \geq 1$ most-talkative interlocutors is always better
- considering the $K = 1$ most-talkative suffices
- performance for $K \geq 1$ flattens out as $T \longrightarrow 10$ seconds

Effect of Adding Other Talkers

DA Type		$K=0$		$K=3$	$\Delta F/F_{orig}$
Statement	s	91.4	$\rightarrow$	91.3	-0.08
Question	q	23.4	$\rightarrow$	26.3	$+12.3\dagger$
Backchannel	b	56.7	$\rightarrow$	57.8	$+1.9\dagger$
Acknowledgment	bk	12.6	$\rightarrow$	14.9	$+18.5$
Assert	aa	8.7	$\rightarrow$	13.0	$+49.4\dagger$
Floor holder	fh	21.7	$\rightarrow$	25.6	$+18.3\dagger$
Floor grabber	fg	10.4	$\rightarrow$	13.7	$+31.8$
Hold	h	1.1	$\rightarrow$	6.3	$+485.6\dagger$

- large improvements for all but statements and backchannels
- for backchannels, already doing well at $K=0$

Further Results

1. by adding speech activity, we achieved improvements over a state-of-the-art lexical DA recognizer
    - particularly for floor grabbers, asserts, and questions
    - remarkable because the lexical system uses true words
2. large and significant improvements for DA-terminal phenomena, in particular for interruption ($F = 10.7\% \rightarrow 22.6\%$)

Summary

- GOAL:
  - given only speech/non-speech activity
  - jointly segment and classify into DAs
- APPROACH:
  - frame-level HMM decoding
  - consider (target speaker and) interlocutor activity
- RESULTS:
  - can actually get a lot out of speech/non-speech
  - it's useful to model the other talkers
  - sufficient to consider the single locally most-talkative interlocutor, $K = 1$
  - sufficient to consider a temporal window of $T = 10$ seconds
  - additional benefit: complimentary to lexical information
  - additional benefit: improved recognition of DA termination

THANK YOU

DA Type		LEXICAL		LEXICAL & VOCINT	ΔF (% rel)
Floor grabber	fg	24.5	→	27.0	+9.8*
Hold	h	41.5	→	42.3	+2.0*
Floor holder	fh	63.5	→	64.5	+1.5
Backchannel	b	77.0	→	77.9	+1.1*
Acknowledgment	bk	56.3	→	56.0	−0.5
Assert	aa	40.0	→	42.0	+5.0*†
Question	q	39.8	→	42.5	+6.8*†
Statement	s	93.3	→	93.5	+0.2*†
Interruption		21.9	→	34.1	+56.0*†
Abandonment		13.0	→	14.4	+10.3 †
Termination		69.1	→	69.6	+0.7 †