Predicting, Detecting & Explaining the Occurrence of Vocal Activity in Multi-Party Conversation

Kornel Laskowski

PhD Defense

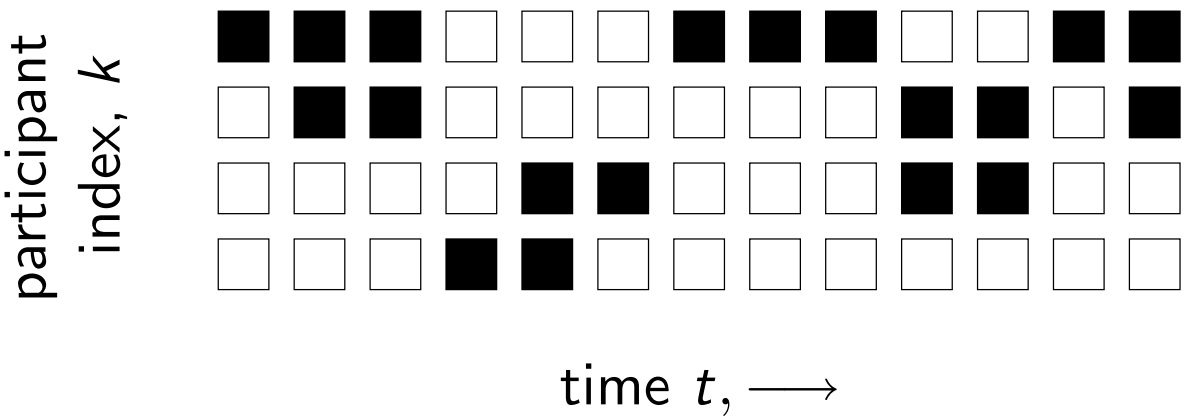
Committee:
Richard Stern, chair
Anton Batliner (FAU)
Alan Black
Alex Waibel

A Multi-Party Conversation

- a **social** event
  - of duration T
  - of $K > 2$ participants
- the predominant activity is **talk**
- **What shapes participants' deployment of talk?**

The Vocal Interaction **Chronogram Q**

(Chapple, 1940; Dabbs & Ruback, 1987)



- ■ ≡ SPEAKING, □ ≡ NOTSPEAKING
- elides content (*"what?"*)
- expresses **form**, via evolving local context
  - chronemics (*"when?"*)
  - attribution (*"who?"*)

Modeling Chronograms

Given a chronogram **Q**, **want the probability** $P(\mathbf{Q})$.

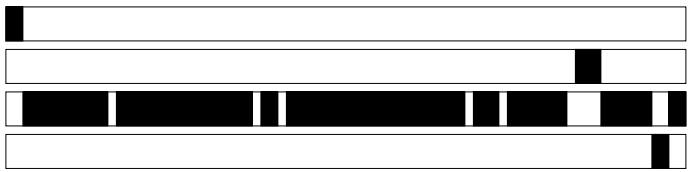
- What does this mean?

Constrastive speech exchange systems (Sacks et al, 1974):

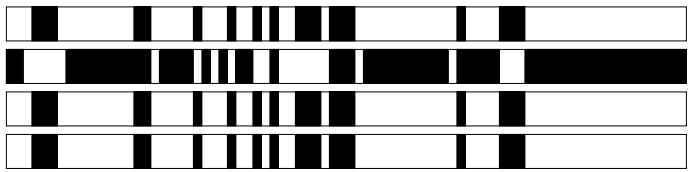
conversation

formal debate

lecture



ritual



Why Model Chronograms?

1. $P(\mathbf{Q})$ can represent a **time- independent and participant-independent** prior for speech activity detection
   - like a language model yields a prior for speech recognition
2. $P(\mathbf{Q}|\mathcal{G})$ can yield a similar prior for **conversational genre** $\mathcal{G}$
   - allows for inference of "what genre $\mathcal{G}$ is this conversation?"
3. $P(\mathbf{Q}|t)$ yields a **time-dependent** prior
   - allows for inference of "what is happening at instant t?"
4. $P(\mathbf{Q}|k)$ yields a **participant-dependent** prior
   - allows for inference of "what is the role of participant k?"

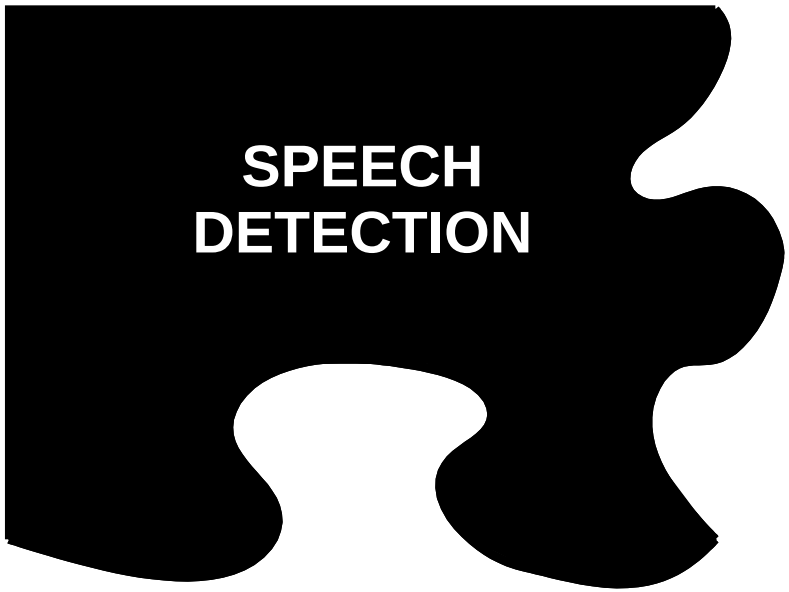
Past Work on Modeling Chronograms

- interaction chronography (Chapple, 1939; Chapple, 1949)
- **modeling in dialogue**: $K = 2$
  - telecomminications (Norwine & Murphy, 1938; Brady, 1969)
  - sociolinguistics (Jaffe & Feldstein, 1970)
  - psycholinguistics (Dabbs & Ruback, 1987)
  - dialogue systems (cf. Raux, 2008)
- **modeling in multi-party settings**: $K > 2$
  - qualitative: Conversation Analysis (Sacks et al, 1974)
  - quantitative: **THIS THESIS**

How to Model Multi-Party Chronograms?

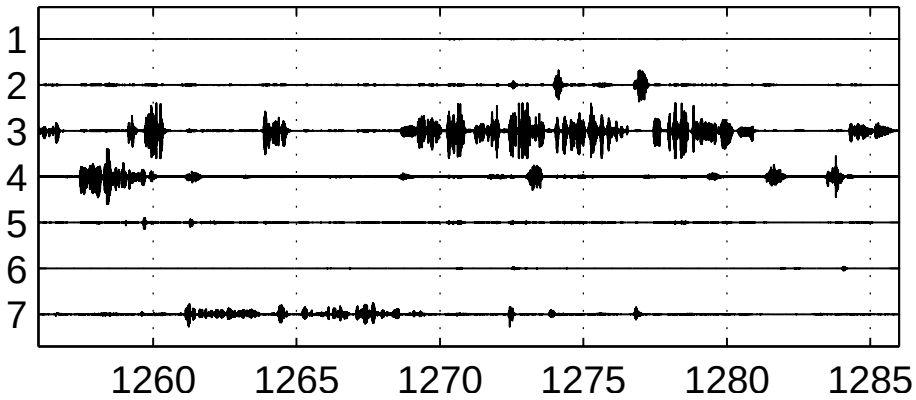
That depends very much on the task.

1. acoustic detection
   1. speech
   2. laughter
2. intent recognition
   1. dialog acts
   2. attempts to amuse
3. participant characterization
   1. diffuse social status
   2. assigned role

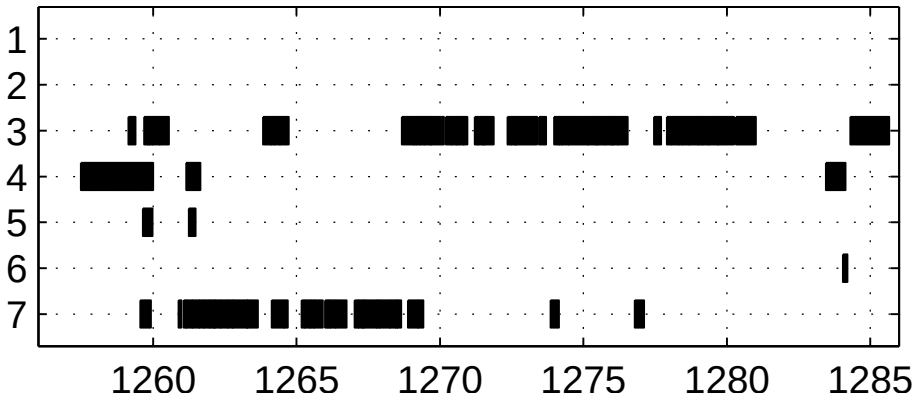

SPEECH
DETECTION

The Goal of Speech Activity Detection (SAD)

Given multichannel nearfield audio **X**:



Produce multi-participant speech chronogram **Q** :



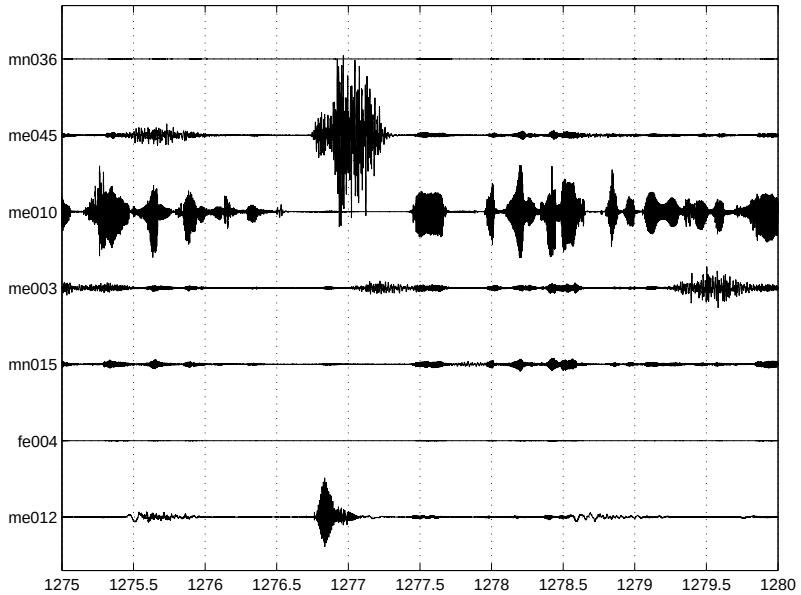
Prior Research in SAD in Multi-Party Meetings

- nearfield, HMM-based speech activity detection (Acero, 1994)
- in meetings: ASR segmentation (Pfau, Ellis & Stolcke, 2001)
  - **crosstalk** is the most serious problem
- in meetings: multiple microphone states
  - 3 states (Huang & Harper, 2005)
  - 4 states (Wrigley et al, 2005)
- in meetings: crosstalk suppression
  - energy normalization (Boakye & Stolcke, 2006)
  - echo cancellation (Dines et al, 2006)
- all of this work decodes participants **one at a time**

The Standard SAD Baseline

- hidden Markov model decoder
- topology enforced minimum duration constraints
  - 16 ms frame step
  - 500 ms for speech ■
  - 500 ms for non-speech □
- acoustic model
  - 32 ms frame size
  - log-energy, MFCCs, Δs, $\Delta\Delta$s (39)
  - Gaussian mixture model (GMM) emissions

The Crosstalk Problem



How Might Chronogram Modeling Help?

Detection is the **inference of the chronogram**:

$$P(\mathbf{Q}|\mathbf{X}) \quad \propto \quad P(\mathbf{X}|\mathbf{Q}) \cdot P(\mathbf{Q})$$

1. treat $\mathbf{Q}$ as a vector-valued process:

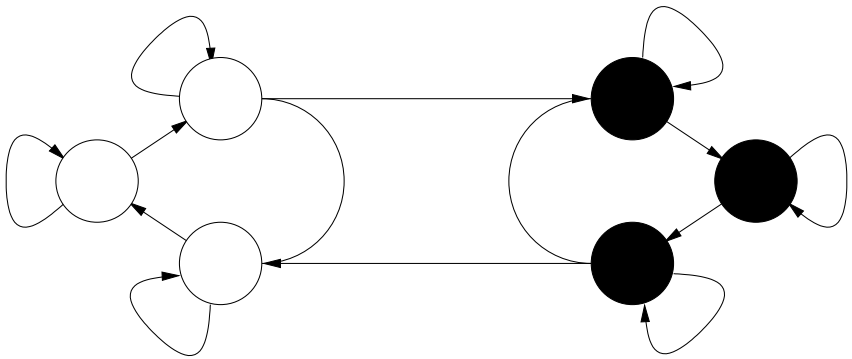
$$\cdots, \quad \mathbf{q}_{t-1} = \begin{bmatrix} \square \\ \square \\ \blacksquare \\ \square \end{bmatrix}, \quad \mathbf{q}_t = \begin{bmatrix} \blacksquare \\ \square \\ \square \\ \square \end{bmatrix}, \quad \mathbf{q}_{t+1} = \begin{bmatrix} \blacksquare \\ \square \\ \square \\ \square \end{bmatrix}, \quad \cdots$$

2. assume process is 1st-order Markovian:

$$P(\mathbf{Q}) \quad = \quad \prod_{t=1}^{T} P(\mathbf{q}_t|\mathbf{q}_{t-1})$$

The Multi-Participant State Space

- if the topology $\mathbb{T}$, of N states, for a single participant is



- then the K-participant topology is the Cartesian product of $\mathbb{T}$

$$\mathbf{q} \in \mathbb{T} \times \mathbb{T} \times \cdots \times \mathbb{T}$$

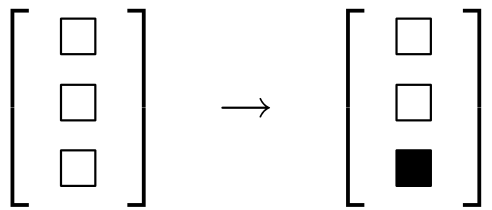
- the number of multi-participant states is N^K

Joint Transition Model: Degree of Overlap

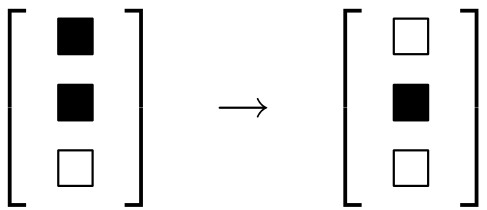
Want the transition from $\mathbf{q}_{t-1}$ to $\mathbf{q}_t$ to be:

- invariant to participant index rotation
- independent of number K of participants

3. replace $\mathbf{q}$ with $\|\mathbf{q}\|$, the **number** of speaking participants



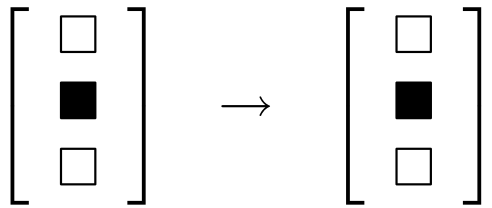
0 → 1



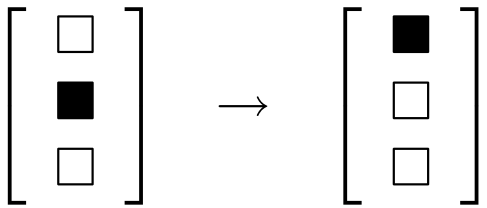
2 → 1

Joint Transition Model: **Extended** Degree of Overlap

Unfortunately,

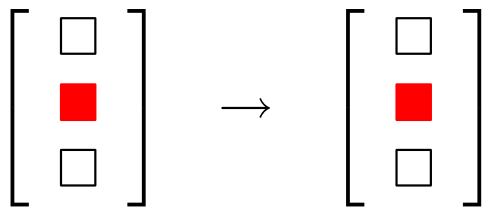


$1 \rightarrow 1$

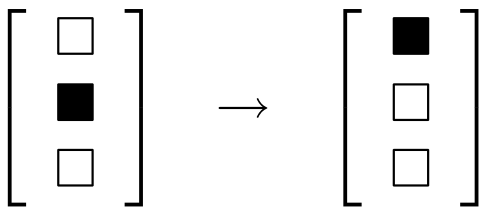


$1 \rightarrow 1$

4. augment "to"-state with **number of same participants** speaking at both t and $t-1$



$1 \rightarrow [1,1]$



$1 \rightarrow [0,1]$

Joint Acoustic Model

Can assume multi-channel acoustics to be independent,

$$P(\mathbf{X}|\mathbf{Q}) = \prod_{k=1}^{K} P(\mathbf{X}[k]|\mathbf{Q}[k])$$

but **crosstalk** proves that they are not.

The covariance matrix Σ of log-energy

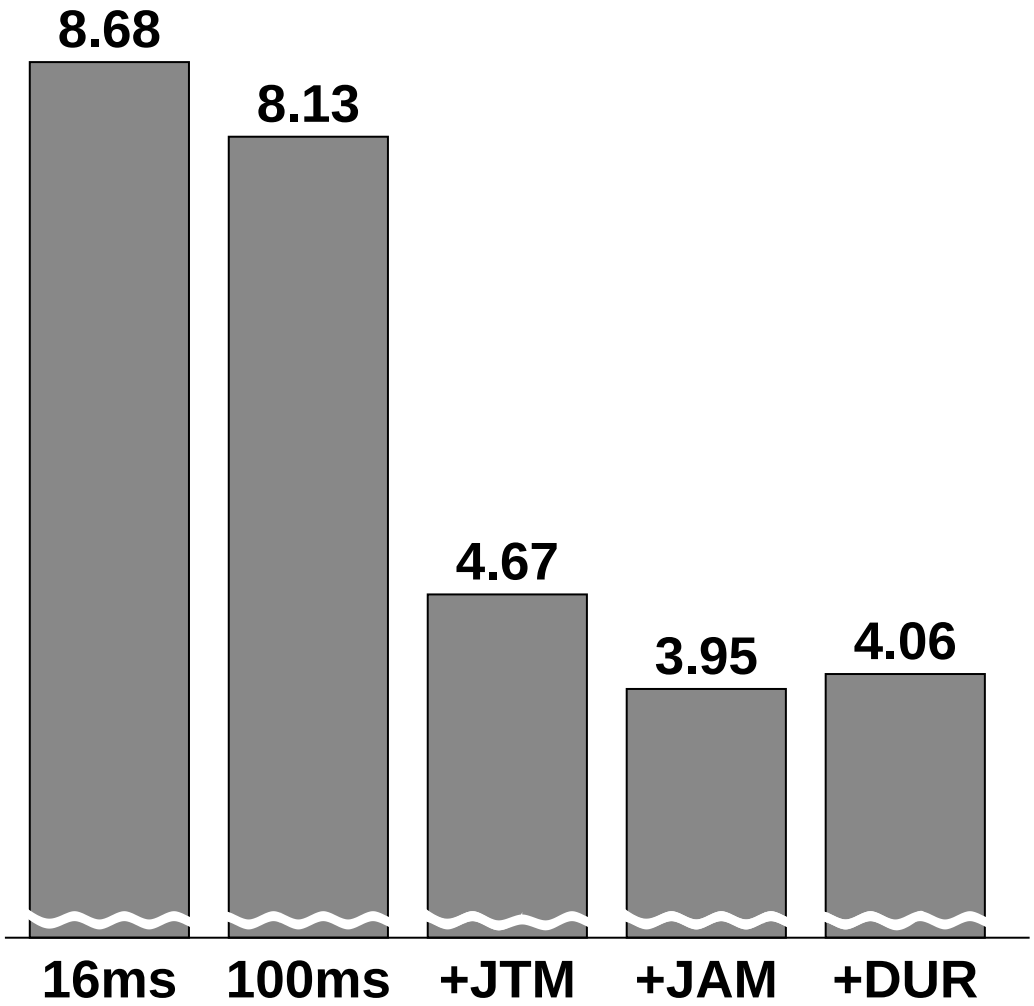
- has size $K \times K$
- off-diagonal entries are non-zero
- off-diagonal entries generalize poorly
  - depend on room acoustics
  - depend on inter-participant proximity

Two-Pass Decoding

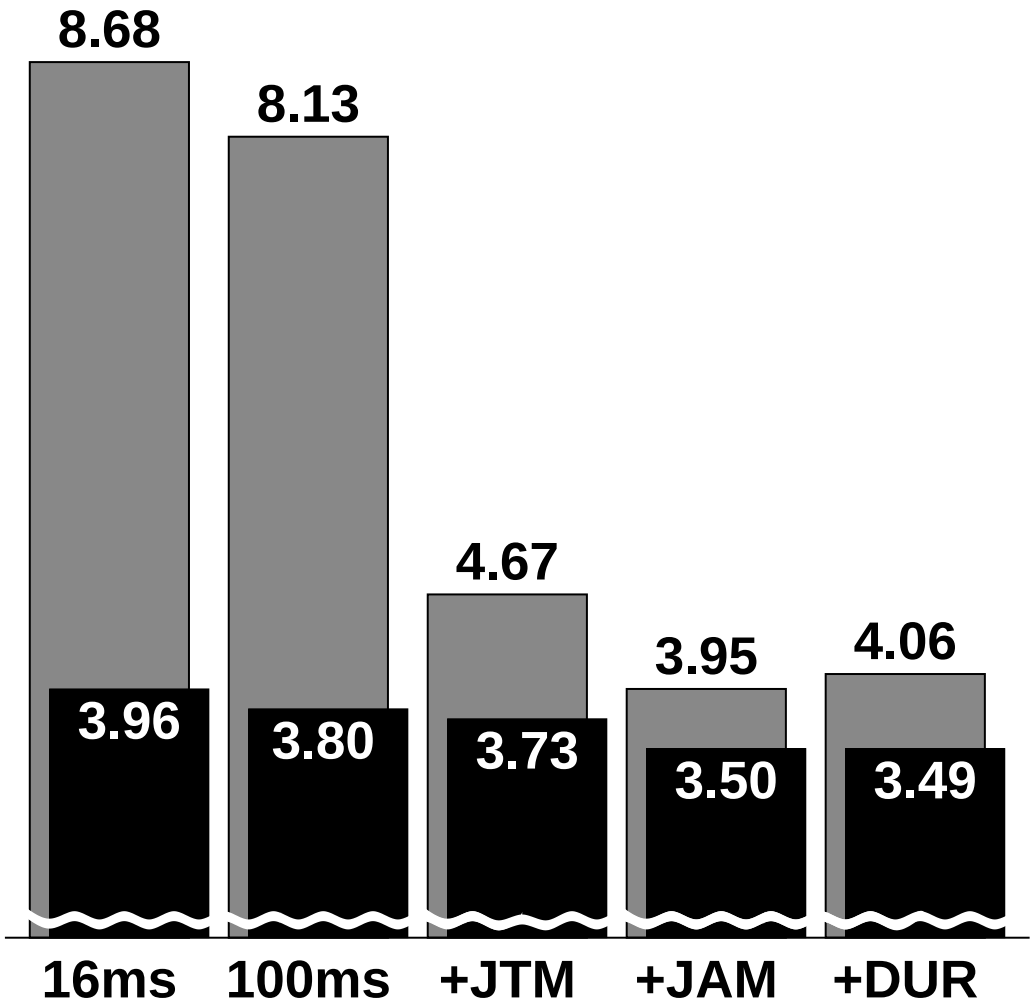
A solution to this problem is to:

1. **obtain models on test conversation**
2. high-precision first pass (Laskowski & Schultz, 2004; 2006)
   - **Non-Target-Normalization of Cross-Correlation Maxima**
   - compute cross-correlation maxima for all channel pairs
   - can infer relative geometry of all participants
3. train full-covariance log-energy model (from scratch)
4. interpolate with supervised single-participant models

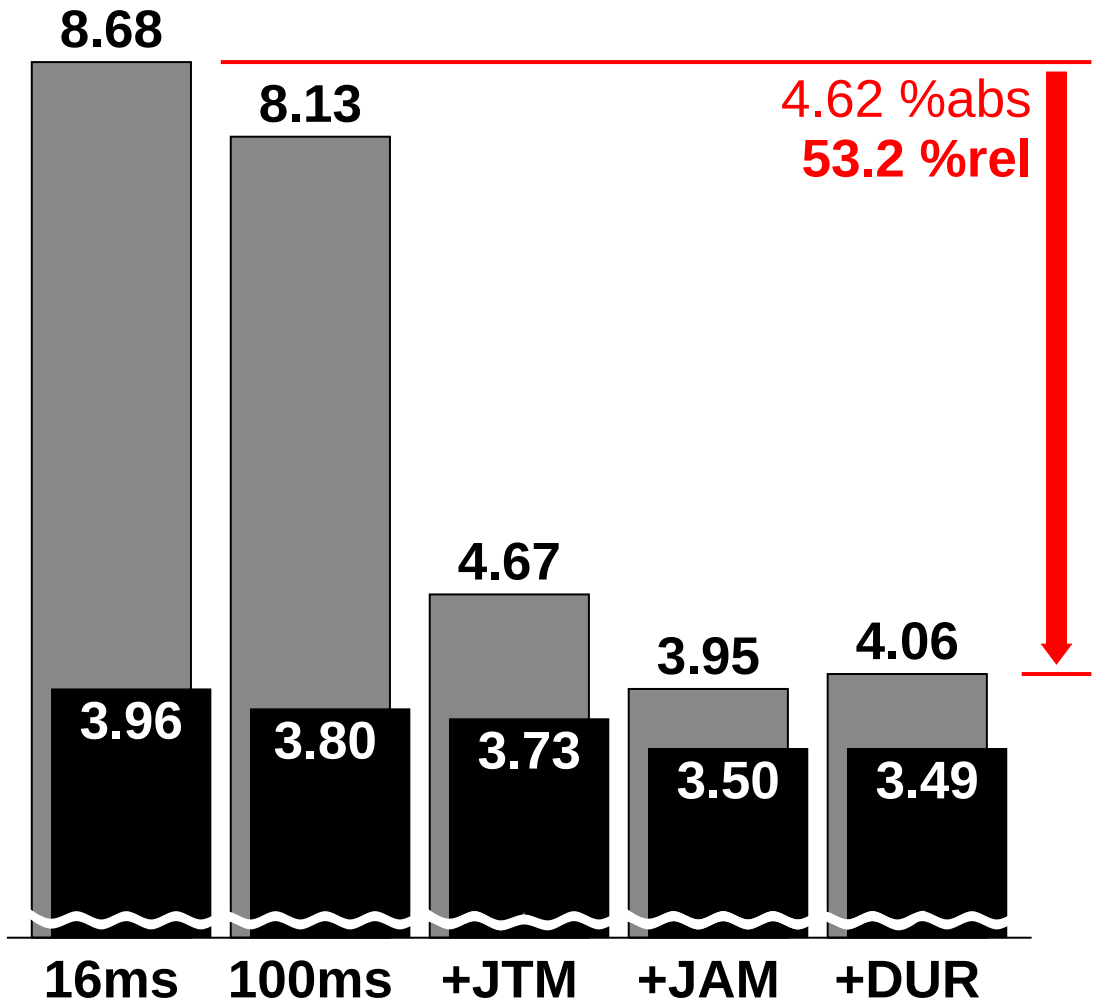
Classification error on EvalSet



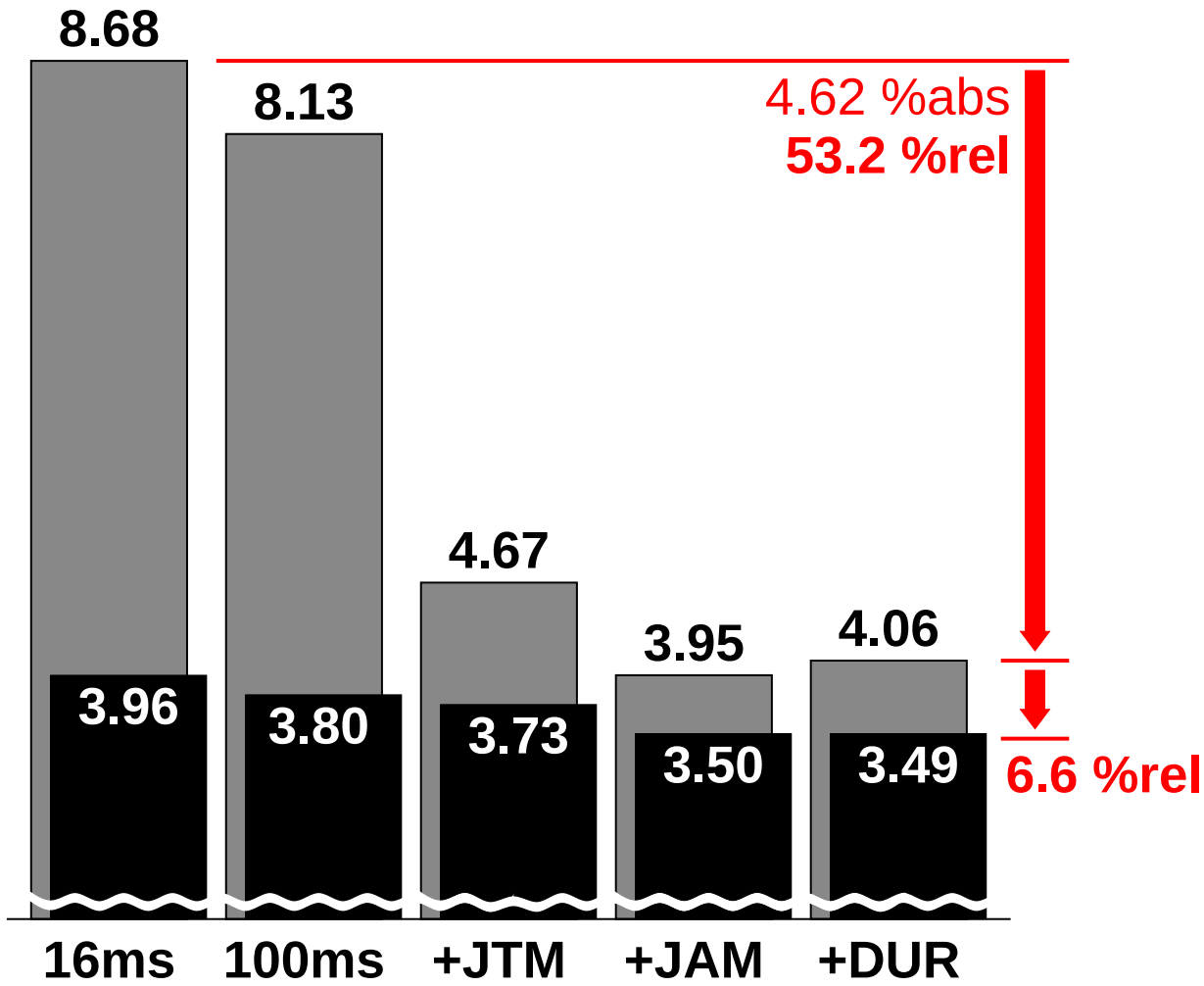
Classification error on EvalSet



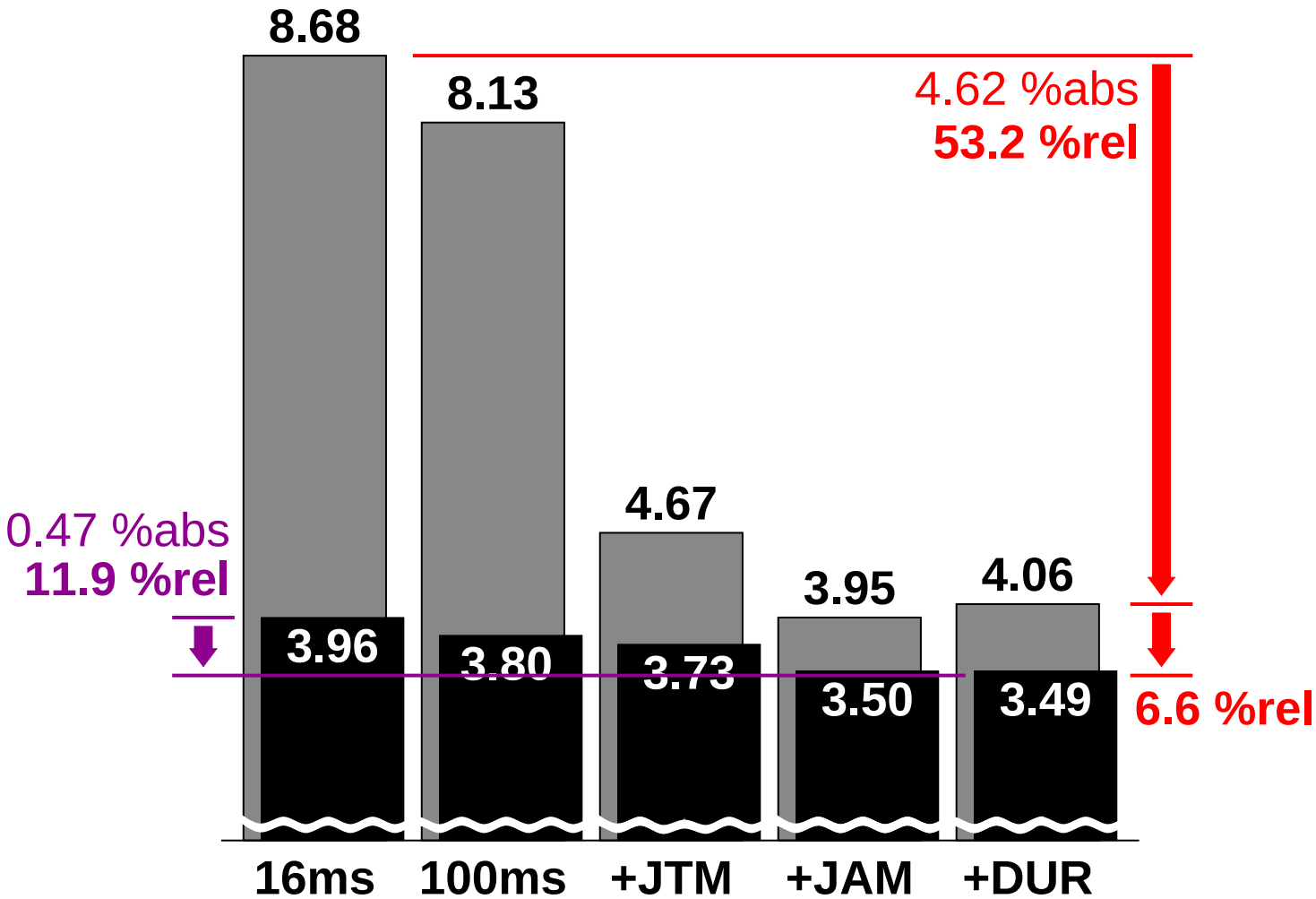
Classification error on EvalSet



Classification error on EvalSet



Classification error on EvalSet



Summary

1. Chronograms make it easy to model **joint behavior**.
2. This enables control over hypothesized **degree of overlap**.
   - participants **take turns** to talk
3. Limiting potential overlap reduces **impact of crosstalk**.
4. Error rates reduced by **40-70%** relative to standard baseline.


LAUGHTER
DETECTION

Laughter is Surprisingly Frequent

- what **else** are participants doing (than can be heard)?

- analysis in ICSI Meeting Corpus (67 hours of conversation)
- laughter is (Laskowski & Burger, 2007):
  - the most frequently transcribed non-verbal vocalization
  - >13,000 bouts of laughter in total
  - accounts for **9% of all vocal effort**
  - bouts containing some voicing: 66%
  - bouts containing no voicing: 34%

Laughter Detection Results, Briefly

- extend 2-class SAD topology to 3-class topology
- achieves F-**scores in the range 30-50%**
  - ERR = misses + false alarms of about 20-30%
  - higher than reported for the 4000 most audible voiced bouts
  - EERs $< 10\%$ (Truong & van Leeuwen, 2007; Knox et al, 2008)
- obtained $F = 47.7\%$ only available baseline for **all laughter**
- joint modeling improves F-scores only by $\approx 2\%$abs
  - and only for small topologies

Why Laughter Detection Poorer than Speech Detection

1. laughter **not** very confusable with speech
2. laughter most confusable **with silence**
   - laughter syllables contain long intervening pauses
   - also, unvoiced syllables sound just like breathing
3. highest F-scores achieved by **extending minimum duration constraints**
   - well beyond the most likely durations of laugh bouts
4. **large topologies prohibit joint participant decoding**
5. also: joint participant decoding of only limited viability
   - participants wait their turn to talk
   - participants **do not** wait their turn to laugh


DIALOG
ACT
RECOGNITION

Time-Dependent Modeling of Chronograms

- condition transition probabilities at instant t:
  - not only on **whether** participants are talking or not talking
  - but on **what they are trying to achieve** by talking
  - $\longrightarrow$ inference of **intent**
  - encoded in content-independent **dialog act (DA)** type
    - e.g., statements, questions, backchannels

- enables text-independent DA recognition (Laskowski & Shriberg, 2009; 2010)
- assign to each instant t, at which a participant is talking,
  - a **DA type**
  - optionally, a **DA boundary type**

- recognition $\equiv$ segmentation AND classification

Time-Dependent Modeling of Chronograms

- condition transition probabilities at instant t:
  - not only on **whether** participants are talking or not talking
  - but on **what they are trying to achieve** by talking
  - $\longrightarrow$ inference of **intent**
  - encoded in content-independent **dialog act (DA)** type
    - e.g., statements, questions, backchannels

- enables text-independent DA recognition (Laskowski & Shriberg, 2009; 2010)
- assign to each instant t, at which a participant is talking,
  - a **DA type**
  - optionally, a **DA boundary type**

- recognition $\equiv$ segmentation AND classification

Time-Dependent Modeling of Chronograms

- condition transition probabilities at instant t:
  - not only on **whether** participants are talking or not talking
  - but on **what they are trying to achieve** by talking
  - $\longrightarrow$ inference of **intent**
  - encoded in content-independent **dialog act (DA)** type
    - e.g., statements, questions, backchannels

- enables text-independent DA recognition (Laskowski & Shriberg, 2009; 2010)
- assign to each instant t, at which a participant is talking,
  - a **DA type**
  - optionally, a **DA boundary type**

- recognition $\equiv$ segmentation AND classification

Prior Research on Dialog Act Recognition

- lots of work in meetings, 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 had not been done

DA Types in ICSI Meetings

Propositional Content DA Types

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

"Short" DA Types

Feedback Types (5.4%)

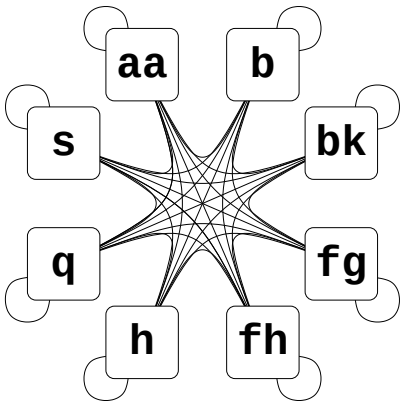
- **backchannel**, `b` (2.8%)
- **acknowledgment**, `bk` (1.5%)
- **assert**, `aa` (1.1%)

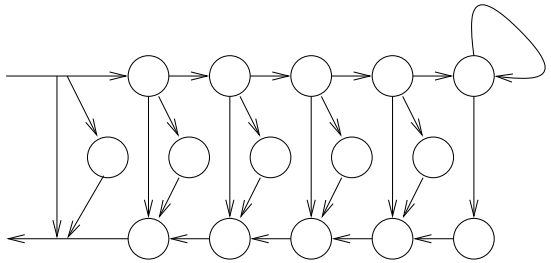
Floor Mechanism Types (3.6%)

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

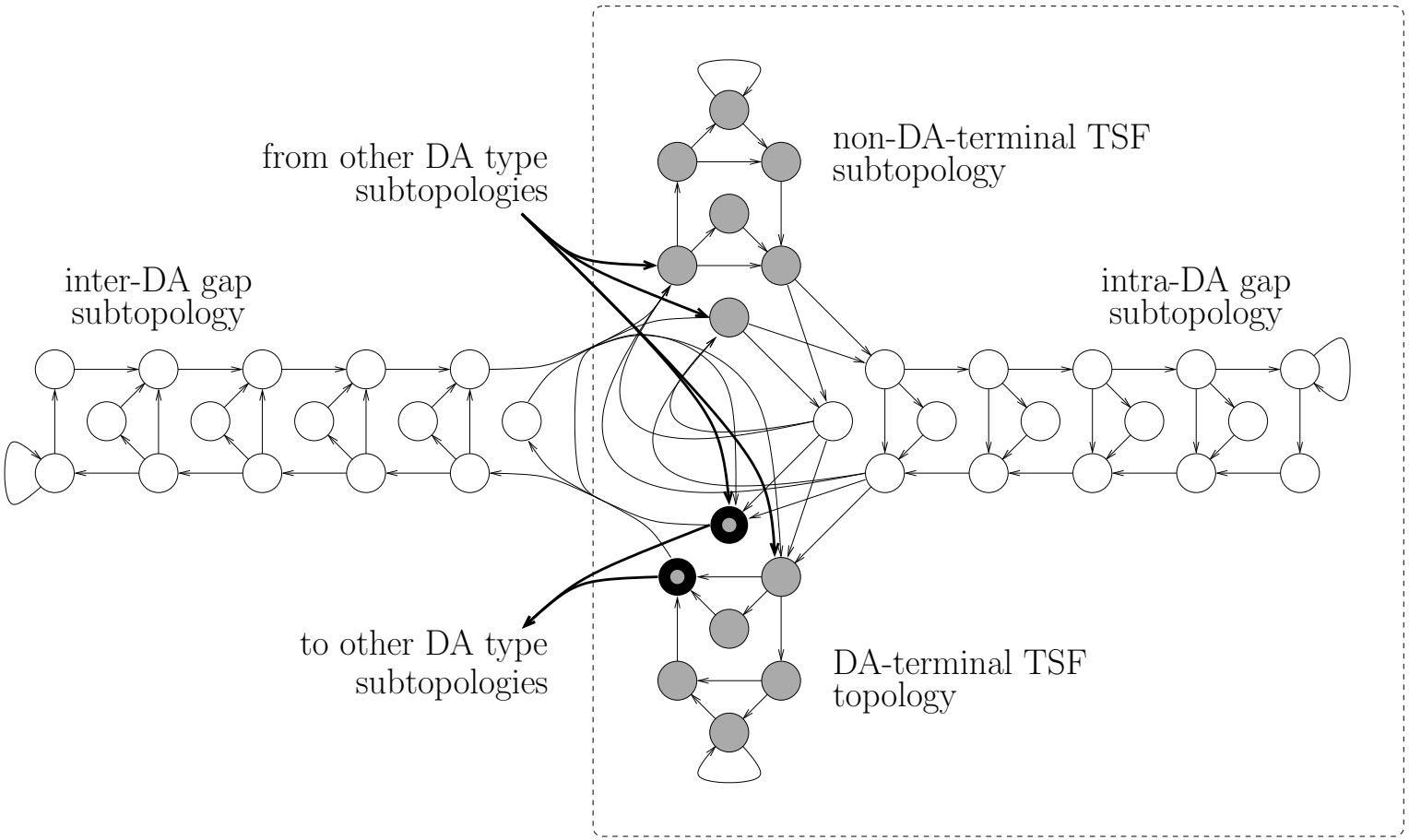
The Single-Participant DA State Space

- one DA-specific sub-topology for each of 8 DA types
- fully connected via silence sub-topologies



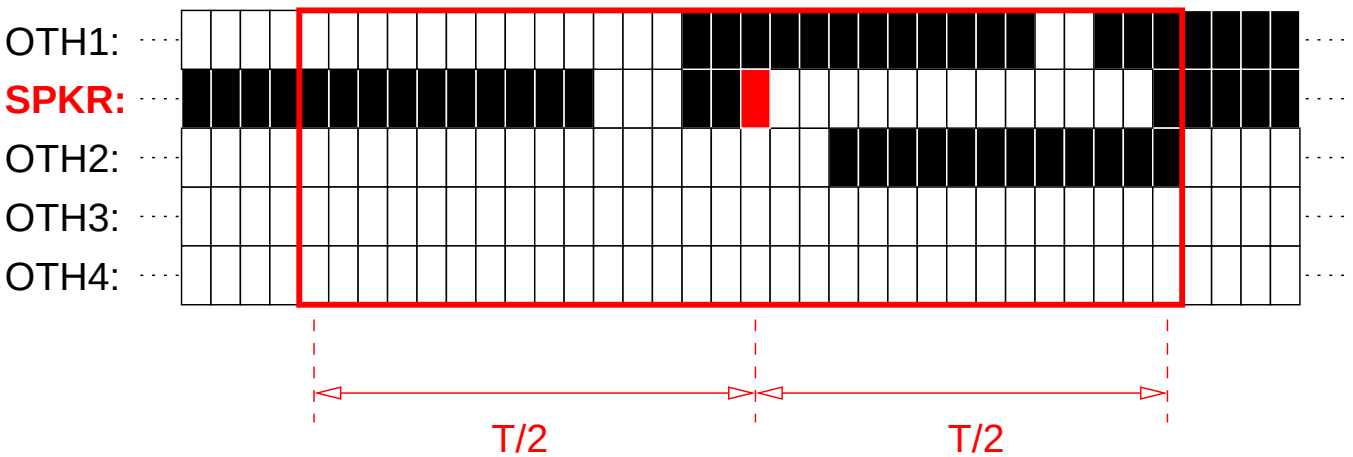


A DA Sub-Topology



Time-Dependent Modeling of Chronograms

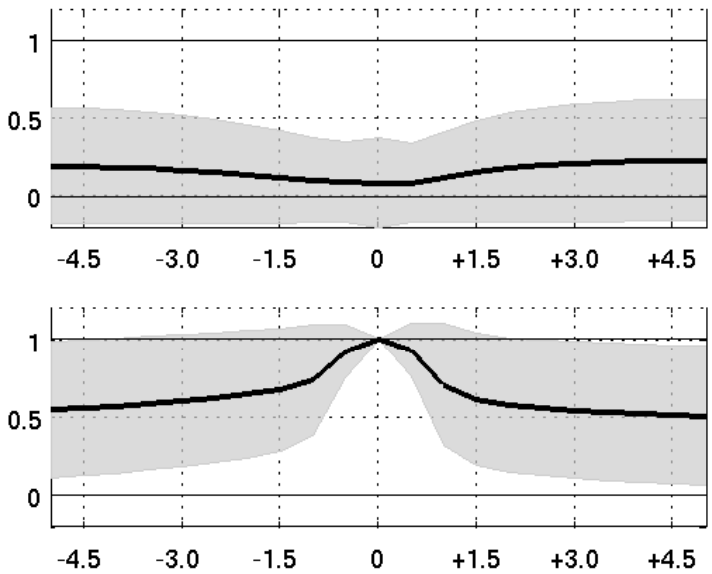
- single-participant state space consists of hundreds of states
- therefore, model participant transitions independently
- but capture a **local chronogram snapshop** as an emission



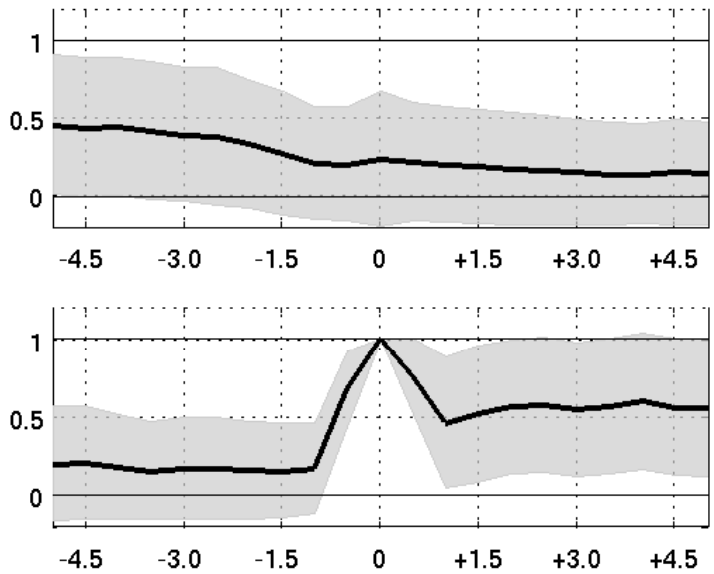
- K-independence: retain only 3 most talkative interlocutors
- rotation invariance: rotate interlocutors by talkativity rank

The Probability of Speaking Near DA Types

- upper panel: most talkative interlocutor
- lower panel: target participant producing the DA

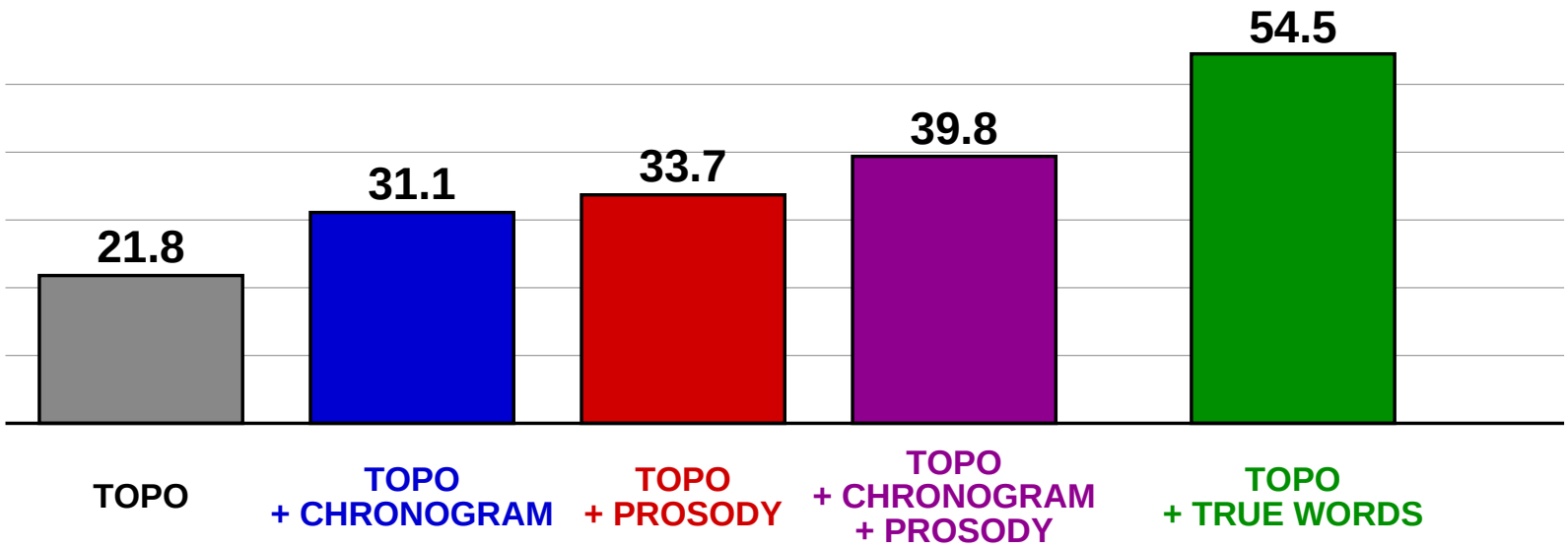


statement

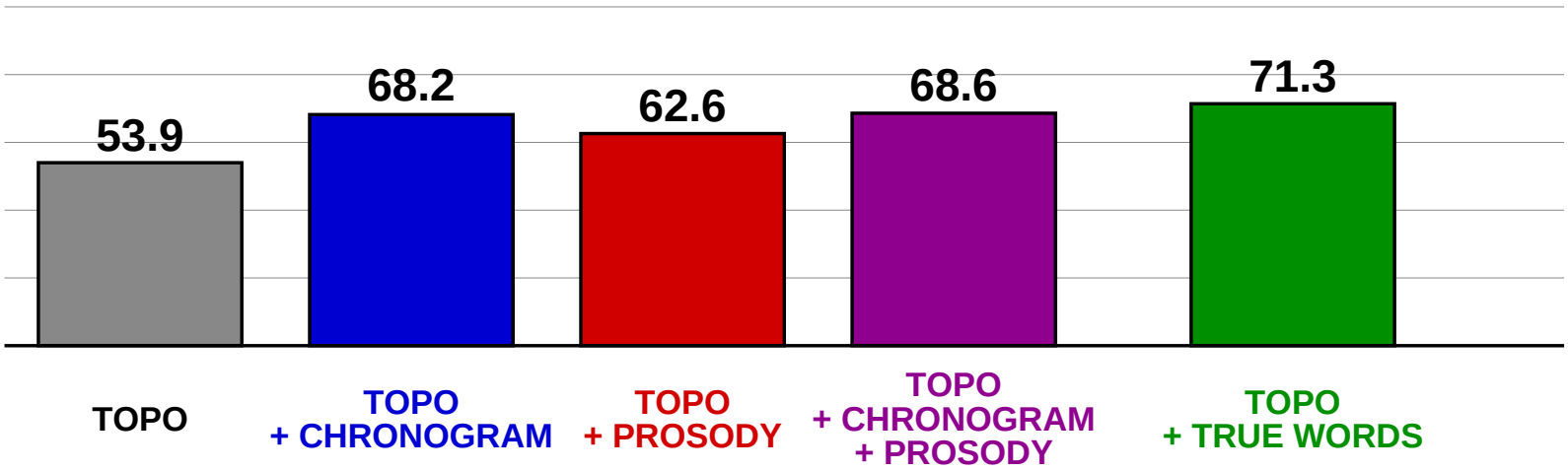


floor grabber

Average 8-class *F*-scores, EvalSet



Punctuation *F*-scores, EvalSet



Summary

1. **local snapshots of speech chronograms** correlate with production of specific dialog act types
2. correlation sufficiently strong to form the basis of a **text-independent DA recognizer**
3. local snapshots of speech chronograms complementary with prosodic features
4. for several dialog act types and dialog act boundary types, performance approaches that using models of manually transcribed word sequences

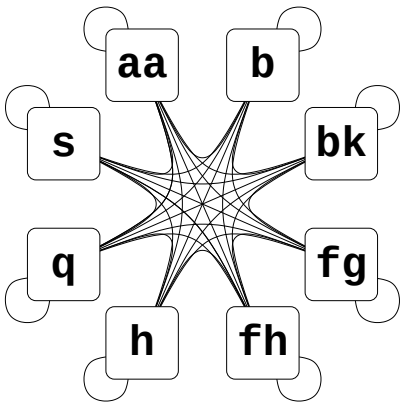

HUMOR
DETECTION

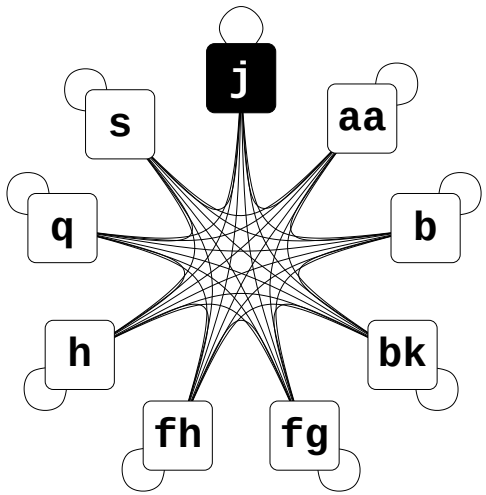
Why Care About Humor?

- talk produced not only to communicate facts or control floor
- also to **regulate socio-emotional state of interlocutors**
- humor **qualifies seriousness** of propositional content
- only prior research in meetings (Clark & Popescu-Belis, 2004) indicated detectability **not above chance**

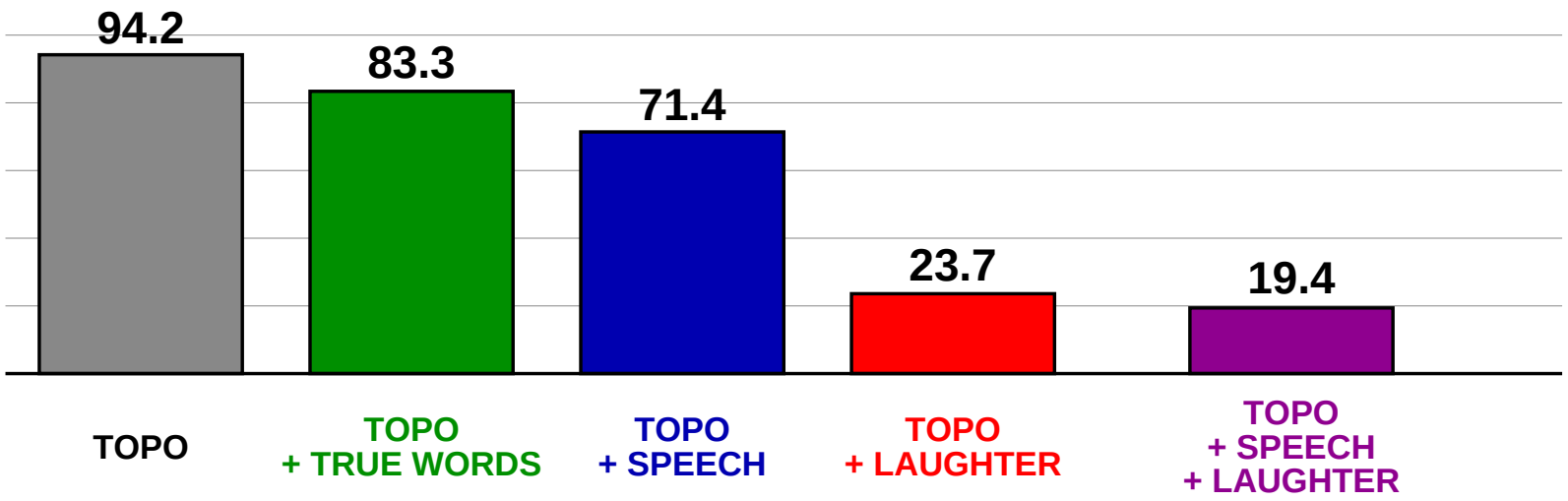
Humor Annotation in ICSI Meetings

- both statements (s) and questions (q) license the optional j
  - *attempts to amuse* or *attempts at sarcasm*
  - accounts for 0.6% of speech by time
  - break j out as **a 9th DA type**
  - then run DA recognition, as shown earlier
  - score only detection of j



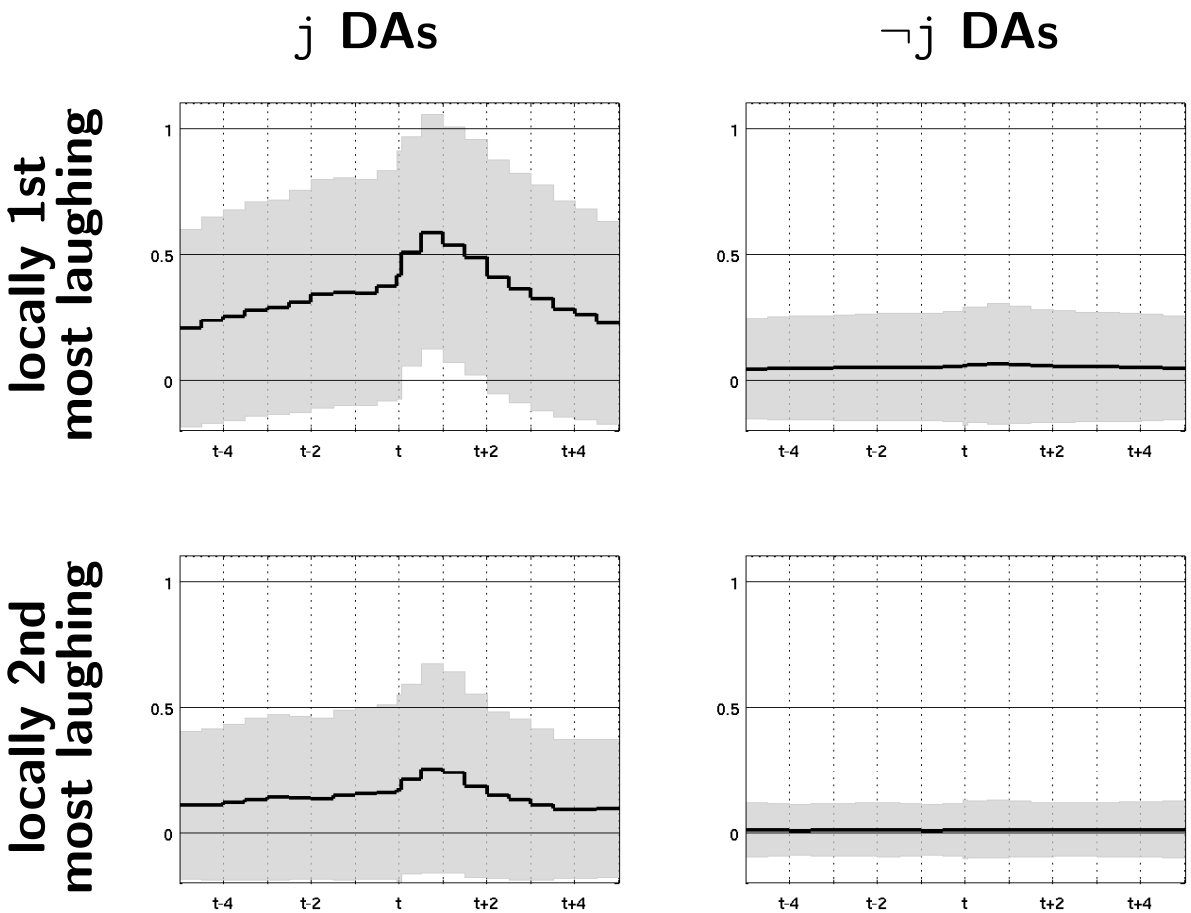


Humor Detection Error Rates, EvalSet

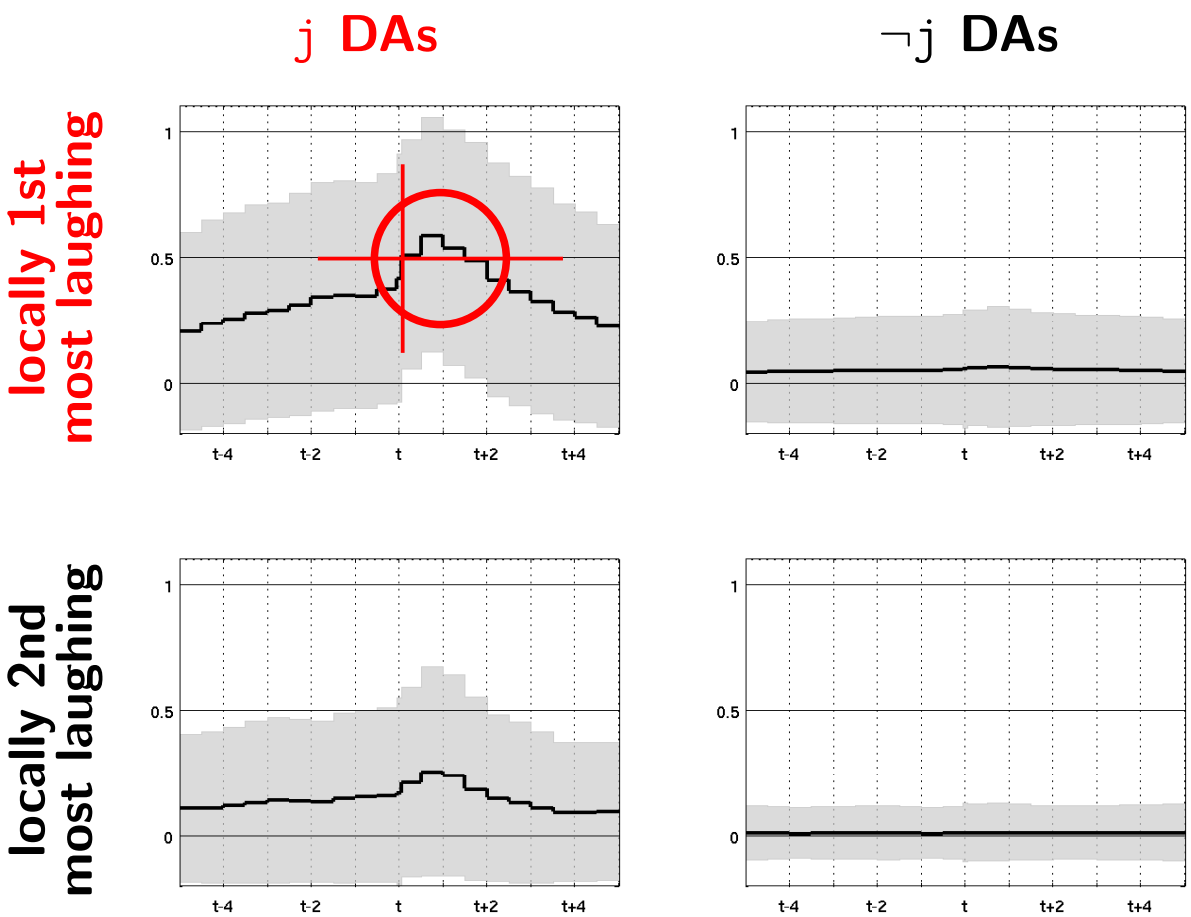


- laughter chronograms = **best single source** of information for detecting humor
- combination with speech chronograms leads to improvement
- combination with lexical system leads to no improvement

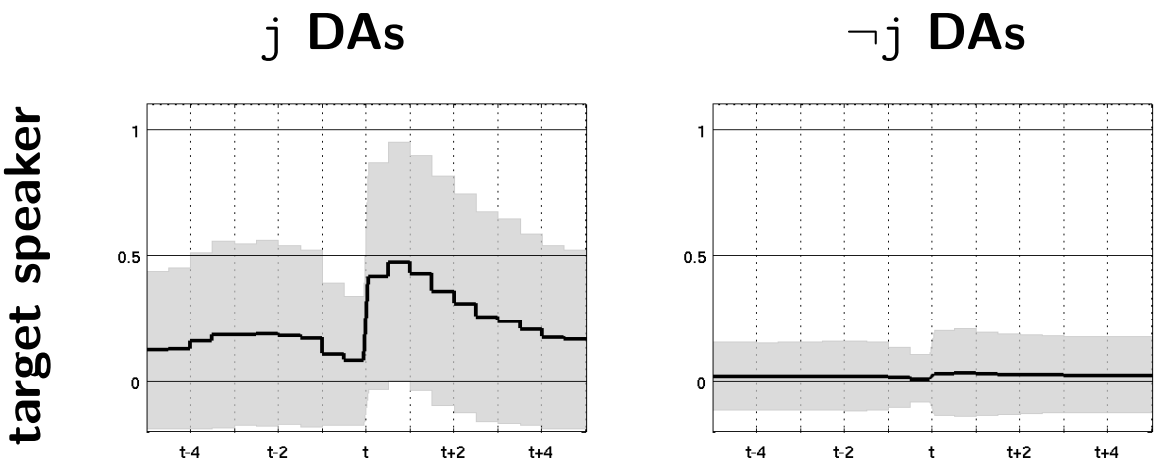
Interlocutor Probability of Laughing



Interlocutor Probability of Laughing

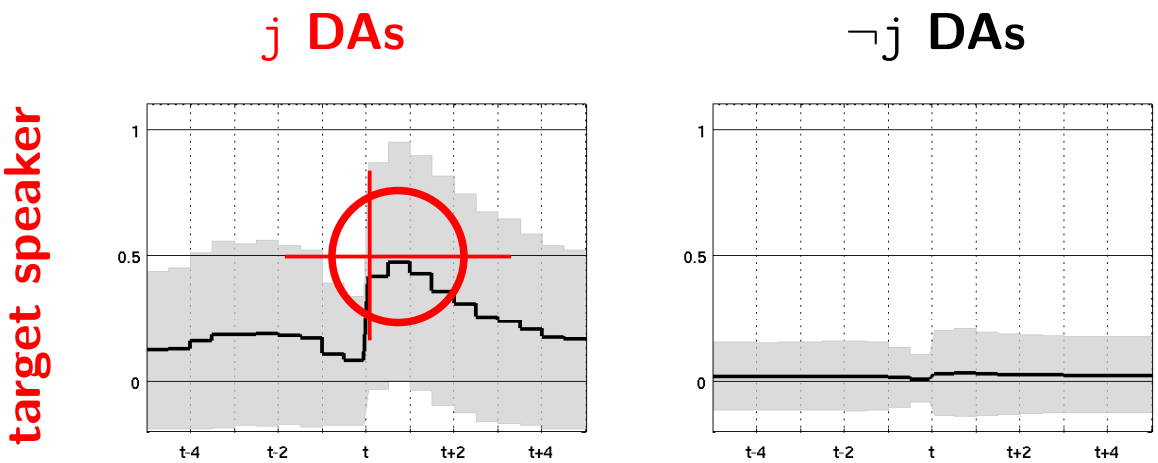


Target Speaker Probability of Laughing



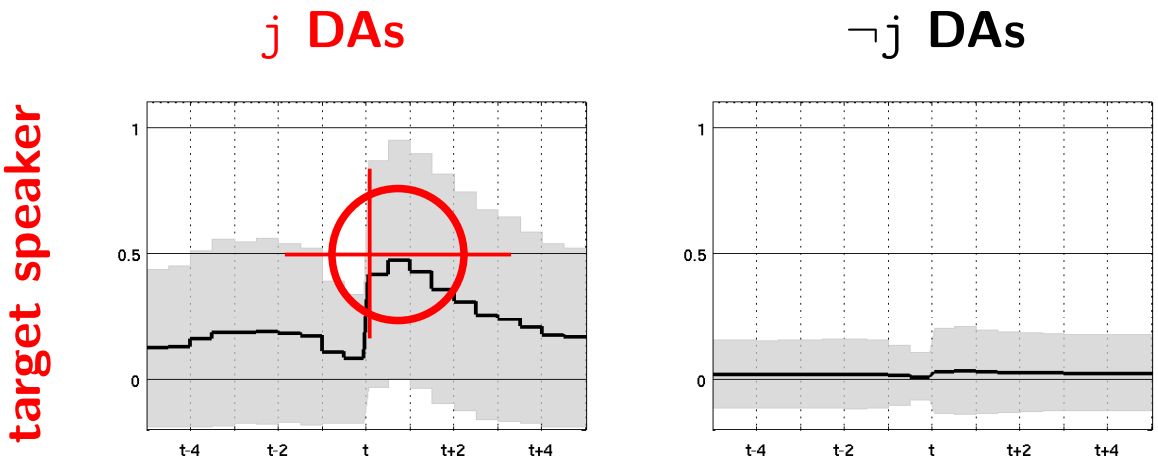
- How well do we do with laughter only from the target speaker?
- ERR = 31% rather than 23.7%

Target Speaker Probability of Laughing



- How well do we do with laughter only from the target speaker?
- ERR = 31% rather than 23.7%

Target Speaker Probability of Laughing

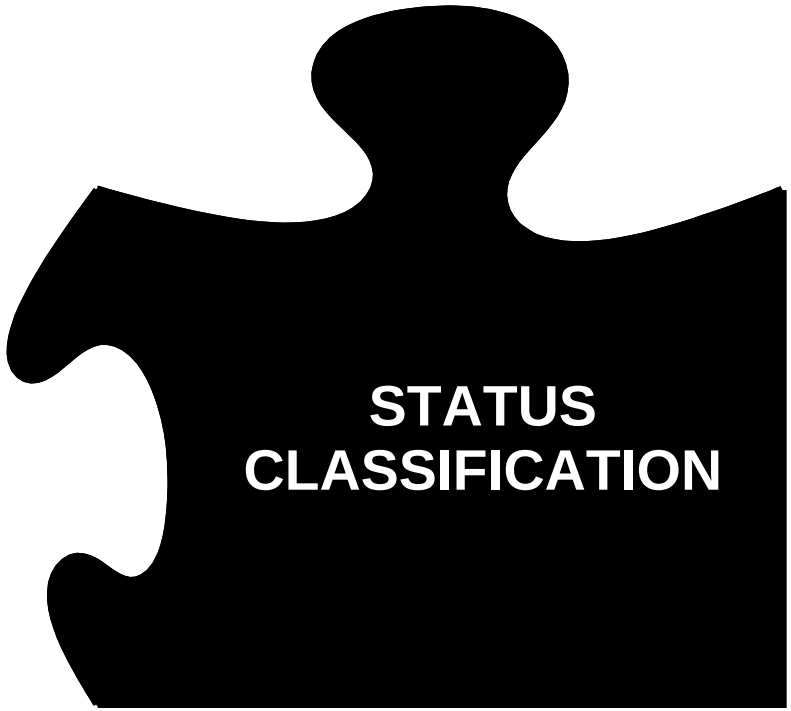


- How well do we do with laughter only from the target speaker?
- ERR = 31% rather than 23.7%

Summary

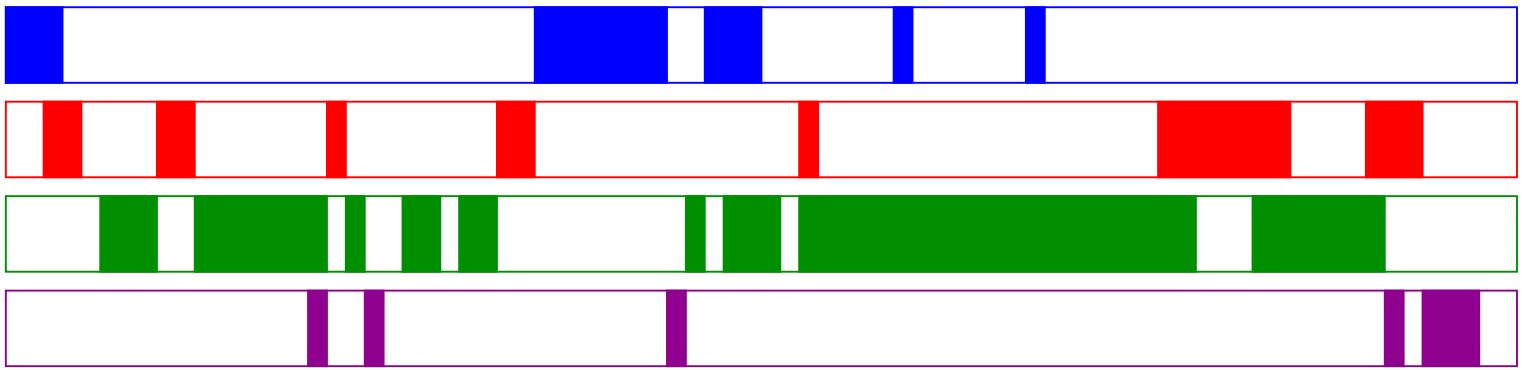
- speech chronograms play an important role in text-independent DA recognition
- text-independent: **without using words**
- system approaches performance achievable of a text-dependent system

- laughter chronograms play a crucial role in detection of attempts to amuse
  - in either text-independent or text-dependent systems
- jokers, in work-place conversations, appear to signal that they have joked by laughing themselves


STATUS
CLASSIFICATION

What can be said of individuals?

- observing only the vocal interaction chronogram (Laskowski, Ostendorf & Schultz, 2007; 2008)



Prior Research

- static characterization of meeting participants
  - dominance rankings: Rienks & Heylen, 2005
  - influence rankings: Rienks *et al.*, 2006
- static characterization of radio talk show participants
  - roles: Vinciarelli, 2007
- dynamic characterization of meeting participants
  - roles: Banerjee & Rudnicky, 2004
  - roles: Zancanaro *et al.*, 2006
  - roles: Rienks *et al.*, 2006
- lots of work in social psychology, for dialogue
  - human resource allocation
  - diagnosis of psychological disorders

Prior Research

- static characterization of meeting participants
  - dominance rankings: Rienks & Heylen, 2005
  - influence rankings: Rienks *et al.*, 2006
- static characterization of radio talk show participants
  - roles: Vinciarelli, 2007
- dynamic characterization of meeting participants
  - roles: Banerjee & Rudnicky, 2004
  - roles: Zancanaro *et al.*, 2006
  - roles: Rienks *et al.*, 2006
- lots of work in social psychology, for dialogue
  - human resource allocation
  - diagnosis of psychological disorders

Prior Research

- static characterization of meeting participants
  - dominance rankings: Rienks & Heylen, 2005
  - influence rankings: Rienks *et al.*, 2006
- static characterization of radio talk show participants
  - roles: Vinciarelli, 2007
- dynamic characterization of meeting participants
  - roles: Banerjee & Rudnicky, 2004
  - roles: Zancanaro *et al.*, 2006
  - roles: Rienks *et al.*, 2006
- lots of work in social psychology, for dialogue
  - human resource allocation
  - diagnosis of psychological disorders

Prior Research

- static characterization of meeting participants
  - dominance rankings: Rienks & Heylen, 2005
  - influence rankings: Rienks *et al.*, 2006
- static characterization of radio talk show participants
  - roles: Vinciarelli, 2007
- dynamic characterization of meeting participants
  - roles: Banerjee & Rudnicky, 2004
  - roles: Zancanaro *et al.*, 2006
  - roles: Rienks *et al.*, 2006
- lots of work in social psychology, for dialogue
  - human resource allocation
  - diagnosis of psychological disorders

Modeling Individual Participation

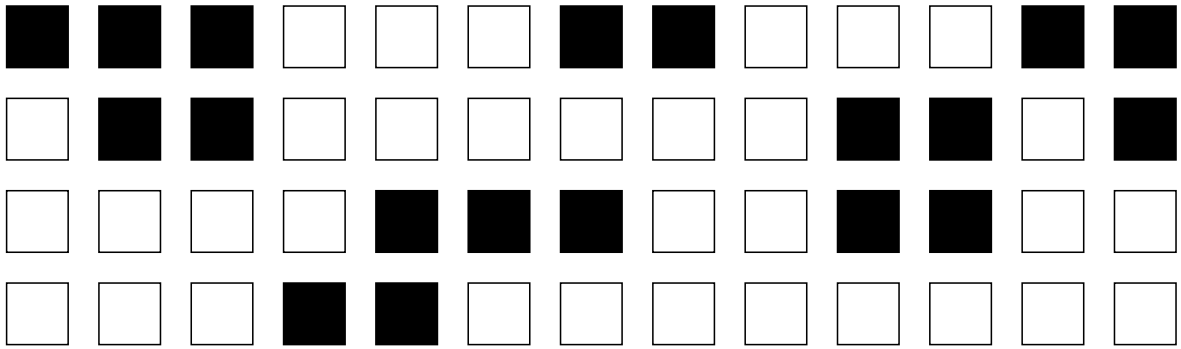
1. assume participant behavior to be conditionally independent, given prior joint participant behavior

$$P\left(\mathbf{q}_t|\mathbf{q}_{t-1}\right) = \prod_{k=1}^{K} P\left(\mathbf{q}_t\left[k\right]|\mathbf{q}_{t-1}\right)$$

2. infer model for each participant, given test conversation
3. extract specific probabilities as features
4. model using independent Gaussian emission probabilities

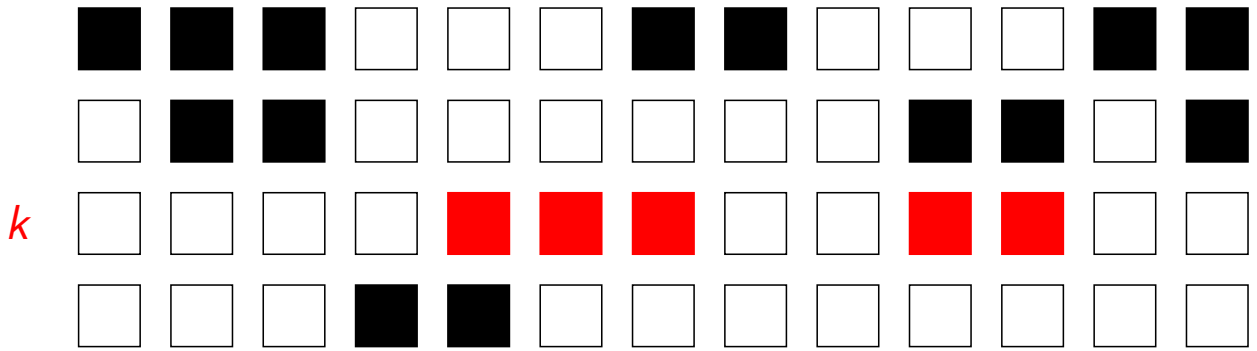
Features **F** Describing Participant Classes

1. probability of vocalizing (V)
2. probability of initiating vocalization (VI) in prior silence
3. probability of continuing vocalization (VC) in prior non-overlap
4. probability of initiating overlap (OI) in prior non-overlap
5. probability of continuing overlap (OC) in prior overlap



Features **F** Describing Participant Classes

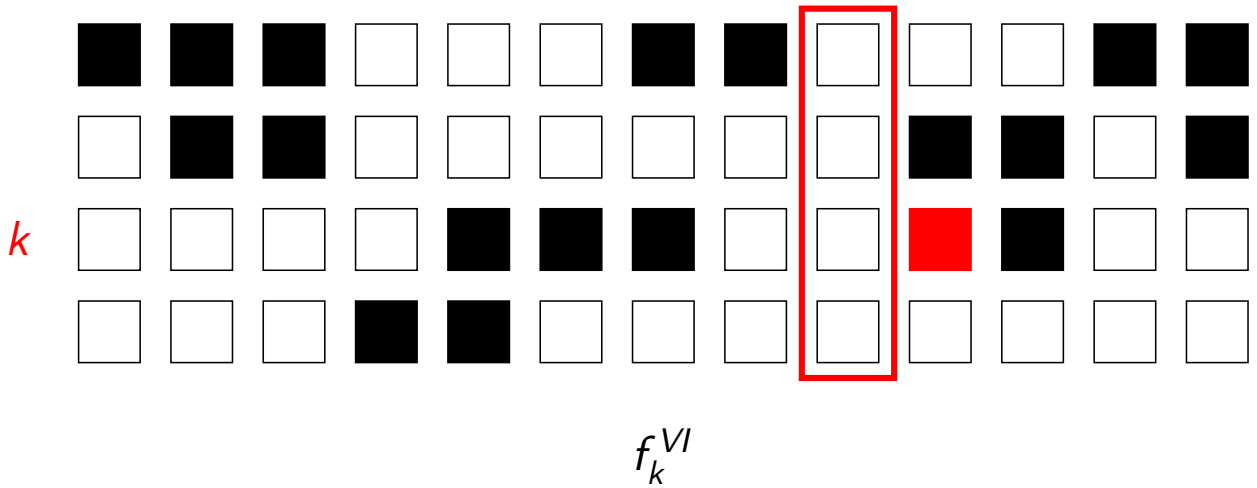
1. probability of vocalizing (V)
2. probability of initiating vocalization (VI) in prior silence
3. probability of continuing vocalization (VC) in prior non-overlap
4. probability of initiating overlap (OI) in prior non-overlap
5. probability of continuing overlap (OC) in prior overlap



f_k^V

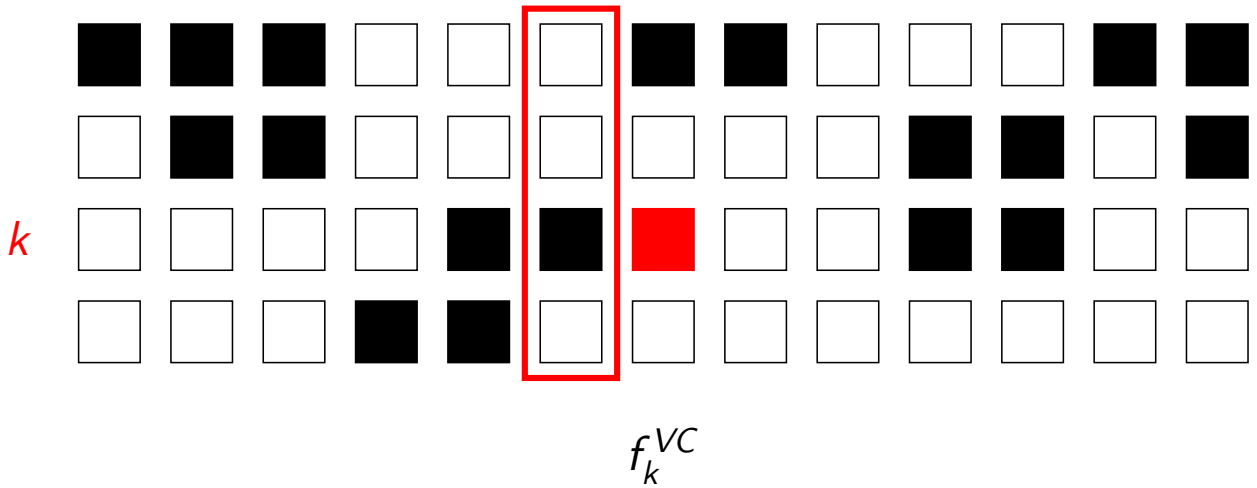
Features **F** Describing Participant Classes

1. probability of vocalizing (V)
2. probability of initiating vocalization (VI) in prior silence
3. probability of continuing vocalization (VC) in prior non-overlap
4. probability of initiating overlap (OI) in prior non-overlap
5. probability of continuing overlap (OC) in prior overlap



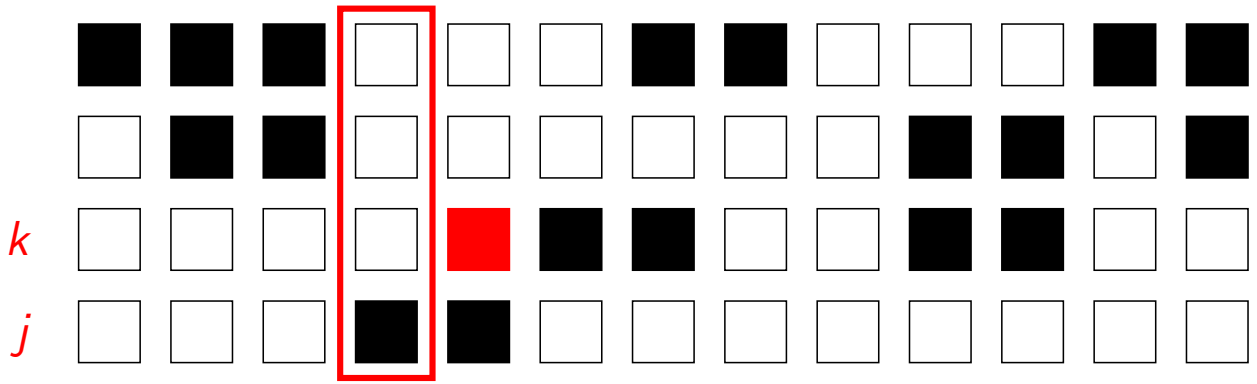
Features **F** Describing Participant Classes

1. probability of vocalizing (V)
2. probability of initiating vocalization (VI) in prior silence
3. probability of continuing vocalization (VC) in prior non-overlap
4. probability of initiating overlap (OI) in prior non-overlap
5. probability of continuing overlap (OC) in prior overlap



Features **F** Describing Participant Classes

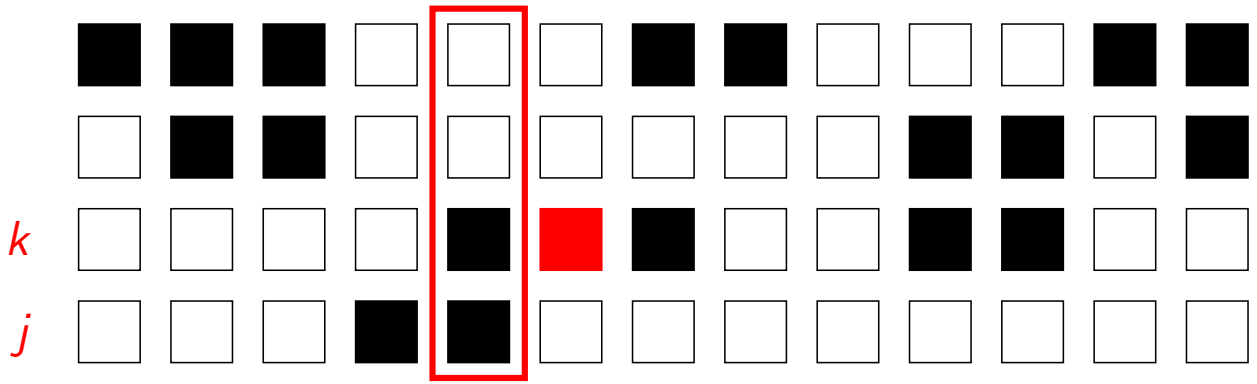
1. probability of vocalizing (V)
2. probability of initiating vocalization (VI) in prior silence
3. probability of continuing vocalization (VC) in prior non-overlap
4. probability of initiating overlap (OI) in prior non-overlap
5. probability of continuing overlap (OC) in prior overlap



$f^{OI}_{k,j}$

Features **F** Describing Participant Classes

1. probability of vocalizing (V)
2. probability of initiating vocalization (VI) in prior silence
3. probability of continuing vocalization (VC) in prior non-overlap
4. probability of initiating overlap (OI) in prior non-overlap
5. probability of continuing overlap (OC) in prior overlap



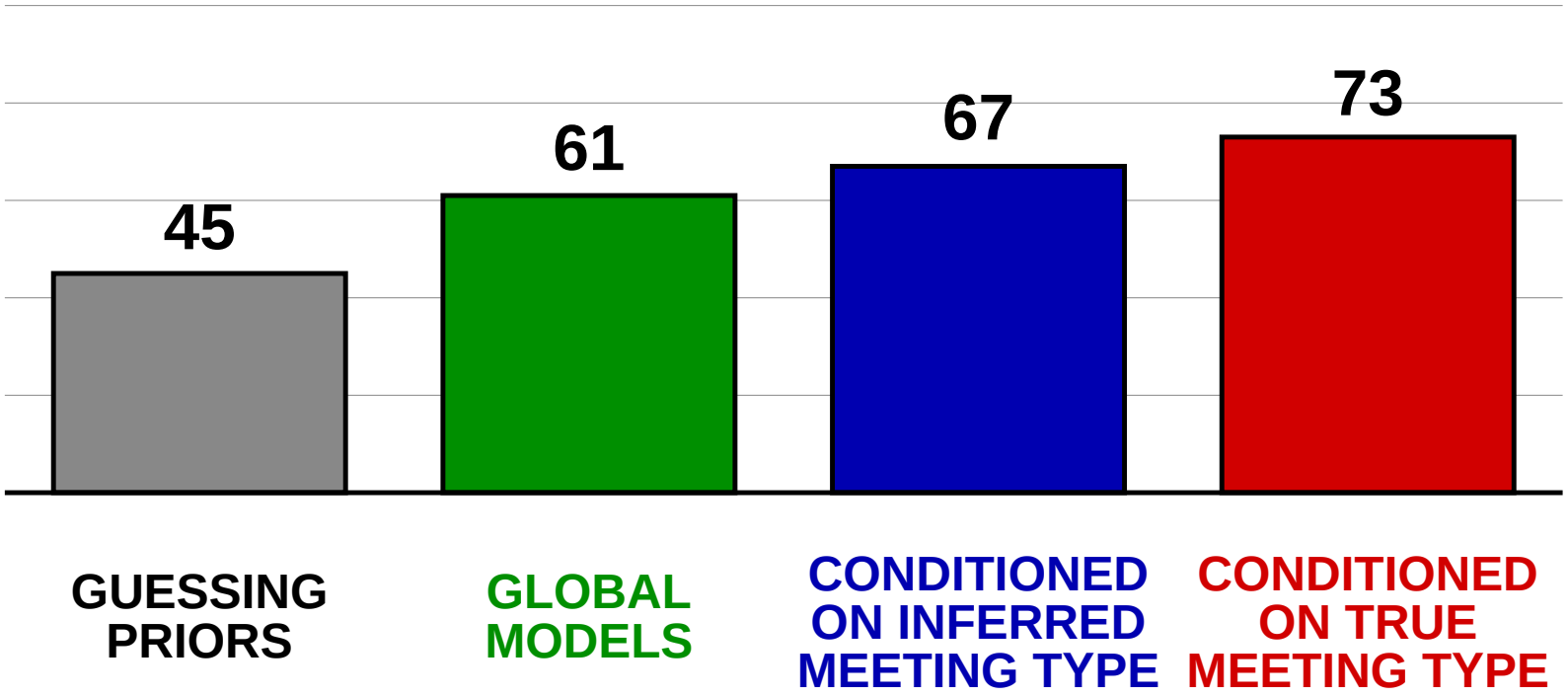
$f^{OC}_{k,j}$

What Participant Classes Can We Identify?

- ICSI Meeting Corpus
- **naturally occurring** meetings
- participants self-reported as one of three of:
  - professor (PROF)
  - possessing PhD (PHD)
  - graduate students (STUD)

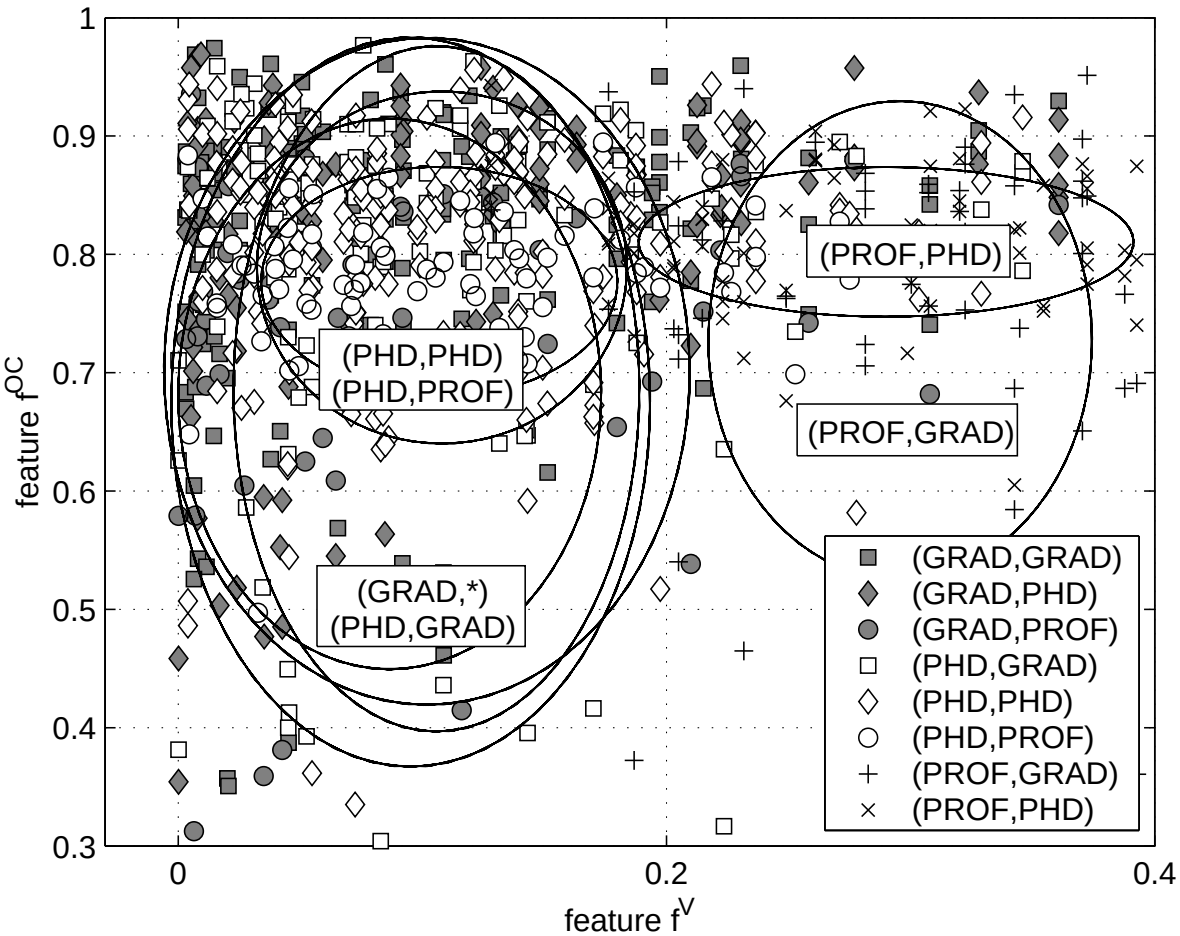
  $\longrightarrow$ **organizational seniority**
- 67 meetings of one of three types:
  - professor-student discussions (Bed)
  - annotation discussions (Bmr)
  - research discussions (Bro)
- presumably, people behave **differently in different settings**

Seniority Classification Accuracy



- 1st-best feature type: continuation of overlap
- 2nd-best feature type: initiation of overlap
- 3rd-best feature type: total speaking time proportion

Seniority Level Feature Distributions




ROLE
CLASSIFICATION

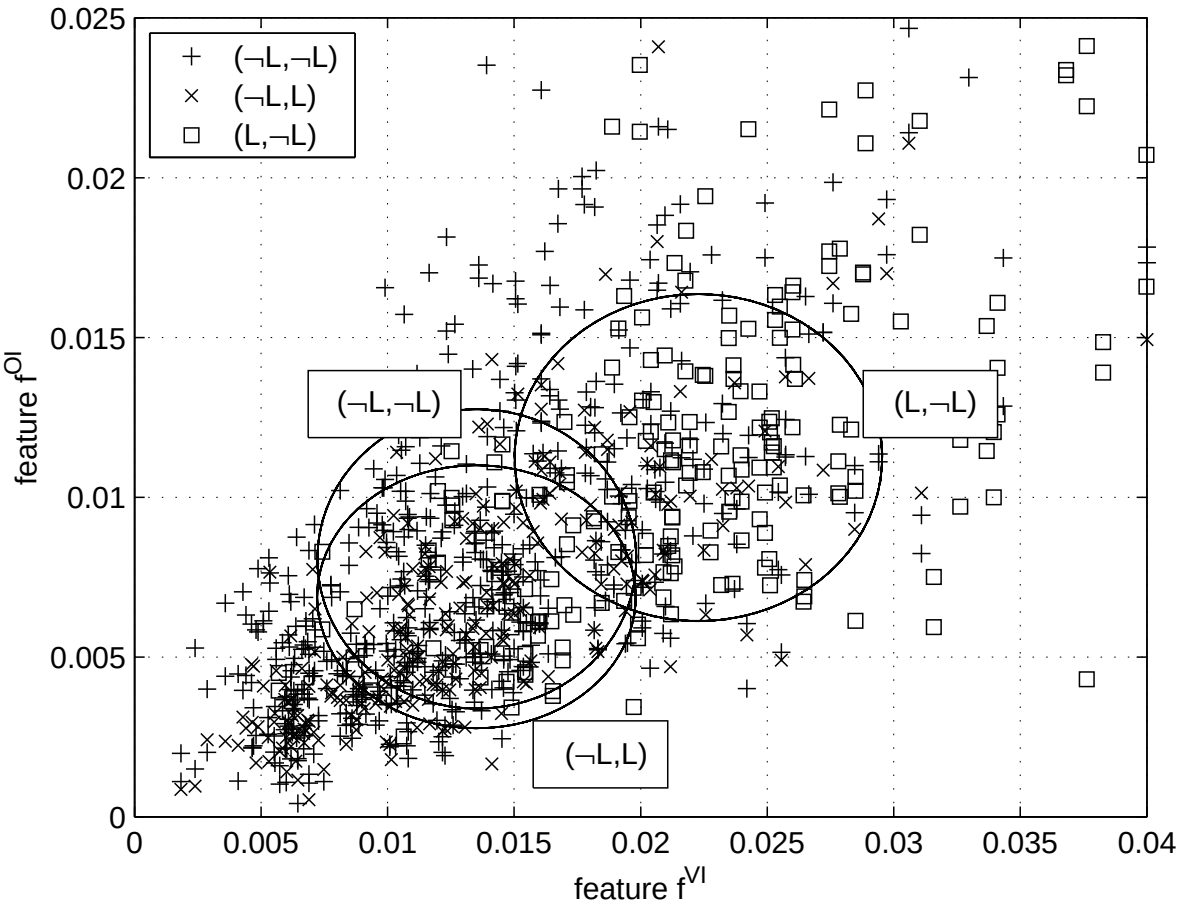
Assigned Role in Meetings

- Can we detect a role **given** to a participant, independent of their diffuse status characteristics?

- AMI Meeting Corpus
- always $K = 4$ participants
- always 4 roles
  - project manager (PM)
  - marketing expert (ME)
  - user interface designer (UI)
  - industrial designed (ID)

- classification paradigm identical to seniority classification,
- except that roles are mutually exclusive
- classification: 53% accuracy (guessing priors: 25%)
- detection of PM: 75% accuracy

Feature Distributions for Finding Project Managers



Participant Characterization Summary

- aspects of chronogram patterns correlated with characterizations of individual participants
  - diffuse characteristics, e.g. seniority
  - (temporarily) assigned roles
- correlation sufficiently strong to allow for inference of participant type
- first baselines for both text-independent tasks

Conclusions

1. the chronogram is a **deceptively sparse** representation
2. appears to contain **very rich information**
   - particularly that information which is not explicitly stated
3. it makes it easy to consider participants' **joint behavior**
   - vocal behavior readily synchronizable across participants

4. chronograms are amenable to various modeling alternatives, leading to successful inference
   1. time- and participant-indendent: **detection of vocal activity**
   2. time-dependent: **recognition of intent**
   3. participant-dependent: **characterization of participants**

Contributions

1. explicit framework and techniques for modeling chronograms
   - in a variety of ways, depending on application
   - shown to corroborate many findings in the social sciences
2. a **text-independent conversation understanding system**
   - allowing inference of many aspects of conversation
   - **without ever needing to recognize a word**
3. first-ever text-independent baselines for several tasks
   - detection of all laughter
   - segmentation and classification of dialog acts
   - detection of attempts to amuse
   - classification of (tacit) participant seniority
   - classification of assigned participant role

Potential Future Impact

1. it is now possible to automatically compare conversations
   - across genres
   - across cultures
   - across languages
2. it is now possible to perform large-scale, automated validation of the qualitative findings of
   - conversation analysis
   - social psychology
   - anthropology
   - and others ...
3. merging **conversational content** with **conversation form** is promising
4. many of the presented systems are amenable to immediate improvement

THANK YOU.

Special thanks to: Anton Batliner, Alan Black, Susi Burger, Jaime Carbonell, Jens Edlund, Christian Fügen, Mattias Heldner, Qin Jin, Rob Malkin, Florian Metze, Mari Ostendorf, Matthias Paulik, Tanja Schultz, Liz Shriberg, Richard Stern, Ashish Venugopal, Stephan Vogel, Alex Waibel & Mattias Wölfel.