# Modeling Vocal Interaction for Text-Independent Classification of Conversation Type

Kornel Laskowski[1,3], Mari Ostendorf[1,2] & Tanja Schultz[1,3]

[1]interACT, Universität Karlsruhe
[2]Dept. Electrical Engineering, University of Washington
[3]interACT, Carnegie Mellon University

September 2, 2007

# What Is Vocal Interaction?

- the patterns of vocal activity for all participants to a conversation
  - no words $\longrightarrow$ a **text-independent** representation of multi-party conversation
- as used in psycholinguistics (Dabbs & Ruback, 1987)
- in telecommunications: "on-off patterns" (Brady, 1967)
- studied since the 1930s

# What Is Vocal Interaction?

- the patterns of vocal activity for all participants to a conversation
  - no words $\longrightarrow$ a **text-independent** representation of multi-party conversation
- as used in psycholinguistics (Dabbs & Ruback, 1987)
- in telecommunications: "on-off patterns" (Brady, 1967)
- studied since the 1930s

# What Is Vocal Interaction?

- the patterns of vocal activity for all participants to a conversation
  - no words $\longrightarrow$ a **text-independent** representation of multi-party conversation
- as used in psycholinguistics (Dabbs & Ruback, 1987)
- in telecommunications: "on-off patterns" (Brady, 1967)
- studied since the 1930s

# What Is Vocal Interaction?

- the patterns of vocal activity for all participants to a conversation
  - no words $\longrightarrow$ a **text-independent** representation of multi-party conversation
- as used in psycholinguistics (Dabbs & Ruback, 1987)
- in telecommunications: "on-off patterns" (Brady, 1967)
- studied since the 1930s

# What Is Vocal Interaction?

- the patterns of vocal activity for all participants to a conversation
  - no words $\longrightarrow$ a **text-independent** representation of multi-party conversation
- as used in psycholinguistics (Dabbs & Ruback, 1987)
- in telecommunications: "on-off patterns" (Brady, 1967)
- studied since the 1930s

# Why Do Classification of Conversation Type?

- a basic competence in conversation understanding
- type is most often taken for granted
  - ie. *"My project is about cocktail parties. Why would I ever need to know that a cocktail party is not a business meeting?"*
- searching & indexing in heterogenous multi-party conversation recordings (or portions)
- **text-independence**: pre-ASR availability of type hypothesis/prior

# Why Do Classification of Conversation Type?

- a basic competence in conversation understanding
- type is most often taken for granted
  - ie. *"My project is about cocktail parties. Why would I ever need to know that a cocktail party is not a business meeting?"*
- searching & indexing in heterogenous multi-party conversation recordings (or portions)
- **text-independence**: pre-ASR availability of type hypothesis/prior

# Why Do Classification of Conversation Type?

- a basic competence in conversation understanding
- type is most often taken for granted
  - ie. *"My project is about cocktail parties. Why would I ever need to know that a cocktail party is not a business meeting?"*
- searching & indexing in heterogenous multi-party conversation recordings (or portions)
- **text-independence**: pre-ASR availability of type hypothesis/prior
  - may contribute to optimal selection of ASR components
  - type classification possible where no ASR is available or processing possible

# Why Do Classification of Conversation Type?

- a basic competence in conversation understanding
- type is most often taken for granted
  - ie. *"My project is about cocktail parties. Why would I ever need to know that a cocktail party is not a business meeting?"*
- searching & indexing in heterogenous multi-party conversation recordings (or portions)
- **text-independence**: pre-ASR availability of type hypothesis/prior
  - may contribute to optimal selection of ASR components
  - type classification possible where no ASR is used (e.g. processing privacy)

# Why Do Classification of Conversation Type?

- a basic competence in conversation understanding
- type is most often taken for granted
  - ie. *"My project is about cocktail parties. Why would I ever need to know that a cocktail party is not a business meeting?"*
- searching & indexing in heterogenous multi-party conversation recordings (or portions)
- **text-independence**: pre-ASR availability of type hypothesis/prior
  - may contribute to optimal selection of ASR components
  - type classification possible where no ASR or upstream processing possible

# Why Do Classification of Conversation Type?

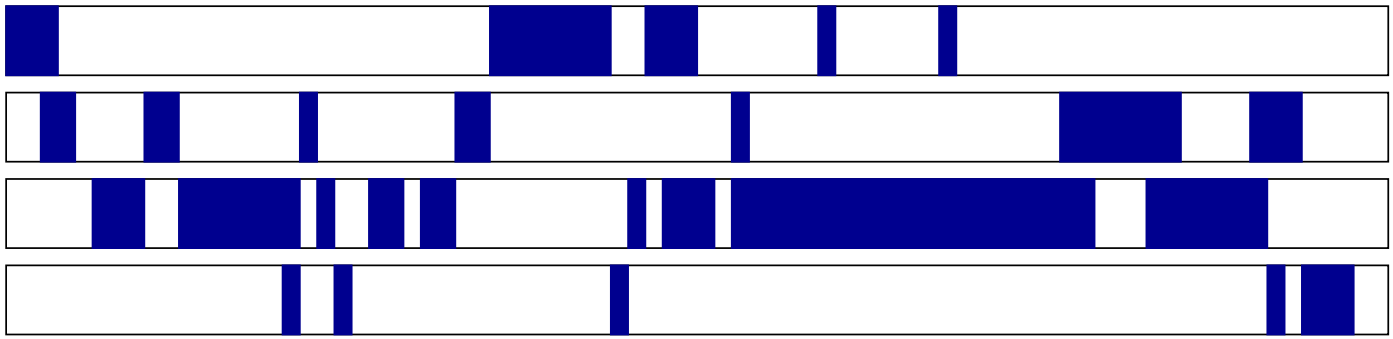
- a basic competence in conversation understanding
- type is most often taken for granted
  - ie. *"My project is about cocktail parties. Why would I ever need to know that a cocktail party is not a business meeting?"*
- searching & indexing in heterogenous multi-party conversation recordings (or portions)
- **text-independence**: pre-ASR availability of type hypothesis/prior
  - may contribute to optimal selection of ASR components
  - type classification possible where no ASR or upstream processing possible

# Why Do Classification of Conversation Type?

- a basic competence in conversation understanding
- type is most often taken for granted
  - ie. *"My project is about cocktail parties. Why would I ever need to know that a cocktail party is not a business meeting?"*
- searching & indexing in heterogenous multi-party conversation recordings (or portions)
- **text-independence**: pre-ASR availability of type hypothesis/prior
  - may contribute to optimal selection of ASR components
  - type classification possible where no ASR or upstream processing possible

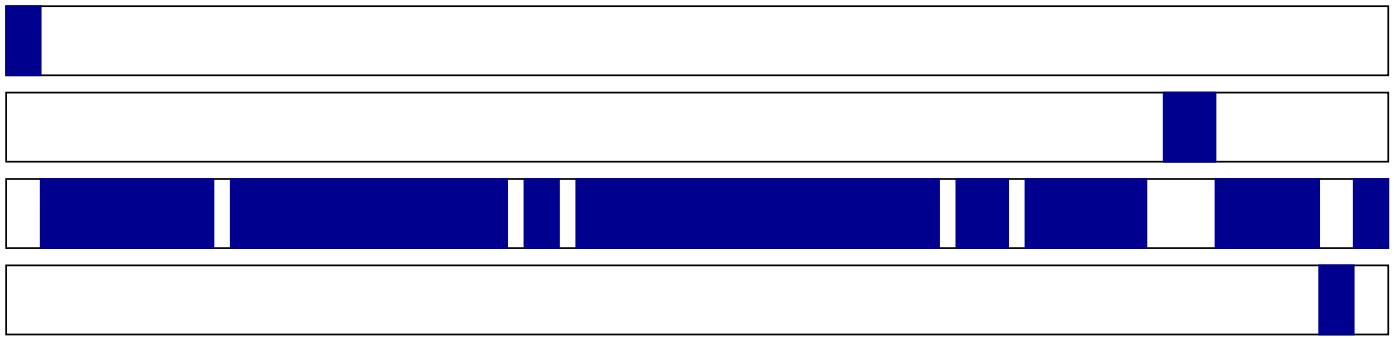
# Defining Conversation Type

- Sacks (1974) viewed conversation as one of several normative **speech-exchange system** types



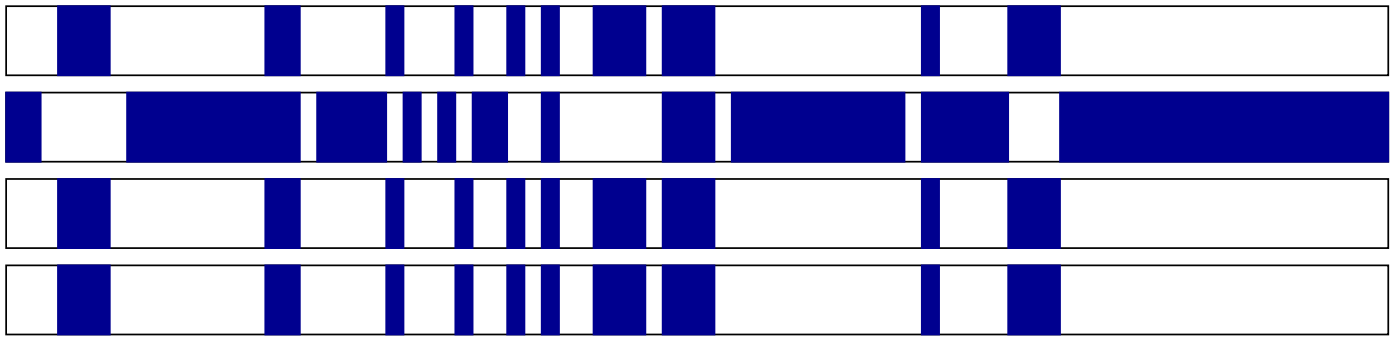
# Defining Conversation Type

- Sacks (1974) viewed conversation as one of several normative **speech-exchange system** types
- others include: **lectures**



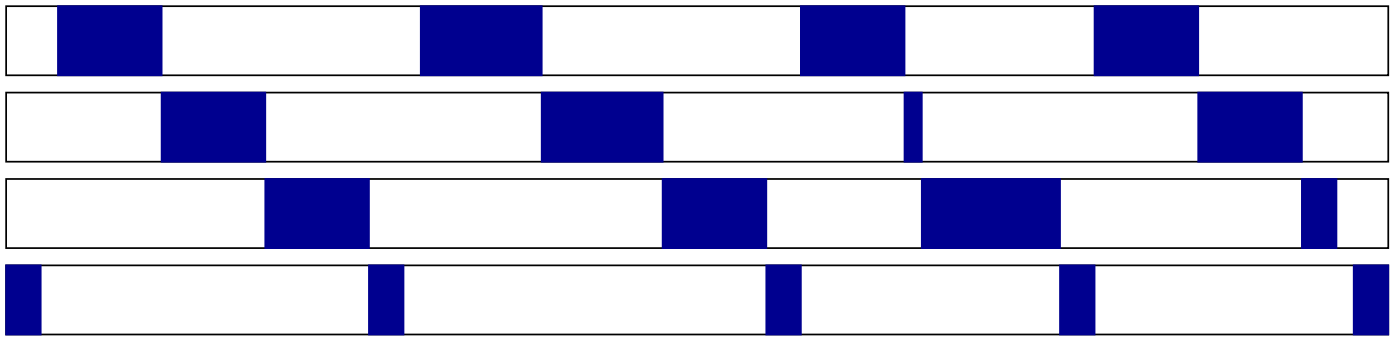
# Defining Conversation Type

- Sacks (1974) viewed conversation as one of several normative **speech-exchange system** types
- others include: lectures, **rituals**



# Defining Conversation Type

- Sacks (1974) viewed conversation as one of several normative **speech-exchange system** types
- others include: lectures, rituals, **debates**, etc.



# Defining Conversation Type

- Sacks (1974) viewed conversation as one of several normative **speech-exchange system** types
- others include: lectures, rituals, **debates**, etc.

- here, **type of conversation** $\equiv$ **subtype of work-related conversation** (meeting)
- implicitly assume that specific activities and specific participant groups and/or roles give rise to vocal interactions which are subtype-specific

# Defining Conversation Type

- Sacks (1974) viewed conversation as one of several normative **speech-exchange system** types
- others include: lectures, rituals, **debates**, etc.

- here, **type of conversation** $\equiv$ **subtype of work-related conversation** (meeting)
- implicitly assume that specific activities and specific participant groups and/or roles give rise to vocal interactions which are subtype-specific

# Related Work

- **none** on conversation type classification
- various, on evolving conversation state
  - (Banerjee & Rudnicky, 2004)
  - (McCowan et al, 2005)
  - (Zancanaro et al, 2006)
- several related text-independent tasks
  - participant dominance detection (Rienks et al, 2005), 4-party
  - interaction group recognition (Brdiczka et al, 2005), 4-party
  - conversational pair detection (Basu, 2002), 2-party
- modeling vocal interaction for vocal activity detection
  - meetings (Laskowski & Schultz, 2006)
  - ambulatory data (Wyatt et al, 2007)

# Related Work

- **none** on conversation type classification
- various, on evolving conversation state
  - (Banerjee & Rudnicky, 2004)
  - (McCowan et al, 2005)
  - (Zancanaro et al, 2006)
- several related text-independent tasks
  - participant dominance detection (Rienks et al, 2005), 4-party
  - interaction group recognition (Brdiczka et al, 2005), 4-party
  - conversational pair detection (Basu, 2002), 2-party
- modeling vocal interaction for vocal activity detection
  - meetings (Laskowski & Schultz, 2006)
  - ambulatory data (Wyatt et al, 2007)

# Related Work

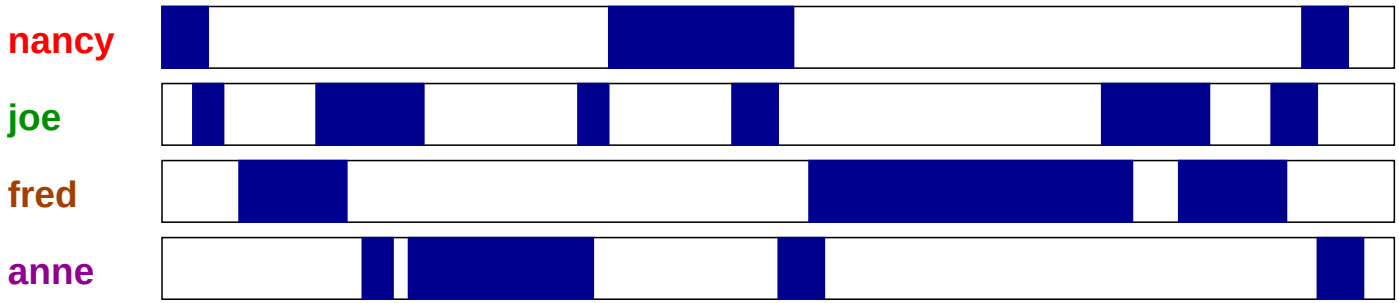
- **none** on conversation type classification
- various, on evolving conversation state
  - (Banerjee & Rudnicky, 2004)
  - (McCowan et al, 2005)
  - (Zancanaro et al, 2006)
- several related text-independent tasks
  - participant dominance detection (Rienks et al, 2005), 4-party
  - interaction group recognition (Brdiczka et al, 2005), 4-party
  - conversational pair detection (Basu, 2002), 2-party
- modeling vocal interaction for vocal activity detection
  - meetings (Laskowski & Schultz, 2006)
  - ambulatory data (Wyatt et al, 2007)

# Related Work

- **none** on conversation type classification
- various, on evolving conversation state
  - (Banerjee & Rudnicky, 2004)
  - (McCowan et al, 2005)
  - (Zancanaro et al, 2006)
- several related text-independent tasks
  - participant dominance detection (Rienks et al, 2005), 4-party
  - interaction group recognition (Brdiczka et al, 2005), 4-party
  - conversational pair detection (Basu, 2002), 2-party
- modeling vocal interaction for vocal activity detection
  - meetings (Laskowski & Schultz, 2006)
  - ambulatory data (Wyatt et al, 2007)

# Observables

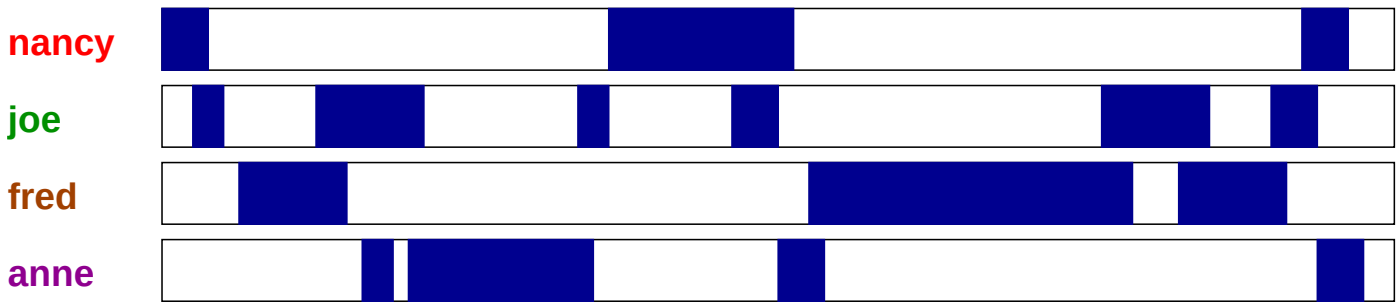
- the **vocal interaction** record of a conversation $\mathcal{C}$, of type $\mathcal{T}$ (of $N_{\mathcal{T}}$ possible conversation types)




- at time $t$, each of $K$ participants is in one of 2 discrete states, vocalizing ($\mathcal{V}$) or not vocalizing ($\mathcal{N}$)
- therefore, at time $t$, the state $\mathbf{q}_t$ of $\mathcal{C}$, as a whole, has one of $2^K$ discrete values

# Observables

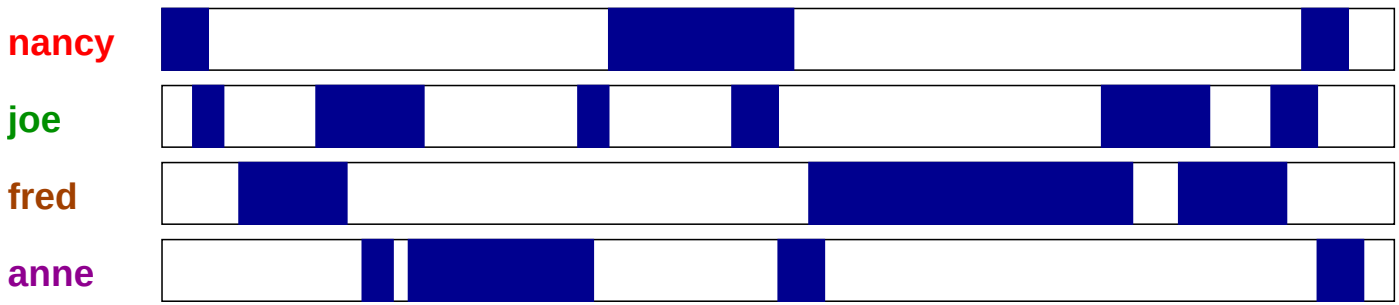
- the **vocal interaction** record of a conversation $\mathcal{C}$, of type $\mathcal{T}$ (of $N_{\mathcal{T}}$ possible conversation types)




- at time $t$, each of $K$ participants is in one of 2 discrete states, vocalizing ($\mathcal{V}$) or not vocalizing ($\mathcal{N}$)
- therefore, at time $t$, the state $\mathbf{q}_t$ of $\mathcal{C}$, as a whole, has one of $2^K$ discrete values

# Observables

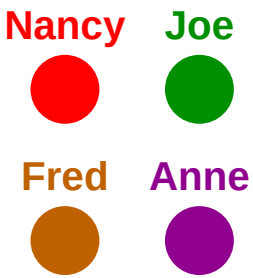
- the **vocal interaction** record of a conversation $\mathcal{C}$, of type $\mathcal{T}$ (of $N_{\mathcal{T}}$ possible conversation types)




- at time $t$, each of $K$ participants is in one of 2 discrete states, vocalizing ($\mathcal{V}$) or not vocalizing ($\mathcal{N}$)
- therefore, at time $t$, the state $\mathbf{q}_t$ of $\mathcal{C}$, as a whole, has one of $2^K$ discrete values

# Modeling Groups

- assume that $\mathcal{C}$ is a 1st order Markov process, produced by the ordered **group** $\mathcal{G}$ of $\|\mathcal{G}\| \equiv K$ specific participants




$\mathcal{G}$

- participants are drawn from a known population $\mathcal{P}$ of size $\|\mathcal{P}\|$
- the number of distinct groups of size $\|\mathcal{G}\| \leq \|\mathcal{P}\|$ is

$$N_{\mathcal{G}} = \frac{\|\mathcal{P}\|!}{(\|\mathcal{P}\| - \|\mathcal{G}\|)!}$$

# Modeling Groups

- assume that $\mathcal{C}$ is a 1st order Markov process, produced by the ordered **group** $\mathcal{G}$ of $\|\mathcal{G}\| \equiv K$ specific participants

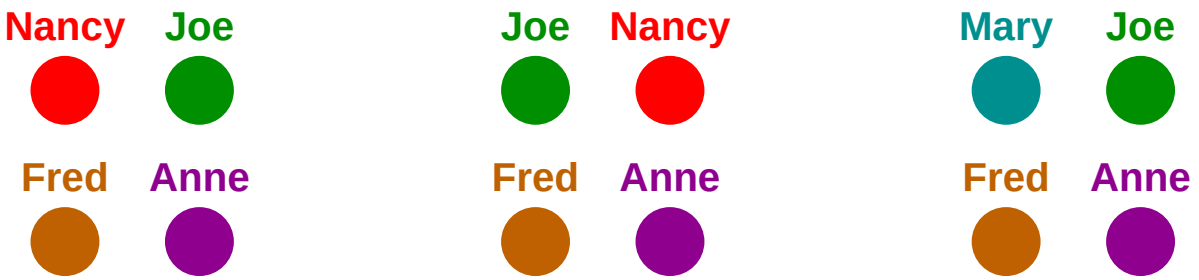

$\mathcal{G}$ $\mathcal{G}'$

- participants are drawn from a known population $\mathcal{P}$ of size $\|\mathcal{P}\|$
- the number of distinct groups of size $\|\mathcal{G}\| \leq \|\mathcal{P}\|$ is

$$N_{\mathcal{G}} = \frac{\|\mathcal{P}\|!}{(\|\mathcal{P}\| - \|\mathcal{G}\|)!}$$

# Modeling Groups

- assume that $\mathcal{C}$ is a 1st order Markov process, produced by the ordered **group** $\mathcal{G}$ of $\|\mathcal{G}\| \equiv K$ specific participants



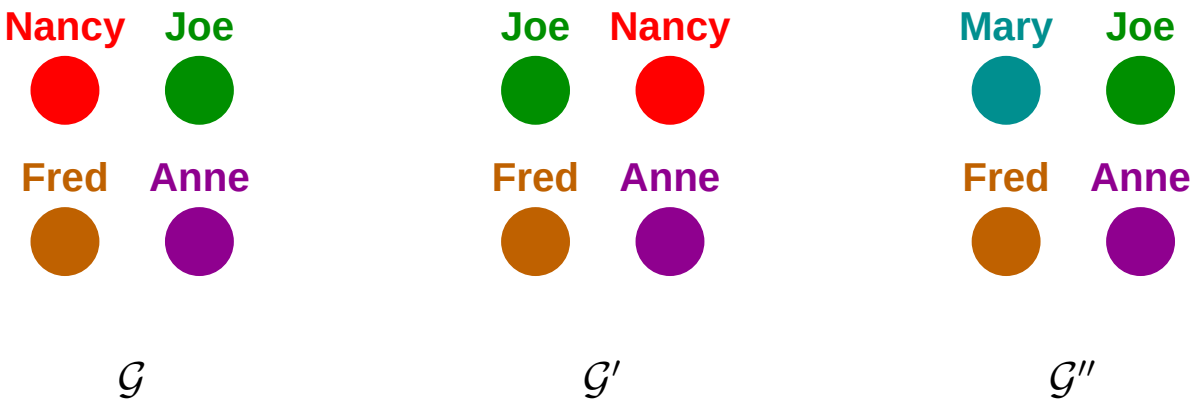

$\mathcal{G}$ $\mathcal{G}'$ $\mathcal{G}''$

- participants are drawn from a known population $\mathcal{P}$ of size $\|\mathcal{P}\|$
- the number of distinct groups of size $\|\mathcal{G}\| \leq \|\mathcal{P}\|$ is

$$N_{\mathcal{G}} = \frac{\|\mathcal{P}\|!}{(\|\mathcal{P}\| - \|\mathcal{G}\|)!}$$

# Modeling Groups

- assume that $\mathcal{C}$ is a 1st order Markov process, produced by the ordered **group** $\mathcal{G}$ of $\|\mathcal{G}\| \equiv K$ specific participants



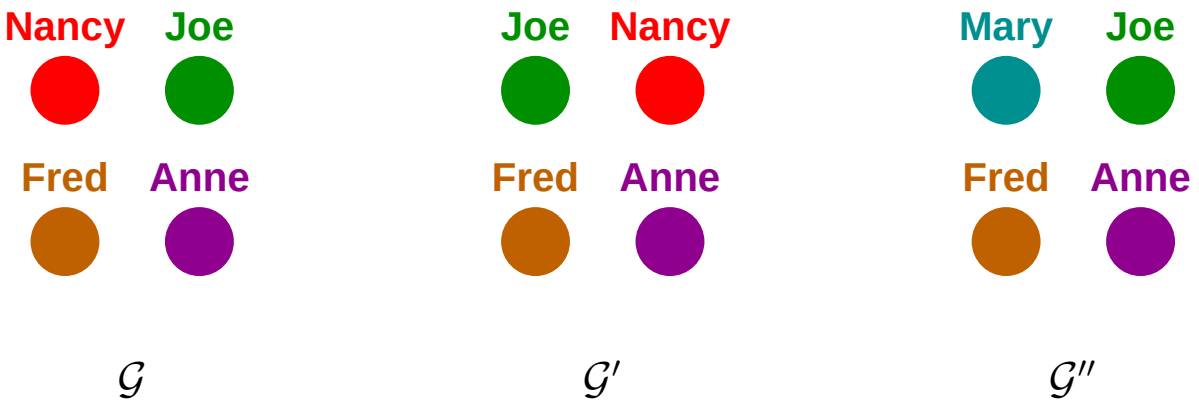

$\mathcal{G}$ $\mathcal{G}'$ $\mathcal{G}''$

- participants are drawn from a known population $\mathcal{P}$ of size $\|\mathcal{P}\|$
- the number of distinct groups of size $\|\mathcal{G}\| \leq \|\mathcal{P}\|$ is

$$N_{\mathcal{G}} = \frac{\|\mathcal{P}\|!}{(\|\mathcal{P}\| - \|\mathcal{G}\|)!}$$

# Modeling Groups

- assume that $\mathcal{C}$ is a 1st order Markov process, produced by the ordered **group** $\mathcal{G}$ of $\|\mathcal{G}\| \equiv K$ specific participants




$\mathcal{G}$ $\mathcal{G}'$ $\mathcal{G}''$

- participants are drawn from a known population $\mathcal{P}$ of size $\|\mathcal{P}\|$
- the number of distinct groups of size $\|\mathcal{G}\| \leq \|\mathcal{P}\|$ is

$$N_{\mathcal{G}} = \frac{\|\mathcal{P}\|!}{(\|\mathcal{P}\| - \|\mathcal{G}\|)!}$$

# Conversation Type Classification

- participant identities, and therefore $\mathcal{G}$, are **hidden variables**
- given a set of features $\mathbf{F}$ extracted from $\mathcal{C}$,

$$\begin{aligned}
\mathcal{T}^* &= \arg\max_{\mathcal{T}} P(\mathcal{T} \mid \mathbf{F}) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P(\mathcal{G}, \mathcal{T}, \mathbf{F}) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P(\mathcal{T}) \times \underbrace{P(\mathcal{G} \mid \mathcal{T})}_{\text{Membership Model}} \times \underbrace{P(\mathbf{F} \mid \mathcal{G}, \mathcal{T})}_{\text{Behavior Model}}
\end{aligned}$$

- hypothesis testing: cycle through $N_{\mathcal{T}}$ types and $N_{\mathcal{G}}$ groups

# Conversation Type Classification

- participant identities, and therefore $\mathcal{G}$, are **hidden variables**
- given a set of features **F** extracted from $\mathcal{C}$,

$$\begin{aligned}
\mathcal{T}^* &= \arg\max_{\mathcal{T}} P(\mathcal{T} \mid \mathbf{F}) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P(\mathcal{G}, \mathcal{T}, \mathbf{F}) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P(\mathcal{T}) \times \underbrace{P(\mathcal{G} \mid \mathcal{T})}_{\text{Membership Model}} \times \underbrace{P(\mathbf{F} \mid \mathcal{G}, \mathcal{T})}_{\text{Behavior Model}}
\end{aligned}$$

- hypothesis testing: cycle through $N_{\mathcal{T}}$ types and $N_{\mathcal{G}}$ groups

# Conversation Type Classification

- participant identities, and therefore $\mathcal{G}$, are **hidden variables**
- given a set of features $\mathbf{F}$ extracted from $\mathcal{C}$,

$$\begin{aligned}
\mathcal{T}^* &= \arg\max_{\mathcal{T}} P(\mathcal{T} \mid \mathbf{F}) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P(\mathcal{G}, \mathcal{T}, \mathbf{F}) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P(\mathcal{T}) \times \underbrace{P(\mathcal{G} \mid \mathcal{T})}_{\text{Membership Model}} \times \underbrace{P(\mathbf{F} \mid \mathcal{G}, \mathcal{T})}_{\text{Behavior Model}}
\end{aligned}$$

- hypothesis testing: cycle through $N_{\mathcal{T}}$ types and $N_{\mathcal{G}}$ groups

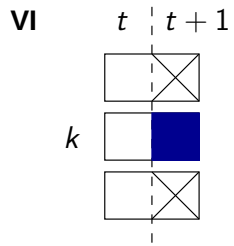
# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)

- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)

# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)




- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)

# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)


- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)

# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)

$$
\begin{aligned}
f_k^{VI} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i\right) \\
f_k^{VC} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[k] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i \neq k\right)
\end{aligned}
$$

- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)

# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)

$$
\begin{aligned}
f_k^{VI} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i\right) \\
f_k^{VC} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[k] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i \neq k\right)
\end{aligned}
$$

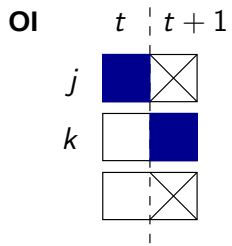
- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)

# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)

$$
\begin{aligned}
f_k^{VI} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i\right) \\
f_k^{VC} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[k] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i \neq k\right)
\end{aligned}
$$

- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)

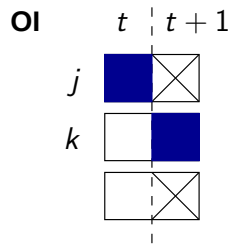


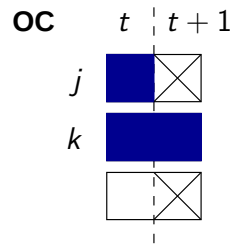
# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)

$$
\begin{aligned}
f_k^{VI} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i\right) \\
f_k^{VC} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[k] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i \neq k\right)
\end{aligned}
$$

- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)



# Features

- probability, when no-one else is vocalizing, that $k$ initiates vocalization (VI) and that $k$ continues vocalization (VC)

$$
\begin{aligned}
f_k^{VI} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i\right) \\
f_k^{VC} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[k] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i \neq k\right)
\end{aligned}
$$

- probability, when $j$ is vocalizing, that $k$ initiates vocalization overlap (OI) and that $k$ continues vocalization overlap (OC)

$$
\begin{aligned}
f_{k,j}^{OI} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[j] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall\, i \neq j\right) \\
f_{k,j}^{OC} &= P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t[k] = \mathbf{q}_t[j] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N}\right. \\
&\qquad\qquad \left.\forall\, i \neq j,\, i \neq k\right)
\end{aligned}
$$

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames

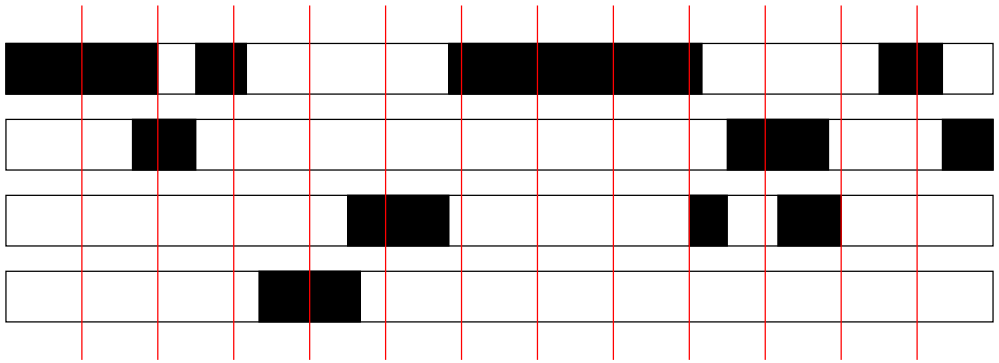
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



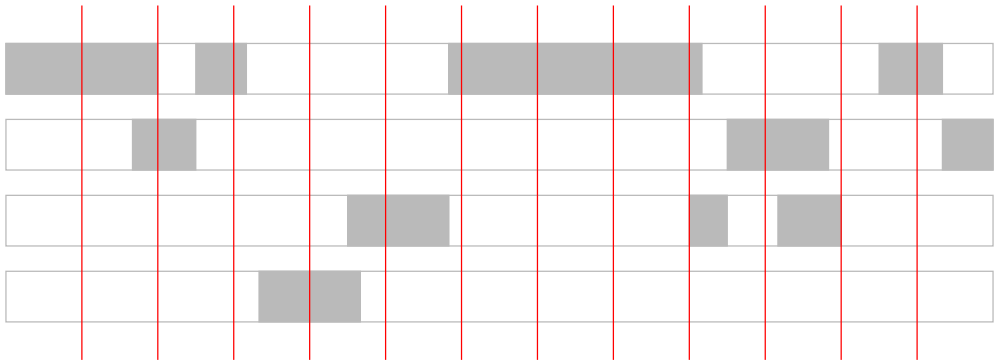
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



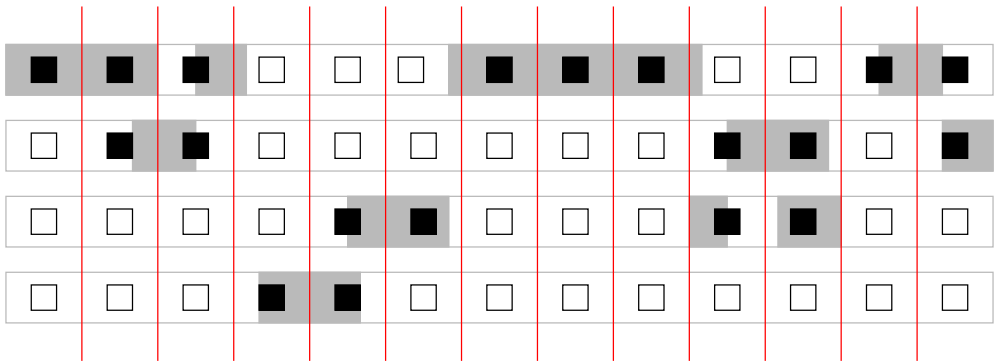
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



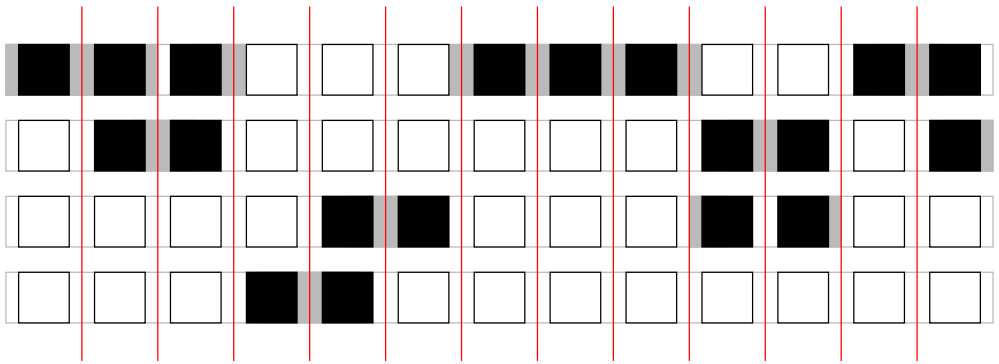
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



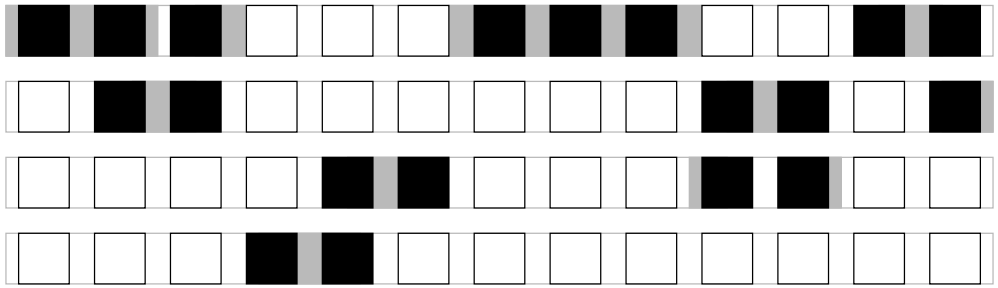
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



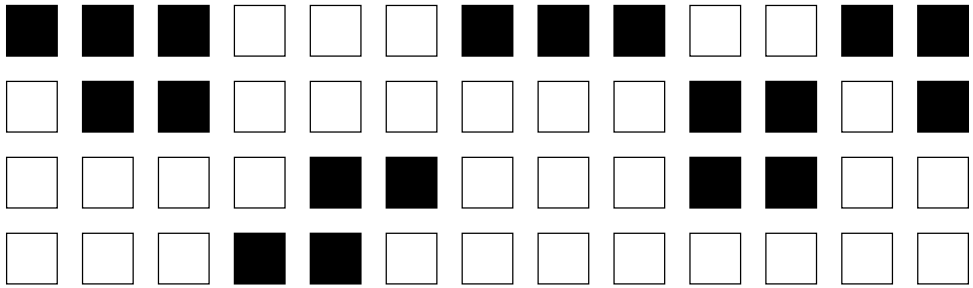
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



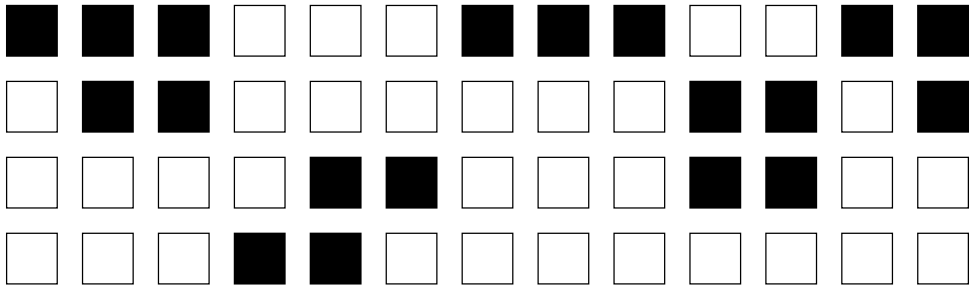
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



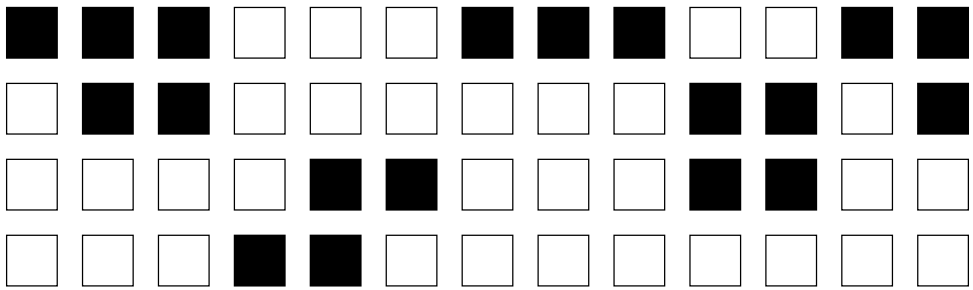
- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Feature Estimation

- need to estimate all probabilities in feature vector **F**:

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\}$$

- discretize the vocal interaction record using 200 ms frames



- estimate features using maximum likelihood (ML)
- probabilities with unseen conditioning contexts are set to 0.5

# Alternate Feature Estimation Method

- use a variant of a model from stochastic dynamics, the **Ising model** (Glauber, 1963)
- assume the conditional probability of vocal activity state transition, for each $k$

$$P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t = \mathbf{S}_i\right) \quad = \quad y_k\left(\mathbf{S}_i\right)$$

- where

$$y_k\left(\mathbf{x}\right) \quad = \quad \frac{1}{1 + e^{-\beta\left(\sum_{j=1}^{K} w_{k,j} x_j + b_k\right)}}$$

- not coincidentally, this is a one-layer neural network
- obviates the need for designing a back-off/smoothing strategy in ML estimation of features

# Alternate Feature Estimation Method

- use a variant of a model from stochastic dynamics, the **Ising model** (Glauber, 1963)
- assume the conditional probability of vocal activity state transition, for each $k$

$$P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t = \mathbf{S}_i\right) \quad = \quad y_k\left(\mathbf{S}_i\right)$$

- where

$$y_k(\mathbf{x}) \quad = \quad \frac{1}{1 + e^{-\beta\left(\sum_{j=1}^{K} w_{k,j} x_j + b_k\right)}}$$

- not coincidentally, this is a one-layer neural network
- obviates the need for designing a back-off/smoothing strategy in ML estimation of features

# Alternate Feature Estimation Method

- use a variant of a model from stochastic dynamics, the **Ising model** (Glauber, 1963)
- assume the conditional probability of vocal activity state transition, for each $k$

$$P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t = \mathbf{S}_i\right) \quad = \quad y_k\left(\mathbf{S}_i\right)$$

- where

$$y_k\left(\mathbf{x}\right) \quad = \quad \frac{1}{1 + e^{-\beta\left(\sum_{j=1}^{K} w_{k,j} x_j + b_k\right)}}$$

- not coincidentally, this is a one-layer neural network
- obviates the need for designing a back-off/smoothing strategy in ML estimation of features

# Alternate Feature Estimation Method

- use a variant of a model from stochastic dynamics, the **Ising model** (Glauber, 1963)
- assume the conditional probability of vocal activity state transition, for each $k$

$$P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t = \mathbf{S}_i\right) \quad = \quad y_k\left(\mathbf{S}_i\right)$$

- where

$$y_k\left(\mathbf{x}\right) \quad = \quad \frac{1}{1 + e^{-\beta\left(\sum_{j=1}^{K} w_{k,j} x_j + b_k\right)}}$$

- not coincidentally, this is a one-layer neural network
- obviates the need for designing a back-off/smoothing strategy in ML estimation of features

# Alternate Feature Estimation Method

- use a variant of a model from stochastic dynamics, the **Ising model** (Glauber, 1963)
- assume the conditional probability of vocal activity state transition, for each $k$

$$P\left(\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t = \mathbf{S}_i\right) \quad = \quad y_k\left(\mathbf{S}_i\right)$$

- where

$$y_k(\mathbf{x}) \quad = \quad \frac{1}{1 + e^{-\beta\left(\sum_{j=1}^{K} w_{k,j} x_j + b_k\right)}}$$

- not coincidentally, this is a one-layer neural network
- obviates the need for designing a back-off/smoothing strategy in ML estimation of features

# Behavior Model

- for each conversation type $\mathcal{T}$ and each group $\mathcal{G}$, require the likelihood of **F** (as estimated from the observed vocal interaction record)

$$P(\mathbf{F} \mid \mathcal{G}, \mathcal{T}) = \prod_{k=1}^{K} P\left(f_k^{VI} \mid \theta^{VI}_{\mathcal{T},\mathcal{G}[k]}\right) P\left(f_k^{VC} \mid \theta^{VC}_{\mathcal{T},\mathcal{G}[k]}\right) \times \prod_{j \neq k}^{K} P\left(f_{k,j}^{OI} \mid \theta^{OI}_{\mathcal{T},\mathcal{G}[k],\mathcal{G}[j]}\right) P\left(f_{k,j}^{OC} \mid \theta^{OC}_{\mathcal{T},\mathcal{G}[k],\mathcal{G}[j]}\right)$$

- each $\theta$ represents a single one-dimensional Gaussian mean $\mu$ and variance $\Sigma$ pair

# Behavior Model

- for each conversation type $\mathcal{T}$ and each group $\mathcal{G}$, require the likelihood of **F** (as estimated from the observed vocal interaction record)

$$P(\mathbf{F} \mid \mathcal{G}, \mathcal{T}) = \prod_{k=1}^{K} P\left(f_k^{VI} \mid \theta^{VI}_{\mathcal{T},\mathcal{G}[k]}\right) P\left(f_k^{VC} \mid \theta^{VC}_{\mathcal{T},\mathcal{G}[k]}\right) \times \prod_{j \neq k}^{K} P\left(f_{k,j}^{OI} \mid \theta^{OI}_{\mathcal{T},\mathcal{G}[k],\mathcal{G}[j]}\right) P\left(f_{k,j}^{OC} \mid \theta^{OC}_{\mathcal{T},\mathcal{G}[k],\mathcal{G}[j]}\right)$$

- each $\theta$ represents a single one-dimensional Gaussian mean $\mu$ and variance $\Sigma$ pair

# Membership Model

- for each conversation type $\mathcal{T}$, require the probability of group $\mathcal{G}$ (as hypothesized)

$$P(\mathcal{G}\,|\,\mathcal{T}) \;=\; \frac{1}{Z_{\mathcal{G}}} \prod_{k=1}^{K} P(\mathcal{G}[k]\,|\,\mathcal{T})$$

- $Z_{\mathcal{G}}$ is a normalization constant, $\sum_{N_{\mathcal{G}}} P(\mathcal{G}|\mathcal{T}) = 1$

# Membership Model

- for each conversation type $\mathcal{T}$, require the probability of group $\mathcal{G}$ (as hypothesized)

$$P(\mathcal{G}\,|\,\mathcal{T}) \;=\; \frac{1}{Z_{\mathcal{G}}} \prod_{k=1}^{K} P(\mathcal{G}[k]\,|\,\mathcal{T})$$

- $Z_{\mathcal{G}}$ is a normalization constant, $\sum_{N_{\mathcal{G}}} P(\mathcal{G}|\mathcal{T}) = 1$

# The ICSI Meeting Corpus

(Janin et al, 2003), (Shriberg et al, 2004)

- naturally occurring project-oriented conversations
- for our purposes, 4 types of longitudinal collections:

| type | # of meetings | # of possible participants | # of participants | | |
|---|---|---|---|---|---|
| | | | mod | min | max |
| Bed | 15 | 13 | 6 | 4 | 7 |
| Bmr | 29 | 15 | 7 | 3 | 9 |
| Bro | 23 | 10 | 6 | 4 | 8 |
| other | 8 | 27 | 6 | 5 | 8 |

- "other" contains types of which there are $\leq 3$ meetings
- rarely, meetings contain additional, uninstrumented participants (whose contributions we ignore)

# The ICSI Meeting Corpus

(Janin et al, 2003), (Shriberg et al, 2004)

- naturally occurring project-oriented conversations
- for our purposes, 4 types of longitudinal collections:

| type | # of meetings | # of possible participants | # of participants | | |
|---|---|---|---|---|---|
| | | | mod | min | max |
| Bed | 15 | 13 | 6 | 4 | 7 |
| Bmr | 29 | 15 | 7 | 3 | 9 |
| Bro | 23 | 10 | 6 | 4 | 8 |
| other | 8 | 27 | 6 | 5 | 8 |

- "other" contains types of which there are $\leq 3$ meetings
- rarely, meetings contain additional, uninstrumented participants (whose contributions we ignore)

# The ICSI Meeting Corpus

(Janin et al, 2003), (Shriberg et al, 2004)

- naturally occurring project-oriented conversations
- for our purposes, 4 types of longitudinal collections:

| type | # of meetings | # of possible participants | # of participants | | |
|---|---|---|---|---|---|
| | | | mod | min | max |
| Bed | 15 | 13 | 6 | 4 | 7 |
| Bmr | 29 | 15 | 7 | 3 | 9 |
| Bro | 23 | 10 | 6 | 4 | 8 |
| other | 8 | 27 | 6 | 5 | 8 |

- "other" contains types of which there are $\leq 3$ meetings
- rarely, meetings contain additional, uninstrumented participants (whose contributions we ignore)

# The ICSI Meeting Corpus

(Janin et al, 2003), (Shriberg et al, 2004)

- naturally occurring project-oriented conversations
- for our purposes, 4 types of longitudinal collections:

| type | # of meetings | # of possible participants | # of participants | | |
|---|---|---|---|---|---|
| | | | mod | min | max |
| Bed | 15 | 13 | 6 | 4 | 7 |
| Bmr | 29 | 15 | 7 | 3 | 9 |
| Bro | 23 | 10 | 6 | 4 | 8 |
| other | 8 | 27 | 6 | 5 | 8 |

- "other" contains types of which there are $\leq$3 meetings
- rarely, meetings contain additional, uninstrumented participants (whose contributions we ignore)

# The ICSI Meeting Corpus

(Janin et al, 2003), (Shriberg et al, 2004)

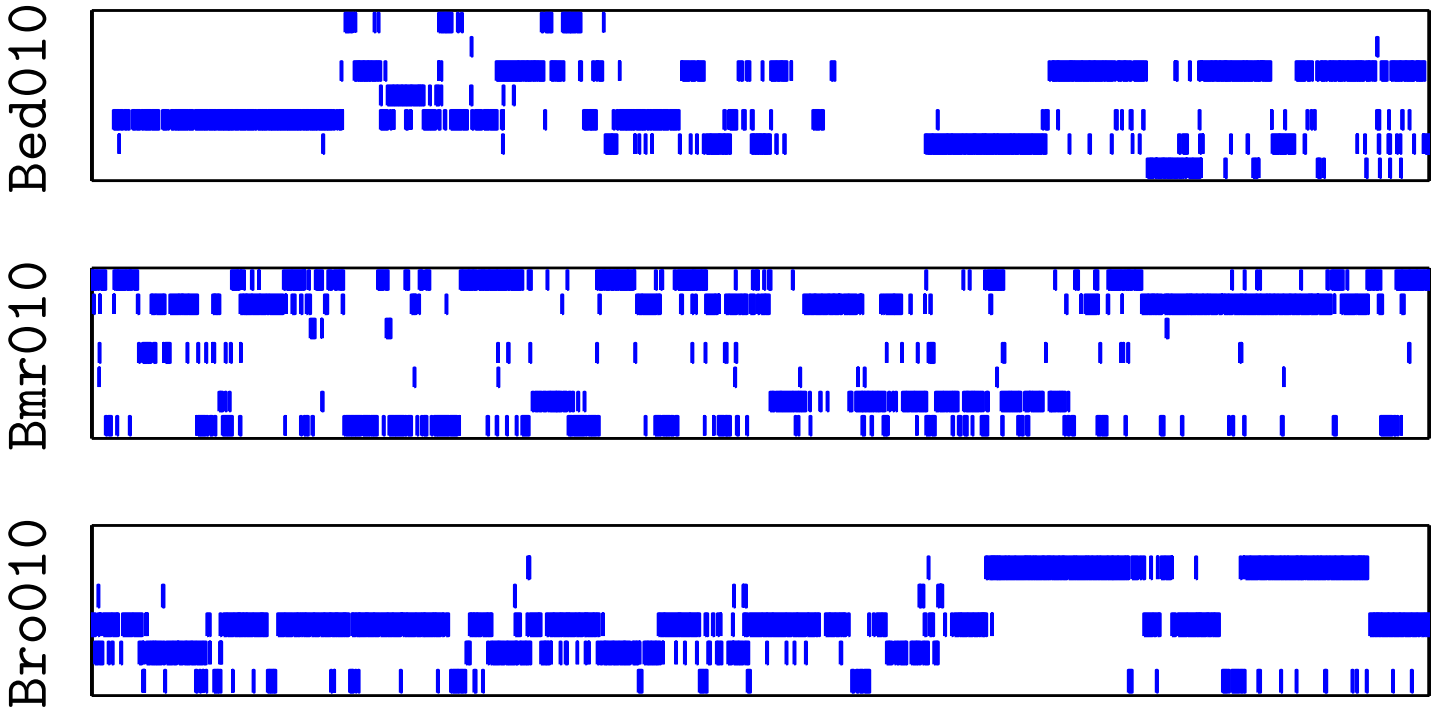
- naturally occurring project-oriented conversations
- for our purposes, 4 types of longitudinal collections:

| type | # of meetings | # of possible participants | # of participants | | |
|---|---|---|---|---|---|
| | | | mod | min | max |
| Bed | 15 | 13 | 6 | 4 | 7 |
| Bmr | 29 | 15 | 7 | 3 | 9 |
| Bro | 23 | 10 | 6 | 4 | 8 |
| other | 8 | 27 | 6 | 5 | 8 |

- "other" contains types of which there are $\leq$3 meetings
- rarely, meetings contain additional, uninstrumented participants (whose contributions we ignore)

# Differences Between Meeting Types

- 36-minute excerpts (from 1000 sec to 2000 sec)



# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
- cluster participants for training the behavior model
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
- cluster participants for training the behavior model
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
  - train on 65 meetings, test on 1 meeting, rotate
  - too little data for a true, unseen evaluation set
- cluster participants for training the behavior model
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
  - train on 65 meetings, test on 1 meeting, rotate
  - too little data for a true, unseen evaluation set
- cluster participants for training the behavior model
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
  - train on 65 meetings, test on 1 meeting, rotate
  - too little data for a true, unseen evaluation set
- cluster participants for training the behavior model
  - side effect: reduces impact of mismatched segmentation
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
  - train on 65 meetings, test on 1 meeting, rotate
  - too little data for a true, unseen evaluation set
- cluster participants for training the behavior model
  - side-effect: renders impact of membership model negligible
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
  - train on 65 meetings, test on 1 meeting, rotate
  - too little data for a true, unseen evaluation set
- cluster participants for training the behavior model
  - side-effect: renders impact of membership model negligible
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
  - train on 65 meetings, test on 1 meeting, rotate
  - too little data for a true, unseen evaluation set
- cluster participants for training the behavior model
  - side-effect: renders impact of membership model negligible
- accuracy: **65.7%** (random guessing: 43%)

# Baseline

- use inclusive-OR of "talk-spurt" (Shriberg et al, 2001) and "laugh-bout" (Laskowski & Burger, 2007) segmentations
- compute a single feature $f_k^T$: **vocalizing time proportion**
  - employed for assessing speaker diarization performance (Jin et al, 2004), (Mirghafori & Wooters, 2006)
  - captures "flatness" of speaking-time distribution across speakers
- leave-one-out classification
  - train on 65 meetings, test on 1 meeting, rotate
  - too little data for a true, unseen evaluation set
- cluster participants for training the behavior model
  - side-effect: renders impact of membership model negligible
- accuracy: **65.7%** (random guessing: 43%)

# Feature Comparison

| Feature(s) | ML Estimation | | NN Estimation | |
|---|---|---|---|---|
| | w/o $f_k^T$ | w/ $f_k^T$ | w/o $f_k^T$ | w/ $f_k^T$ |
| baseline | — | 65.7 | — | 65.7 |
| $f_k^{VI}$ | 59.7 | 67.2 | 56.7 | 65.7 |
| $f_k^{VC}$ | 62.7 | 77.6 | 56.7 | 71.6 |
| $\langle f_{k,j}^{OI} \rangle_j$ | 35.8 | 52.2 | 64.2 | 67.2 |
| $\langle f_{k,j}^{OC} \rangle_j$ | 53.7 | 67.2 | 64.2 | 80.6 |
| $f_{k,j}^{OI}$ | 41.8 | 46.3 | 67.2 | 64.2 |
| $f_{k,j}^{OC}$ | 61.2 | 68.7 | 73.1 | 79.1 |

# Feature Comparison

| Feature(s) | ML Estimation | | NN Estimation | |
|---|---|---|---|---|
| | w/o $f_k^T$ | w/ $f_k^T$ | w/o $f_k^T$ | w/ $f_k^T$ |
| baseline | — | **65.7** | — | **65.7** |
| $f_k^{VI}$ | **59.7** | 67.2 | **56.7** | 65.7 |
| $f_k^{VC}$ | **62.7** | 77.6 | **56.7** | 71.6 |
| $\langle f_{k,j}^{OI} \rangle_j$ | **35.8** | 52.2 | **64.2** | 67.2 |
| $\langle f_{k,j}^{OC} \rangle_j$ | **53.7** | 67.2 | **64.2** | 80.6 |
| $f_{k,j}^{OI}$ | **41.8** | 46.3 | 67.2 | 64.2 |
| $f_{k,j}^{OC}$ | **61.2** | 68.7 | 73.1 | 79.1 |

- the baseline feature $f_k^T$ outperforms most other features

# Feature Comparison

| Feature(s) | ML Estimation | | NN Estimation | |
|---|---|---|---|---|
| | w/o $f_k^T$ | w/ $f_k^T$ | w/o $f_k^T$ | w/ $f_k^T$ |
| baseline | — | 65.7 | — | 65.7 |
| $f_k^{VI}$ | 59.7 | 67.2 | 56.7 | 65.7 |
| $f_k^{VC}$ | 62.7 | 77.6 | 56.7 | 71.6 |
| $\langle f_{k,j}^{OI} \rangle_j$ | 35.8 | 52.2 | 64.2 | 67.2 |
| $\langle f_{k,j}^{OC} \rangle_j$ | 53.7 | 67.2 | 64.2 | 80.6 |
| $f_{k,j}^{OI}$ | **41.8** | 46.3 | **67.2** | 64.2 |
| $f_{k,j}^{OC}$ | **61.2** | 68.7 | **73.1** | 79.1 |

- by themselves, specific participant-pair features outperform each participant's average participant-pair features

# Feature Comparison

| Feature(s) | ML Estimation | | NN Estimation | |
|---|---|---|---|---|
| | w/o $f_k^T$ | w/ $f_k^T$ | w/o $f_k^T$ | w/ $f_k^T$ |
| baseline | — | 65.7 | — | 65.7 |
| $f_k^{VI}$ | 59.7 | 67.2 | 56.7 | 65.7 |
| $f_k^{VC}$ | 62.7 | 77.6 | 56.7 | 71.6 |
| $\langle f_{k,j}^{OI} \rangle_j$ | 35.8 | 52.2 | 64.2 | **67.2** |
| $\langle f_{k,j}^{OC} \rangle_j$ | 53.7 | **67.2** | 64.2 | **80.6** |
| $f_{k,j}^{OI}$ | 41.8 | 46.3 | 67.2 | 64.2 |
| $f_{k,j}^{OC}$ | 61.2 | **68.7** | 73.1 | **79.1** |

- most features, when combined with $f_k^T$, lead to improved performance

# Feature Comparison

| Feature(s) | ML Estimation | | NN Estimation | |
|---|---|---|---|---|
| | w/o $f_k^T$ | w/ $f_k^T$ | w/o $f_k^T$ | w/ $f_k^T$ |
| baseline | — | 65.7 | — | 65.7 |
| $f_k^{VI}$ | 59.7 | 67.2 | 56.7 | 65.7 |
| $f_k^{VC}$ | 62.7 | 77.6 | 56.7 | 71.6 |
| $\langle f_{k,j}^{OI} \rangle_j$ | 35.8 | 52.2 | 64.2 | 67.2 |
| $\langle f_{k,j}^{OC} \rangle_j$ | 53.7 | 67.2 | 64.2 | 80.6 |
| $f_{k,j}^{OI}$ | 41.8 | 46.3 | 67.2 | 64.2 |
| $f_{k,j}^{OC}$ | 61.2 | 68.7 | 73.1 | 79.1 |

- all NN-estimated features together yield **82.1%**
- an optimal NN-estimated feature subset (forward selection) yields **83.6%**

# Feature Comparison

| Feature(s) | ML Estimation | | NN Estimation | |
|---|---|---|---|---|
| | w/o $f_k^T$ | w/ $f_k^T$ | w/o $f_k^T$ | w/ $f_k^T$ |
| baseline | — | 65.7 | — | 65.7 |
| $f_k^{VI}$ | 59.7 | 67.2 | 56.7 | 65.7 |
| $f_k^{VC}$ | 62.7 | 77.6 | 56.7 | 71.6 |
| $\langle f_{k,j}^{OI} \rangle_j$ | 35.8 | 52.2 | 64.2 | 67.2 |
| $\langle f_{k,j}^{OC} \rangle_j$ | 53.7 | 67.2 | 64.2 | 80.6 |
| $f_{k,j}^{OI}$ | 41.8 | 46.3 | 67.2 | 64.2 |
| $f_{k,j}^{OC}$ | 61.2 | 68.7 | 73.1 | 79.1 |

- all NN-estimated features together yield **82.1%**
- an optimal NN-estimated feature subset (forward selection) yields **83.6%**

# Discussion

- 3-way confusion matrix, optimal NN-estimated feature subset

| Estimated | Actual Type | | |
|---|---|---|---|
| | Bed | Bmr | Bro |
| Bed | **11** | 1 | 3 |
| Bmr | 2 | **26** | 1 |
| Bro | 3 | 1 | **19** |

  - Bmr (discussions among peers) is the most distinct type
  - Bed and Bro (both more structured meetings) are more similar to each other than to Bmr

- miss-classification pattern reflects intuition

# Discussion

- 3-way confusion matrix, optimal NN-estimated feature subset

| Estimated | Actual Type | | |
|---|---|---|---|
| | Bed | Bmr | Bro |
| Bed | **11** | 1 | 3 |
| Bmr | 2 | **26** | 1 |
| Bro | 3 | 1 | **19** |

  - Bmr (discussions among peers) is the most distinct type
  - Bed and Bro (both more structured meetings) are more similar to each other than to Bmr

- miss-classification pattern reflects intuition

# Discussion

- 3-way confusion matrix, optimal NN-estimated feature subset

| Estimated | Actual Type | | |
|---|---|---|---|
| | Bed | Bmr | Bro |
| Bed | **11** | 1 | 3 |
| Bmr | 2 | **26** | 1 |
| Bro | 3 | 1 | **19** |

  - Bmr (discussions among peers) is the most distinct type
  - Bed and Bro (both more structured meetings) are more similar to each other than to Bmr

- miss-classification pattern reflects intuition

# Discussion

- 3-way confusion matrix, optimal NN-estimated feature subset

| Estimated | Actual Type | | |
|---|---|---|---|
| | Bed | Bmr | Bro |
| Bed | **11** | 1 | 3 |
| Bmr | 2 | **26** | 1 |
| Bro | 3 | 1 | **19** |

  - Bmr (discussions among peers) is the most distinct type
  - Bed and Bro (both more structured meetings) are more similar to each other than to Bmr

- miss-classification pattern reflects intuition

# Conclusions

- classification paradigm with several novel elements:
  1. exclusively text-independent features, from vocal interaction patterns
  2. participant groups, allowing for modeling multi-participant behaviors
  3. Ising model assumption of $\mathcal{C}$ transition probabilities
- meeting sub-type classification accuracy: 83%
- relative error reduction of 52% over the baseline
- (specific) multi-participant interaction features play a large role in this improvement

# Conclusions

- classification paradigm with several novel elements:
  1. exclusively text-independent features, from vocal interaction patterns
  2. participant groups, allowing for modeling multi-participant behaviors
  3. Ising model assumption of $\mathcal{C}$ transition probabilities
- meeting sub-type classification accuracy: 83%
- relative error reduction of 52% over the baseline
- (specific) multi-participant interaction features play a large role in this improvement

# Conclusions

- classification paradigm with several novel elements:
  1. exclusively text-independent features, from vocal interaction patterns
  2. participant groups, allowing for modeling multi-participant behaviors
  3. Ising model assumption of $\mathcal{C}$ transition probabilities
- meeting sub-type classification accuracy: 83%
- relative error reduction of 52% over the baseline
- (specific) multi-participant interaction features play a large role in this improvement

# Conclusions

- classification paradigm with several novel elements:
  1. exclusively text-independent features, from vocal interaction patterns
  2. participant groups, allowing for modeling multi-participant behaviors
  3. Ising model assumption of $\mathcal{C}$ transition probabilities
- meeting sub-type classification accuracy: 83%
- relative error reduction of 52% over the baseline
- (specific) multi-participant interaction features play a large role in this improvement

# Conclusions

- classification paradigm with several novel elements:
  1. exclusively text-independent features, from vocal interaction patterns
  2. participant groups, allowing for modeling multi-participant behaviors
  3. Ising model assumption of $\mathcal{C}$ transition probabilities
- meeting sub-type classification accuracy: 83%
- relative error reduction of 52% over the baseline
- (specific) multi-participant interaction features play a large role in this improvement

# Conclusions

- classification paradigm with several novel elements:
  1. exclusively text-independent features, from vocal interaction patterns
  2. participant groups, allowing for modeling multi-participant behaviors
  3. Ising model assumption of $\mathcal{C}$ transition probabilities
- meeting sub-type classification accuracy: 83%
- relative error reduction of 52% over the baseline
- (specific) multi-participant interaction features play a large role in this improvement

# Conclusions

- classification paradigm with several novel elements:
  1. exclusively text-independent features, from vocal interaction patterns
  2. participant groups, allowing for modeling multi-participant behaviors
  3. Ising model assumption of $\mathcal{C}$ transition probabilities
- meeting sub-type classification accuracy: 83%
- relative error reduction of 52% over the baseline
- (specific) multi-participant interaction features play a large role in this improvement

# Future Work

- use automatic, rather than manual, segmentation
- include verbal (words, DAs) features
- explore the dual problem of role/participant detection:

$$\begin{aligned} \mathcal{G}^* &= \arg\max_{\mathcal{G}} P(\mathcal{G} \mid \mathbf{F}) \\ &= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{G}, \mathcal{T}, \mathbf{F}) \\ &= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{T}) \times P(\mathcal{G} \mid \mathcal{T}) \times P(\mathbf{F} \mid \mathcal{G}, \mathcal{T}) \end{aligned}$$

# Future Work

- use automatic, rather than manual, segmentation
- include verbal (words, DAs) features
- explore the dual problem of role/participant detection:

$$\begin{aligned} \mathcal{G}^* &= \arg\max_{\mathcal{G}} P(\mathcal{G} \mid \mathbf{F}) \\ &= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{G}, \mathcal{T}, \mathbf{F}) \\ &= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{T}) \times P(\mathcal{G} \mid \mathcal{T}) \times P(\mathbf{F} \mid \mathcal{G}, \mathcal{T}) \end{aligned}$$

# Future Work

- use automatic, rather than manual, segmentation
- include verbal (words, DAs) features
- explore the dual problem of role/participant detection:

$$
\begin{aligned}
\mathcal{G}^* &= \arg\max_{\mathcal{G}} P(\mathcal{G} \mid \mathbf{F}) \\
&= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{G}, \mathcal{T}, \mathbf{F}) \\
&= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{T}) \times P(\mathcal{G} \mid \mathcal{T}) \times P(\mathbf{F} \mid \mathcal{G}, \mathcal{T})
\end{aligned}
$$

# Thanks!

We'd also like to thank:

- Liz Shriberg
  - lots of helpful discussion
  - access to the ICSI MRDA annotation
- CHIL project for funding