Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings

Kornel Laskowski and Tanja Schultz

Cognitive Systems Lab, Universität Karlsruhe
Language Technologies Institute, Carnegie Mellon University

09 September, 2008

Why Detect Laughter in Meetings?

- evidence suggests that it is the most frequently occurring and most robust behavior which external observers associate with perceived emotion
  - marked valence
  - marked activation

1. automatic emotion recognition in meetings
   - enable indexing, search and summary, mediated by para-propositional content
   - also necessary for autonomous machine participation
2. detection and tracking of humour and seriousness

Why Detect Laughter in Meetings?

- evidence suggests that it is the most frequently occurring and most robust behavior which external observers associate with perceived emotion
  - marked valence
  - marked activation

1. automatic emotion recognition in meetings
   - enable indexing, search and summary, mediated by para-propositional content
   - also necessary for autonomous machine participation
2. detection and tracking of humour and seriousness

Why Detect Laughter in Meetings?

- evidence suggests that it is the most frequently occurring and most robust behavior which external observers associate with perceived emotion
    - marked valence
    - marked activation

1. automatic emotion recognition in meetings
    - enable indexing, search and summary, mediated by para-propositional content
    - also necessary for autonomous machine participation
2. detection and tracking of humour and seriousness

Why Detect Laughter in Meetings?

- evidence suggests that it is the most frequently occurring and most robust behavior which external observers associate with perceived emotion
  - marked valence
  - marked activation

1. automatic emotion recognition in meetings
   - enable indexing, search and summary, mediated by para-propositional content
   - also necessary for autonomous machine participation
2. detection and tracking of humour and seriousness

Why Detect Laughter in Meetings?

- evidence suggests that it is the most frequently occurring and most robust behavior which external observers associate with perceived emotion
  - marked valence
  - marked activation

1. automatic emotion recognition in meetings
   - enable indexing, search and summary, mediated by para-propositional content
   - also necessary for autonomous machine participation
2. detection and tracking of humour and seriousness

Classifying Emotional Valence

- data: ISL Meeting Corpus (Burger et al, 2002)
- annotation: perceived valence (Laskowski & Burger, 2006)
- task: classify **segmented utteraces** as exhibiting one of $\{\texttt{negative}, \texttt{neutral}, \texttt{positive}\}$

Classification	Accuracy, %
	EVAL
guessing with uniform prior	33.3
guessing with TRAINSET prior	≈67
guessing majority TRAINSET class	≈81
presence of $\mathcal{L}$ only	**≈92**
prosody features	≈84
all features (except presence of $\mathcal{L}$)	≈87

Classifying Emotional Activation

- also known as emotional *arousal*
- data: ICSI Meeting Corpus (Janin et al, 2003)
- annotation: hotspots (Wrede & Shriberg, 2004; Wrede et al, 2005)
- task: classify **60-second intervals** as one of $\{\texttt{involvementContaining}, \neg\texttt{involvementContaining}\}$

Classification	Accuracy, %		
	TRAIN	DEV	EVAL
guessing with uniform prior	50.0	50.0	50.0
guessing with TRAINSET prior	61.3	60.9	61.2
guessing majority TRAINSET class	73.7	72.9	73.7
features from $\mathcal{L}$ only	79.2	80.0	**80.6**
features from $\mathcal{L}$ and $\mathcal{S}$	84.3	82.7	**83.0**

Goals of this Work

Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings

1. propose a framework for detecting laughter from audio only
   - close-talk microphones on all participants
2. attempt to detect **all** laughter
   1. temporally isolated from the laugher's speech
   2. occurring within dialog acts among verbal productions
3. attempt to detect **without prior knowledge**
   - no inactive channel exclusion
   - expect to encounter many false alarms
4. attribute laughter to specific participants

Goals of this Work

Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings

1. propose a framework for detecting laughter from audio only
   - close-talk microphones on all participants
2. attempt to detect **all** laughter
   1. temporally isolated from the laugher's speech
   2. occurring within dialog acts among verbal productions
3. attempt to detect **without prior knowledge**
   - no inactive channel exclusion
   - expect to encounter many false alarms
4. attribute laughter to specific participants

Goals of this Work

Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings

1. propose a framework for detecting laughter from audio only
   - close-talk microphones on all participants
2. attempt to detect **all** laughter
   1. temporally isolated from the laugher's speech
   2. occurring within dialog acts among verbal productions
3. attempt to detect **without prior knowledge**
   - no inactive channel exclusion
   - expect to encounter many false alarms
4. attribute laughter to specific participants

Goals of this Work

Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings

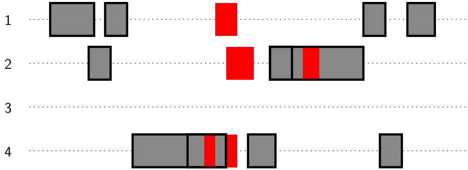
1. propose a framework for detecting laughter from audio only
   - close-talk microphones on all participants
2. attempt to detect **all** laughter
   1. temporally isolated from the laugher's speech
   2. occurring within dialog acts among verbal productions
3. attempt to detect **without prior knowledge**
   - no inactive channel exclusion
   - expect to encounter many false alarms
4. attribute laughter to specific participants

Goals of this Work

Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings

1. propose a framework for detecting laughter from audio only
   - close-talk microphones on all participants
2. attempt to detect **all** laughter
   1. temporally isolated from the laugher's speech
   2. occurring within dialog acts among verbal productions
3. attempt to detect **without prior knowledge**
   - no inactive channel exclusion
   - expect to encounter many false alarms
4. attribute laughter to specific participants

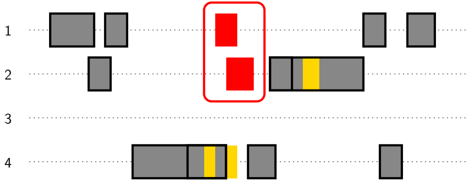
Detecting All Laughter from All Audio



Past work has focused on:

- a subset of laughter (improving recall)
  - isolated laughter
  - loud, clear, unambiguous laughter
- and/or a subset of audio (improving precision)

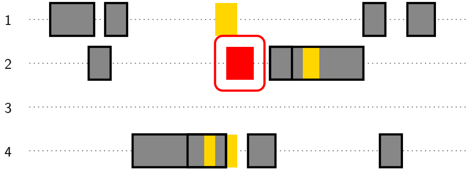
Detecting All Laughter from All Audio



Past work has focused on:

- a subset of laughter (improving recall)
  - isolated laughter
  - loud, clear, unambiguous laughter
- and/or a subset of audio (improving precision)
  - segmented intervals
  - only active channels

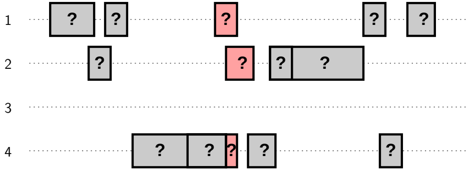
Detecting All Laughter from All Audio



Past work has focused on:

- a subset of laughter (improving recall)
  - isolated laughter
  - loud, clear, unambiguous laughter
- and/or a subset of audio (improving precision)
  - segmented intervals
  - only active channels

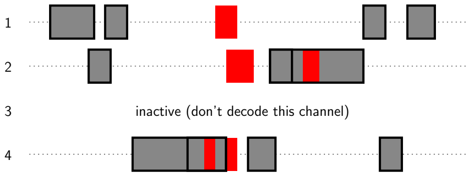
Detecting All Laughter from All Audio



Past work has focused on:

- a subset of laughter (improving recall)
  - isolated laughter
  - loud, clear, unambiguous laughter
- and/or a subset of audio (improving precision)
  - segmented intervals
  - only active channels

Detecting All Laughter from All Audio



Past work has focused on:

- a subset of laughter (improving recall)
  - isolated laughter
  - loud, clear, unambiguous laughter
- and/or a subset of audio (improving precision)
  - segmented intervals
  - only active channels

Brief Comparison with Related Work

Aspect	$\mathcal{L}/\mathcal{S}$ class. [1]	[2]	$\mathcal{L}/\neg\mathcal{L}$ segm. [3]	[4]	[5]	this work
close-talk microphones	✓	✓	✓	✓		✓
farfield microphones					✓	
single channel at-a-time	✓	✓	✓	✓		
multi-channel at-a-time					✓	✓
participant attribution	✓	✓	✓	✓		✓
only group laughter					✓	
only isolated laughter	✓		✓	✓		
only clear laughter		✓				
rely on pre-segmentation	✓	✓	?			
rely on channel exclusion			?	✓		

[1] (Truong & van Leeuwen, 2005); [2] (Truong & van Leeuwen, 2007a); [3] (Truong & van Leeuwen, 2007b); [4] (Knox & Mirghafori, 2007); [5] (Kennedy & Ellis, 2004).

Outline of this Talk

1. Introduction (about to be over)
2. Data
3. Multiparticipant 3-state Vocal Activity Detector
4. Experiments
5. Analysis
6. Conclusions (& Unqualified Recommendations)

ICSI Meeting Corpus

- the complete corpus (Janin et al, 2003)
  - 75 naturally occurring meetings
  - longitudinal CTM recordings of several work groups
  - 3-9 instrumented participants per meeting
- we use a subset of 67 meetings
  - types: `Bed` (15), `Bmr` (29), `Bro` (23)
  - 23 unique participants
  - 3 participants attend both `Bmr` and `Bro`
  - 1 participant attends both `Bmr` and `Bed`
- in particular, as elsewhere,
  - TRAINSET: 26 `Bmr` meetings
  - TESTSET: 3 `Bmr` meetings

ICSI Meeting Corpus

- the complete corpus (Janin et al, 2003)
  - 75 naturally occurring meetings
  - longitudinal CTM recordings of several work groups
  - 3-9 instrumented participants per meeting
- we use a subset of 67 meetings
  - types: `Bed` (15), `Bmr` (29), `Bro` (23)
  - 23 unique participants
  - 3 participants attend both `Bmr` and `Bro`
  - 1 participant attends both `Bmr` and `Bed`
- in particular, as elsewhere,
  - TrainSet: 26 `Bmr` meetings
  - TestSet: 3 `Bmr` meetings

ICSI Meeting Corpus

- the complete corpus (Janin et al, 2003)
  - 75 naturally occurring meetings
  - longitudinal CTM recordings of several work groups
  - 3-9 instrumented participants per meeting
- we use a subset of 67 meetings
  - types: `Bed` (15), `Bmr` (29), `Bro` (23)
  - 23 unique participants
  - 3 participants attend both `Bmr` and `Bro`
  - 1 participant attends both `Bmr` and `Bed`
- in particular, as elsewhere,
  - TRAINSET: 26 `Bmr` meetings
  - TESTSET: 3 `Bmr` meetings

Reference Segmentation

- speech, $\mathcal{S}$
  - forced alignment of words and word fragments
  - available in the ICSI MRDA Corpus (Shriberg et al, 2004)
  - bridge inter-lexeme gaps shorter than 300 ms
  - as in NIST Rich Transcription Meeting Recognition evaluations
- laughter, $\mathcal{L}$
  - produced semi-automatically (Laskowski & Burger, 2007d)
  - $\geq$ 99% of laughter markup, as originally transcribed
  - bouts include terminal "recovery" in-/ehxalation, if present
  - augmented with voicing classification, $\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$
- "laughed speech" (Nwokah et al, 1999), $\mathcal{S} \cap \mathcal{L}$
  - here, mapped to laughter $\mathcal{L}$
  - each participant can be producing $\mathcal{L}$, $\mathcal{S}$, or neither

Reference Segmentation

- speech, $\mathcal{S}$
  - forced alignment of words and word fragments
  - available in the ICSI MRDA Corpus (Shriberg et al, 2004)
  - bridge inter-lexeme gaps shorter than 300 ms
  - as in NIST Rich Transcription Meeting Recognition evaluations
- laughter, $\mathcal{L}$
  - produced semi-automatically (Laskowski & Burger, 2007d)
  - $\geq$ 99% of laughter markup, as originally transcribed
  - bouts include terminal "recovery" in-/ehxalation, if present
  - augmented with voicing classification, $\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$
- "laughed speech" (Nwokah et al, 1999), $\mathcal{S} \cap \mathcal{L}$
  - here, mapped to laughter $\mathcal{L}$
  - each participant can be producing $\mathcal{L}$, $\mathcal{S}$, or neither

Reference Segmentation

- speech, $\mathcal{S}$
  - forced alignment of words and word fragments
  - available in the ICSI MRDA Corpus (Shriberg et al, 2004)
  - bridge inter-lexeme gaps shorter than 300 ms
  - as in NIST Rich Transcription Meeting Recognition evaluations
- laughter, $\mathcal{L}$
  - produced semi-automatically (Laskowski & Burger, 2007d)
  - $\geq$ 99% of laughter markup, as originally transcribed
  - bouts include terminal "recovery" in-/ehxalation, if present
  - augmented with voicing classification, $\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$
- "laughed speech" (Nwokah et al, 1999), $\mathcal{S} \cap \mathcal{L}$
  - here, mapped to laughter $\mathcal{L}$
  - each participant can be producing $\mathcal{L}$, $\mathcal{S}$, or neither

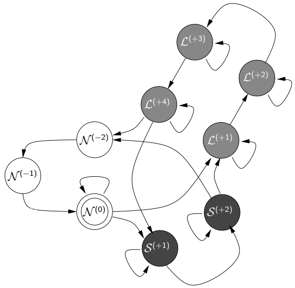
Reference Segmentation

- speech, $\mathcal{S}$
  - forced alignment of words and word fragments
  - available in the ICSI MRDA Corpus (Shriberg et al, 2004)
  - bridge inter-lexeme gaps shorter than 300 ms
  - as in NIST Rich Transcription Meeting Recognition evaluations
- laughter, $\mathcal{L}$
  - produced semi-automatically (Laskowski & Burger, 2007d)
  - $\geq$ 99% of laughter markup, as originally transcribed
  - bouts include terminal "recovery" in-/ehxalation, if present
  - augmented with voicing classification, $\mathcal{L} \equiv \mathcal{L}_V \cup \mathcal{L}_U$
- "laughed speech" (Nwokah et al, 1999), $\mathcal{S} \cap \mathcal{L}$
  - here, mapped to laughter $\mathcal{L}$
  - each participant can be producing $\mathcal{L}$, $\mathcal{S}$, or neither

Multiparticipant 3-state Vocal Activity Detector

- hidden Markov model
- pruned Viterbi (beam) decoding
- topology
  - single participant state subspace
  - multiparticipant state space, pruning
- multiparticipant transition probability model
- standard MFCC features, plus crosstalk suppression features
- multiparticipant emission probability model

Single Participant (SP) State Subspace



- each participant can be
  - speaking, $\mathcal{S}$
  - laughing, $\mathcal{L}$
  - silent, $\mathcal{N}$

- frame step $\Delta T = 0.1$ s
- explicit minimum duration constraints

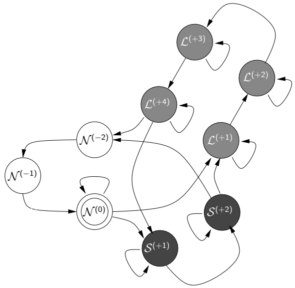
$$\begin{aligned} \mathbf{T}_{min} &= \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min} \right) \\ &= \Delta T \cdot \left(N^{\mathcal{S}}_{min}, N^{\mathcal{L}}_{min}, N^{\mathcal{N}}_{min} \right) \end{aligned}$$

- number of states in 1-participant subspace

$$N = N^{\mathcal{S}}_{min} + N^{\mathcal{L}}_{min} + N^{\mathcal{N}}_{min}$$

- in example shown, $N = 9$

Single Participant (SP) State Subspace



- each participant can be
  - speaking, $\mathcal{S}$
  - laughing, $\mathcal{L}$
  - silent, $\mathcal{N}$

- frame step $\Delta T = 0.1$ s
- explicit minimum duration constraints

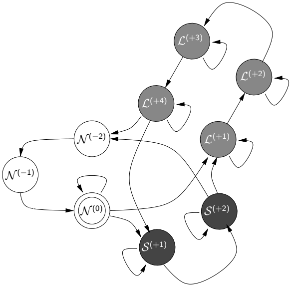
$$\begin{aligned} \mathbf{T}_{min} &= \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min} \right) \\ &= \Delta T \cdot \left(N^{\mathcal{S}}_{min}, N^{\mathcal{L}}_{min}, N^{\mathcal{N}}_{min} \right) \end{aligned}$$

- number of states in 1-participant subspace

$$N = N^{\mathcal{S}}_{min} + N^{\mathcal{L}}_{min} + N^{\mathcal{N}}_{min}$$

- in example shown, $N = 9$

Single Participant (SP) State Subspace



- each participant can be
  - speaking, $\mathcal{S}$
  - laughing, $\mathcal{L}$
  - silent, $\mathcal{N}$

- frame step $\Delta T = 0.1$ s
- explicit minimum duration constraints

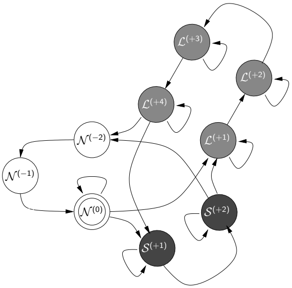
$$
\begin{aligned}
\mathbf{T}_{min} &\equiv \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min}\right) \\
&= \Delta T \cdot \left(N^{\mathcal{S}}_{min}, N^{\mathcal{L}}_{min}, N^{\mathcal{N}}_{min}\right)
\end{aligned}
$$

- number of states in 1-participant subspace

$$
N = N^{\mathcal{S}}_{min} + N^{\mathcal{L}}_{min} + N^{\mathcal{N}}_{min}
$$

- in example shown, $N = 9$

Single Participant (SP) State Subspace



- each participant can be
  - speaking, $\mathcal{S}$
  - laughing, $\mathcal{L}$
  - silent, $\mathcal{N}$

- frame step $\Delta T = 0.1$ s
- explicit minimum duration constraints

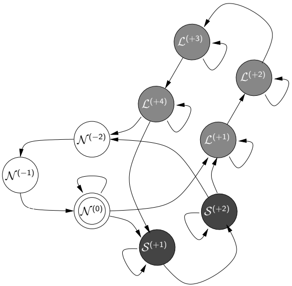
$$\begin{aligned} \mathbf{T}_{min} &\equiv \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min}\right) \\ &= \Delta T \cdot \left(N^{\mathcal{S}}_{min}, N^{\mathcal{L}}_{min}, N^{\mathcal{N}}_{min}\right) \end{aligned}$$

- number of states in 1-participant subspace

$$N = N^{\mathcal{S}}_{min} + N^{\mathcal{L}}_{min} + N^{\mathcal{N}}_{min}$$

- in example shown, $N = 9$

Single Participant (SP) State Subspace



- each participant can be
  - speaking, $\mathcal{S}$
  - laughing, $\mathcal{L}$
  - silent, $\mathcal{N}$

- frame step $\Delta T = 0.1$ s
- explicit minimum duration constraints

$$\begin{aligned} \mathbf{T}_{min} &\equiv \left(T^{\mathcal{S}}_{min}, T^{\mathcal{L}}_{min}, T^{\mathcal{N}}_{min}\right) \\ &= \Delta T \cdot \left(N^{\mathcal{S}}_{min}, N^{\mathcal{L}}_{min}, N^{\mathcal{N}}_{min}\right) \end{aligned}$$

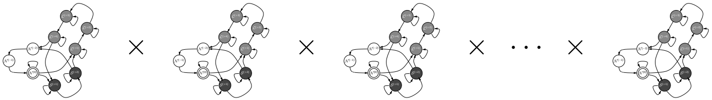
- number of states in 1-participant subspace

$$N = N^{\mathcal{S}}_{min} + N^{\mathcal{L}}_{min} + N^{\mathcal{N}}_{min}$$

- in example shown, $N = 9$

Multiparticipant (MP) State Space

- for a conversation of K participants,
- form the Cartesian product of K factors:

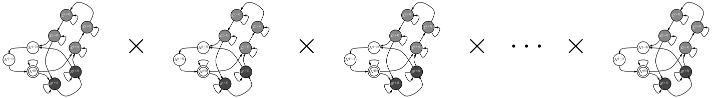


- each MP state: K-vector of N SP states
- total number of MP states in topology: N^K
- impose maximum simultaneous vocalization constraints

$$\mathbf{K}_{max} = \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max}\right)$$

Multiparticipant (MP) State Space

- for a conversation of K participants,
- form the Cartesian product of K factors:



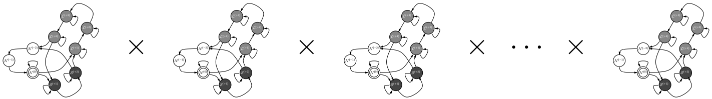
- each MP state: K-vector of N SP states
- total number of MP states in topology: N^K
- impose maximum simultaneous vocalization constraints

$$\mathbf{K}_{max} = \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max}\right)$$

- ie. $K^{\mathcal{L}}_{max}$: max. # participants laughing at the same time

Multiparticipant (MP) State Space

- for a conversation of K participants,
- form the Cartesian product of K factors:



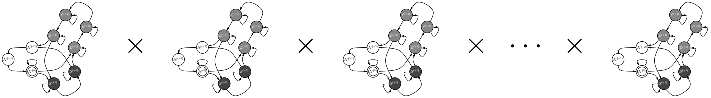
- each MP state: K-vector of N SP states
- total number of MP states in topology: N^K
- impose maximum simultaneous vocalization constraints

$$\mathbf{K}_{max} = \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max}\right)$$

- i.e. $K^{\mathcal{L}}_{max}$: max. # participants laughing at the same time

Multiparticipant (MP) State Space

- for a conversation of K participants,
- form the Cartesian product of K factors:



- each MP state: K-vector of N SP states
- total number of MP states in topology: N^K
- impose maximum simultaneous vocalization constraints

$$\mathbf{K}_{max} \quad = \quad \left(K^{\mathcal{S}}_{max}, K^{\mathcal{L}}_{max}, K^{\neg\mathcal{N}}_{max} \right)$$

  - ie. $K^{\mathcal{L}}_{max}$: max. # participants laughing at the same time

Transition Probability Model

- Example, $K = 5$:
  - at time t, $\mathbf{q}_t = \mathbf{S}_i = \left[\mathcal{S}^{(2)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}\right]$
  - at time $t+1$, $\mathbf{q}_{t+1} = \mathbf{S}_j = \left[\mathcal{N}^{(-2)}, \mathcal{N}^{(0)}, \mathcal{S}^{(1)}, \mathcal{L}^{(1)}\right]$
  - what is $a_{ij} = P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\right)$?

1. $a_{ij} = 0$ if the SP transition from $\mathbf{S}_i$ to $\mathbf{S}_j$ for any participant is not licensed by the SP topology
2. otherwise, ML estimate using ngram counts from best flat-start Viterbi path over training corpus
3. NOTE: each participant's index k in $\mathbf{S}$ is arbitrary
   - for all K-symbol permutations/rotations $\mathbf{R}$
   - want $P\left(\mathbf{S}_j \,|\, \mathbf{S}_i\right) \equiv P\left(\mathbf{R} \cdot \mathbf{S}_j \,|\, \mathbf{R} \cdot \mathbf{S}_i\right)$
   - during model training & querying, rotate each $\mathbf{q}_t$ into a fixed ordering of the N single-participant states

Transition Probability Model

- Example, $K = 5$:
  - at time t, $\mathbf{q}_t = \mathbf{S}_i = \left[\mathcal{S}^{(2)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}\right]$
  - at time $t+1$, $\mathbf{q}_{t+1} = \mathbf{S}_j = \left[\mathcal{N}^{(-2)}, \mathcal{N}^{(0)}, \mathcal{S}^{(1)}, \mathcal{L}^{(1)}\right]$
  - what is $a_{ij} = P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\right)$?

1. $a_{ij} = 0$ if the SP transition from $\mathbf{S}_i$ to $\mathbf{S}_j$ for any participant is not licensed by the SP topology
2. otherwise, ML estimate using ngram counts from best flat-start Viterbi path over training corpus
3. NOTE: each participant's index k in $\mathbf{S}$ is arbitrary
   - for all K-symbol permutations/rotations $\mathbf{R}$
   - want $P\left(\mathbf{S}_j \,|\, \mathbf{S}_i\right) \equiv P\left(\mathbf{R}\cdot\mathbf{S}_j \,|\, \mathbf{R}\cdot\mathbf{S}_i\right)$
   - during model training & querying, rotate each $\mathbf{q}_t$ into a fixed ordering of the N single-participant states

Transition Probability Model

- Example, $K = 5$:
  - at time t, $\mathbf{q}_t = \mathbf{S}_i = \left[\mathcal{S}^{(2)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}\right]$
  - at time $t+1$, $\mathbf{q}_{t+1} = \mathbf{S}_j = \left[\mathcal{N}^{(-2)}, \mathcal{N}^{(0)}, \mathcal{S}^{(1)}, \mathcal{L}^{(1)}\right]$
  - what is $a_{ij} = P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\right)$?

1. $a_{ij} = 0$ if the SP transition from $\mathbf{S}_i$ to $\mathbf{S}_j$ for any participant is not licensed by the SP topology
2. otherwise, ML estimate using ngram counts from best flat-start Viterbi path over training corpus
3. NOTE: each participant's index k in $\mathbf{S}$ is arbitrary
   - for all K-symbol permutations/rotations $\mathbf{R}$
   - want $P\left(\mathbf{S}_j \,|\, \mathbf{S}_i\right) \equiv P\left(\mathbf{R}\cdot\mathbf{S}_j \,|\, \mathbf{R}\cdot\mathbf{S}_i\right)$
   - during model training & querying, rotate each $\mathbf{q}_t$ into a fixed ordering of the N single-participant states

Transition Probability Model

- Example, $K = 5$:
  - at time t, $\mathbf{q}_t = \mathbf{S}_i = \left[\mathcal{S}^{(2)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}, \mathcal{N}^{(0)}\right]$
  - at time $t+1$, $\mathbf{q}_{t+1} = \mathbf{S}_j = \left[\mathcal{N}^{(-2)}, \mathcal{N}^{(0)}, \mathcal{S}^{(1)}, \mathcal{L}^{(1)}\right]$
  - what is $a_{ij} = P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\right)$?

1. $a_{ij} = 0$ if the SP transition from $\mathbf{S}_i$ to $\mathbf{S}_j$ for any participant is not licensed by the SP topology
2. otherwise, ML estimate using ngram counts from best flat-start Viterbi path over training corpus
3. NOTE: each participant's index k in $\mathbf{S}$ is arbitrary
   - for all K-symbol permutations/rotations $\mathbf{R}$
   - want $P\left(\mathbf{S}_j \,|\, \mathbf{S}_i\right) \equiv P\left(\mathbf{R} \cdot \mathbf{S}_j \,|\, \mathbf{R} \cdot \mathbf{S}_i\right)$
   - during model training & querying, rotate each $\mathbf{q}_t$ into a fixed ordering of the N single-participant states

Observables

- each of K participants is wearing a close-talk mic (CTM)
- extract 41 features from every CTM channel
  - log energy + MFCCs
  - Δs and $\Delta\Delta$s
  - min and max normalized log energy differences (NLEDs) (Boakye & Stolcke, 2006)
- $41 \cdot K$ features per frame
- may vary from meeting to meeting (as K does)

Observables

- each of K participants is wearing a close-talk mic (CTM)
- extract 41 features from every CTM channel
  - log energy + MFCCs
  - Δs and $\Delta\Delta$s
  - min and max normalized log energy differences (NLEDs) (Boakye & Stolcke, 2006)
- $41 \cdot K$ features per frame
- may vary from meeting to meeting (as K does)

Observables

- each of K participants is wearing a close-talk mic (CTM)
- extract 41 features from every CTM channel
  - log energy + MFCCs
  - Δs and $\Delta\Delta$s
  - min and max normalized log energy differences (NLEDs) (Boakye & Stolcke, 2006)
- $41 \cdot K$ features per frame
- may vary from meeting to meeting (as K does)

Observables

- each of K participants is wearing a close-talk mic (CTM)
- extract 41 features from every CTM channel
  - log energy + MFCCs
  - Δs and $\Delta\Delta$s
  - min and max normalized log energy differences (NLEDs) (Boakye & Stolcke, 2006)
- $41 \cdot K$ features per frame
- may vary from meeting to meeting (as K does)

Emission Probability Model

- variable feature length vector $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_K]$
- train a single-channel GMM (64 components)
  - for $\mathcal{S}$ and $\mathcal{L}$
  - for $\mathcal{N}_{all}$ and $\mathcal{N}_{nearfield}$
- then approximate the joint MP emission with

$$P(\mathbf{X} \mid \mathbf{S}_i) = \prod_{k=1}^{K} P(\mathbf{X}[k] \mid \mathbf{S}_i[k]) \quad (1)$$

Emission Probability Model

- variable feature length vector $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_K]$
- train a single-channel GMM (64 components)
  - for $\mathcal{S}$ and $\mathcal{L}$
  - for $\mathcal{N}_{all}$ and $\mathcal{N}_{nearfield}$
- then approximate the joint MP emission with

$$P(\mathbf{X} \mid \mathbf{S}_i) = \prod_{k=1}^{K} P(\mathbf{X}[k] \mid \mathbf{S}_i[k]) \tag{1}$$

Emission Probability Model

- variable feature length vector $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_K]$
- train a single-channel GMM (64 components)
  - for $\mathcal{S}$ and $\mathcal{L}$
  - for $\mathcal{N}_{all}$ and $\mathcal{N}_{nearfield}$
- then approximate the joint MP emission with

$$P(\mathbf{X} \mid \mathbf{S}_i) = \prod_{k=1}^{K} P(\mathbf{X}[k] \mid \mathbf{S}_i[k]) \quad (1)$$

Described Experiments

1. independent versus joint participant decoding
2. sensitivity to minimum duration constraints
3. sensitivity to maximum overlap constraints
4. generalization to other (non-Bmr) meetings

Evaluation: recall (R), precision (P), and unweighted F

- goal here: $\mathcal{L}$ versus $\neg\mathcal{L} = \mathcal{S} \cup \mathcal{N}$
- sanity: $\mathcal{S}$ versus $\neg\mathcal{S} = \mathcal{L} \cup \mathcal{N}$
- sanity: $\mathcal{V} = \mathcal{S} \cup \mathcal{L}$ versus $\neg\mathcal{V} = \mathcal{N}$

Described Experiments

1. independent versus joint participant decoding
2. sensitivity to minimum duration constraints
3. sensitivity to maximum overlap constraints
4. generalization to other (non-Bmr) meetings

Evaluation: recall (R), precision (P), and unweighted F

- goal here: $\mathcal{L}$ versus $\neg\mathcal{L} = \mathcal{S} \cup \mathcal{N}$
- sanity: $\mathcal{S}$ versus $\neg\mathcal{S} = \mathcal{L} \cup \mathcal{N}$
- sanity: $\mathcal{V} = \mathcal{S} \cup \mathcal{L}$ versus $\neg\mathcal{V} = \mathcal{N}$

Described Experiments

1. independent versus joint participant decoding
2. sensitivity to minimum duration constraints
3. sensitivity to maximum overlap constraints
4. generalization to other (non-Bmr) meetings

Evaluation: recall (R), precision (P), and unweighted F

- goal here: $\mathcal{L}$ versus $\neg\mathcal{L} = \mathcal{S} \cup \mathcal{N}$
- sanity: $\mathcal{S}$ versus $\neg\mathcal{S} = \mathcal{L} \cup \mathcal{N}$
- sanity: $\mathcal{V} = \mathcal{S} \cup \mathcal{L}$ versus $\neg\mathcal{V} = \mathcal{N}$

Single-participant vs Multiparticipant Decoding

- for decoding participants independently
  - $\mathcal{N}_{all}$ and $\mathcal{N}_{farfield}$ both represent nearfield silence $\mathcal{N}$
  - $\rightarrow$ 3 competing models, rather than 4
- for decoding participant jointly, can use either 3 or 4 models

Decoding	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
indep, 3 AM	76.3	90.3	85.0	87.6	80.9	20.4	32.6
joint, 3 AM	78.8	89.7	86.0	87.8	59.2	20.6	30.6
▶ joint, 4 AM	**79.5**	83.6	90.0	86.7	55.2	**25.1**	**34.5**

1. joint decoding improves precision by reducing potential overlap
2. modeling farfield vocalization on CTMs significantly improves precision for $\mathcal{S}$ and $\mathcal{L}$

Single-participant vs Multiparticipant Decoding

- for decoding participants independently
  - $\mathcal{N}_{all}$ and $\mathcal{N}_{farfield}$ both represent nearfield silence $\mathcal{N}$
  - $\rightarrow$ 3 competing models, rather than 4
- for decoding participant jointly, can use either 3 or 4 models

Decoding	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
indep, 3 AM	76.3	90.3	85.0	87.6	80.9	20.4	32.6
joint, 3 AM	78.8	89.7	86.0	87.8	59.2	20.6	30.6
▶ joint, 4 AM	**79.5**	83.6	90.0	86.7	55.2	**25.1**	**34.5**

1. joint decoding improves precision by reducing potential overlap
2. modeling farfield vocalization on CTMs significantly improves precision for $\mathcal{S}$ and $\mathcal{L}$

Single-participant vs Multiparticipant Decoding

- for decoding participants independently
  - $\mathcal{N}_{all}$ and $\mathcal{N}_{farfield}$ both represent nearfield silence $\mathcal{N}$
  - $\rightarrow$ 3 competing models, rather than 4
- for decoding participant jointly, can use either 3 or 4 models

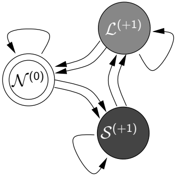
Decoding	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
indep, 3 AM	76.3	90.3	85.0	87.6	80.9	20.4	32.6
joint, 3 AM	78.8	89.7	86.0	87.8	59.2	20.6	30.6
▶ joint, 4 AM	**79.5**	83.6	90.0	86.7	55.2	**25.1**	**34.5**

1. joint decoding improves precision by reducing potential overlap
2. modeling farfield vocalization on CTMs significantly improves precision for $\mathcal{S}$ and $\mathcal{L}$

Alternative Minimum Duration Constraints $\mathbf{T}_{min}$

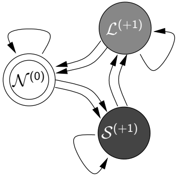
$\mathbf{T}_{min} = (0.1, 0.1, 0.1)$ $\mathbf{T}_{min} = (0.3, 0.3, 0.3)$ $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

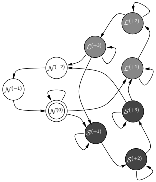
Alternative Minimum Duration Constraints $\mathbf{T}_{min}$



$\mathbf{T}_{min} = (0.1, 0.1, 0.1)$ $\quad$ $\mathbf{T}_{min} = (0.3, 0.3, 0.3)$ $\quad$ $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

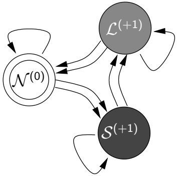
Alternative Minimum Duration Constraints $\mathbf{T}_{min}$



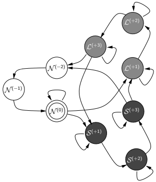


$\mathbf{T}_{min} = (0.1, 0.1, 0.1)$ $\mathbf{T}_{min} = (0.3, 0.3, 0.3)$ $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

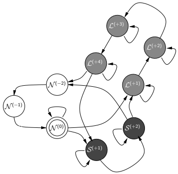
Alternative Minimum Duration Constraints $\mathbf{T}_{min}$



$\mathbf{T}_{min} = (0.1, 0.1, 0.1)$



$\mathbf{T}_{min} = (0.3, 0.3, 0.3)$



$\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

Alternative Minimum Duration Constraints $\mathbf{T}_{min}$

- hold maximum overlap constraints fixed, $\mathbf{K}_{max} = (2, 3, 3)$

$\mathbf{T}_{min}$ (s)	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
$(0.1, 0.1, 0.1)$	78.1	82.3	89.9	86.0	**55.9**	22.1	31.7
$(0.3, 0.3, 0.3)$	**79.5**	**83.7**	**90.4**	**86.9**	54.7	24.2	33.6
▶ $(0.2, 0.4, 0.3)$	**79.5**	83.6	90.0	86.7	55.2	**25.1**	**34.5**

1. increasing all T_{min} from 0.1s to 0.3s improves all F measures
2. allowing $T^{\mathcal{L}}_{min} > T^{\mathcal{S}}_{min}$ can result in higher F $(\mathcal{L})$

Alternative Minimum Duration Constraints $\mathbf{T}_{min}$

- hold maximum overlap constraints fixed, $\mathbf{K}_{max} = (2, 3, 3)$

$\mathbf{T}_{min}$ (s)	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
$(0.1, 0.1, 0.1)$	78.1	82.3	89.9	86.0	**55.9**	22.1	31.7
$(0.3, 0.3, 0.3)$	**79.5**	**83.7**	**90.4**	**86.9**	54.7	24.2	33.6
▶ $(0.2, 0.4, 0.3)$	**79.5**	83.6	90.0	86.7	55.2	**25.1**	**34.5**

1. increasing all T_{min} from 0.1s to 0.3s improves all F measures
2. allowing $T^{\mathcal{L}}_{min} > T^{\mathcal{S}}_{min}$ can result in higher $F(\mathcal{L})$

Alternative Minimum Duration Constraints $\mathbf{T}_{min}$

- hold maximum overlap constraints fixed, $\mathbf{K}_{max} = (2, 3, 3)$

$\mathbf{T}_{min}$ (s)	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
$(0.1, 0.1, 0.1)$	78.1	82.3	89.9	86.0	**55.9**	22.1	31.7
$(0.3, 0.3, 0.3)$	**79.5**	**83.7**	**90.4**	**86.9**	54.7	24.2	33.6
▶ $(0.2, 0.4, 0.3)$	**79.5**	83.6	90.0	86.7	55.2	**25.1**	**34.5**

1. increasing all T_{min} from 0.1s to 0.3s improves all F measures
2. allowing $T^{\mathcal{L}}_{min} > T^{\mathcal{S}}_{min}$ can result in higher $F(\mathcal{L})$

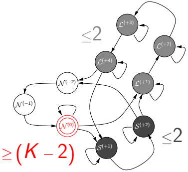
Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$

$\mathbf{K}_{max} = (3, 2, 3)$

$\mathbf{K}_{max} = (2, 2, 2)$ $\mathbf{K}_{max} = (2, 2, 3)$

$\mathbf{K}_{max} = (2, 3, 3)$

Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$



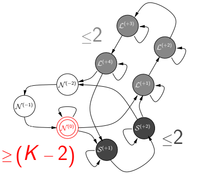
$\mathbf{K}_{max} = (2, 2, 2)$

$\mathbf{K}_{max} = (3, 2, 3)$

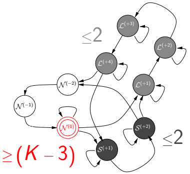
$\mathbf{K}_{max} = (2, 2, 3)$

$\mathbf{K}_{max} = (2, 3, 3)$

Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$



$\mathbf{K}_{max} = (2, 2, 2)$

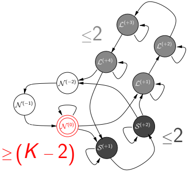


$\mathbf{K}_{max} = (2, 2, 3)$

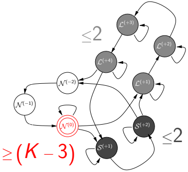
$\mathbf{K}_{max} = (3, 2, 3)$

$\mathbf{K}_{max} = (2, 3, 3)$

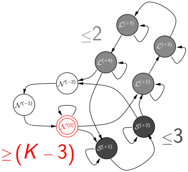
Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$



$\mathbf{K}_{max} = (2, 2, 2)$



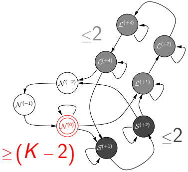
$\mathbf{K}_{max} = (2, 2, 3)$



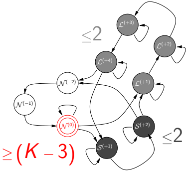
$\mathbf{K}_{max} = (3, 2, 3)$

$\mathbf{K}_{max} = (2, 3, 3)$

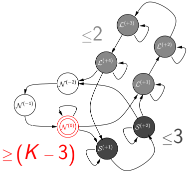
Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$



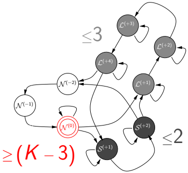
$$\mathbf{K}_{max} = (2, 2, 2)$$



$$\mathbf{K}_{max} = (2, 2, 3)$$



$$\mathbf{K}_{max} = (3, 2, 3)$$



$$\mathbf{K}_{max} = (2, 3, 3)$$

Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$

- minimum duration constraints fixed, $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

$\mathbf{K}_{max}$	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
(2, 2, 2)	**81.3**	83.3	**90.6**	**86.8**	36.9	**27.8**	31.7
(2, 2, 3)	79.9	84.0	89.0	86.4	48.8	24.3	32.4
(3, 2, 3)	79.9	**84.2**	88.6	86.4	49.1	24.6	32.8
► (2, 3, 3)	79.5	83.6	90.0	86.7	**55.2**	25.1	**34.5**

1. increasing K_{max} generally leads to higher R and lower P
2. increasing $K^{\mathcal{S}}_{max}$ from 2 to 3 has negligible impact
3. increasing $K^{\mathcal{L}}_{max}$ from 2 to 3 has significant impact

★ because a higher proportion of $\mathcal{L}$ is produced in overlap

Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$

- minimum duration constraints fixed, $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

$\mathbf{K}_{max}$	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
(2, 2, 2)	**81.3**	83.3	**90.6**	**86.8**	36.9	**27.8**	31.7
(2, 2, 3)	79.9	84.0	89.0	86.4	48.8	24.3	32.4
(3, 2, 3)	79.9	**84.2**	88.6	86.4	49.1	24.6	32.8
▶ (2, 3, 3)	79.5	83.6	90.0	86.7	**55.2**	25.1	**34.5**

1. increasing K_{max} generally leads to higher R and lower P
2. increasing $K^{\mathcal{S}}_{max}$ from 2 to 3 has negligible impact
3. increasing $K^{\mathcal{L}}_{max}$ from 2 to 3 has significant impact

★ because a higher proportion of $\mathcal{L}$ is produced in overlap

Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$

- minimum duration constraints fixed, $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

$\mathbf{K}_{max}$	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
$(2, 2, 2)$	**81.3**	83.3	**90.6**	**86.8**	36.9	**27.8**	31.7
$(2, 2, 3)$	79.9	84.0	89.0	86.4	48.8	24.3	32.4
$(3, 2, 3)$	79.9	**84.2**	88.6	86.4	49.1	24.6	32.8
▶ $(2, 3, 3)$	79.5	83.6	90.0	86.7	**55.2**	25.1	**34.5**

1. increasing K_{max} generally leads to higher R and lower P
2. increasing $K_{max}^{\mathcal{S}}$ from 2 to 3 has negligible impact
3. increasing $K_{max}^{\mathcal{L}}$ from 2 to 3 has significant impact

★ because a higher proportion of $\mathcal{L}$ is produced in overlap

Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$

- minimum duration constraints fixed, $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

$\mathbf{K}_{max}$	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
$(2, 2, 2)$	**81.3**	83.3	**90.6**	**86.8**	36.9	**27.8**	31.7
$(2, 2, 3)$	79.9	84.0	89.0	86.4	48.8	24.3	32.4
$(3, 2, 3)$	79.9	**84.2**	88.6	86.4	49.1	24.6	32.8
► $(2, 3, 3)$	79.5	83.6	90.0	86.7	**55.2**	25.1	**34.5**

1. increasing K_{max} generally leads to higher R and lower P
2. increasing $K^{\mathcal{S}}_{max}$ from 2 to 3 has negligible impact
3. increasing $K^{\mathcal{L}}_{max}$ from 2 to 3 has significant impact

★ because a higher proportion of $\mathcal{L}$ is produced in overlap

Alternative Maximum Overlap Constraints $\mathbf{K}_{max}$

- minimum duration constraints fixed, $\mathbf{T}_{min} = (0.2, 0.4, 0.3)$

$\mathbf{K}_{max}$	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
	F	R	P	F	R	P	F
$(2,2,2)$	**81.3**	83.3	**90.6**	**86.8**	36.9	**27.8**	31.7
$(2,2,3)$	79.9	84.0	89.0	86.4	48.8	24.3	32.4
$(3,2,3)$	79.9	**84.2**	88.6	86.4	49.1	24.6	32.8
► $(2,3,3)$	79.5	83.6	90.0	86.7	**55.2**	25.1	**34.5**

1. increasing K_{max} generally leads to higher R and lower P
2. increasing $K^{\mathcal{S}}_{max}$ from 2 to 3 has negligible impact
3. increasing $K^{\mathcal{L}}_{max}$ from 2 to 3 has significant impact

★ because a higher proportion of $\mathcal{L}$ is produced in overlap

Generalization to Other Meetings

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
			F	R	P	F	R	P	F
Bmr	train	10.91	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
	test	14.94	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
Bro	(all)	5.94	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
Bed	(all)	7.53	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

1. F ($\mathcal{V}$): Bmr(train) > Bmr(test) > Bro > Bed
   - Bmr(train) and Bmr(test) have lots of participants in common
2. F ($\mathcal{L}$) on Bmr(test) higher than on Bmr(train)

Generalization to Other Meetings

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V}$ F	$\mathcal{S}$ R	$\mathcal{S}$ P	$\mathcal{S}$ F	$\mathcal{L}$ R	$\mathcal{L}$ P	$\mathcal{L}$ F
Bmr	train	10.91	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
	test	14.94	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
Bro	(all)	5.94	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
Bed	(all)	7.53	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

1. F ($\mathcal{V}$): Bmr(train) > Bmr(test) > Bro > Bed
   - Bmr(train) and Bmr(test) have lots of participants in common
   - with Bmr, Bro shares 3 participants, and Bed 1 participant
2. F ($\mathcal{L}$) on Bmr(test) higher than on Bmr(train)

Generalization to Other Meetings

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V}$	$\mathcal{S}$			$\mathcal{L}$		
			F	R	P	F	R	P	F
Bmr	train	10.91	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
	test	14.94	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
Bro	(all)	5.94	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
Bed	(all)	7.53	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

1. F ($\mathcal{V}$): Bmr(train) > Bmr(test) > Bro > Bed
   - Bmr(train) and Bmr(test) have lots of participants in common
   - with Bmr, Bro shares 3 participants, and Bed 1 participant
2. F ($\mathcal{L}$) on Bmr(test) higher than on Bmr(train)

Generalization to Other Meetings

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V}$ F	$\mathcal{S}$ R	$\mathcal{S}$ P	$\mathcal{S}$ F	$\mathcal{L}$ R	$\mathcal{L}$ P	$\mathcal{L}$ F
Bmr	train	10.91	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
Bmr	test	14.94	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
Bro	(all)	5.94	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
Bed	(all)	7.53	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

1. F $(\mathcal{V})$: Bmr(train) > Bmr(test) > Bro > Bed
   - Bmr(train) and Bmr(test) have lots of participants in common
   - with Bmr, Bro shares 3 participants, and Bed 1 participant
2. F $(\mathcal{L})$ on Bmr(test) higher than on Bmr(train)
   - appears to correlate with $p_{\mathcal{V}}(\mathcal{L})$, the proportion of vocalization effort spent on laughter
   - this test set is not *typical* of the corpus

Generalization to Other Meetings

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V}$ F	$\mathcal{S}$ R	$\mathcal{S}$ P	$\mathcal{S}$ F	$\mathcal{L}$ R	$\mathcal{L}$ P	$\mathcal{L}$ F
Bmr	train	10.91	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
	test	14.94	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
Bro	(all)	5.94	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
Bed	(all)	7.53	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

1. F ($\mathcal{V}$): Bmr(train) > Bmr(test) > Bro > Bed
   - Bmr(train) and Bmr(test) have lots of participants in common
   - with Bmr, Bro shares 3 participants, and Bed 1 participant
2. F ($\mathcal{L}$) on Bmr(test) higher than on Bmr(train)
   - appears to correlate with $p_{\mathcal{V}}(\mathcal{L})$, the proportion of vocalization effort spent on laughter
   - this test set is *not typical* of the corpus

Generalization to Other Meetings

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V}$ F	$\mathcal{S}$ R	$\mathcal{S}$ P	$\mathcal{S}$ F	$\mathcal{L}$ R	$\mathcal{L}$ P	$\mathcal{L}$ F
`Bmr`	train	10.91	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
	test	14.94	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
`Bro`	(all)	5.94	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
`Bed`	(all)	7.53	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

1. F ($\mathcal{V}$): `Bmr`(train) > `Bmr`(test) > `Bro` > `Bed`
   - `Bmr`(train) and `Bmr`(test) have lots of participants in common
   - with `Bmr`, `Bro` shares 3 participants, and `Bed` 1 participant
2. F ($\mathcal{L}$) on `Bmr`(test) higher than on `Bmr`(train)
   - appears to correlate with $p_{\mathcal{V}}(\mathcal{L})$, the proportion of vocalization effort spent on laughter
   - this test set is *not typical* of the corpus

Generalization to Other Meetings

Test data		$p_{\mathcal{V}}(\mathcal{L})$	$\mathcal{V}$ F	$\mathcal{S}$ R	$\mathcal{S}$ P	$\mathcal{S}$ F	$\mathcal{L}$ R	$\mathcal{L}$ P	$\mathcal{L}$ F
Bmr	train	10.91	**80.1**	83.4	89.8	86.5	53.0	19.4	28.4
	test	14.94	79.5	83.6	90.0	**86.7**	55.2	**25.1**	**34.5**
Bro	(all)	5.94	78.1	81.1	**90.6**	85.6	57.8	11.4	19.0
Bed	(all)	7.53	75.1	**84.6**	85.7	85.2	**58.7**	10.0	17.0

1. F ($\mathcal{V}$): Bmr(train) > Bmr(test) > Bro > Bed
   - Bmr(train) and Bmr(test) have lots of participants in common
   - with Bmr, Bro shares 3 participants, and Bed 1 participant
2. F ($\mathcal{L}$) on Bmr(test) higher than on Bmr(train)
   - appears to correlate with $p_{\mathcal{V}}(\mathcal{L})$, the proportion of vocalization effort spent on laughter
   - this test set is *not typical* of the corpus

Confusion Matrix Analysis

	hypothesized as $\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}$	6.5	**9.1**	1.0	10.4
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

- final system on test set (13.8 hours)
- all quantities in minutes

Confusion Matrix Analysis

	hypothesized as $\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}$	6.5	**9.1**	1.0	16.6
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

- break down references $\mathcal{L} \equiv \{ \mathcal{L}'_U, \mathcal{L}'_V, \mathcal{L} \cap \mathcal{S} \}$
  - $\mathcal{L}'_U \equiv \mathcal{L}_U - \mathcal{L} \cap \mathcal{S}$: unvoiced laughter less "laughed speech"
  - $\mathcal{L}'_V \equiv \mathcal{L}_V - \mathcal{L} \cap \mathcal{S}$: voiced laughter less "laughed speech"

Confusion Matrix Analysis

	hypothesized as $\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}$	6.5	**9.1**	1.0	16.6
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

- break down references $\mathcal{L} \equiv \{ \mathcal{L}'_U, \mathcal{L}'_V, \mathcal{L} \cap \mathcal{S} \}$
  - $\mathcal{L}'_U \equiv \mathcal{L}_U - \mathcal{L} \cap \mathcal{S}$: unvoiced laughter less “laughed speech”
  - $\mathcal{L}'_V \equiv \mathcal{L}_V - \mathcal{L} \cap \mathcal{S}$: voiced laughter less “laughed speech”

Confusion Matrix Analysis

	hypothesized as $\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}$	6.5	**9.1**	1.0	16.6
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

- break down references $\mathcal{L} \equiv \{ \mathcal{L}'_U, \mathcal{L}'_V, \mathcal{L} \cap \mathcal{S} \}$
  - $\mathcal{L}'_U \equiv \mathcal{L}_U - \mathcal{L} \cap \mathcal{S}$: unvoiced laughter less "laughed speech"
  - $\mathcal{L}'_V \equiv \mathcal{L}_V - \mathcal{L} \cap \mathcal{S}$: voiced laughter less "laughed speech"

Confusion Matrix Analysis

	hypothesized as $\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	**2.8**	2.4	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	0.2	**0.5**	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

1. most unvoiced laughter ($\mathcal{L}'_U$) is classified as silence ($\mathcal{N}$)
2. most "laughed speech" ($\mathcal{L} \cap \mathcal{S}$) is classified as speech ($\mathcal{S}$)

Confusion Matrix Analysis

	hypothesized as $\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	**2.8**	2.4	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	0.2	**0.5**	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

1. most unvoiced laughter ($\mathcal{L}'_U$) is classified as silence ($\mathcal{N}$)
2. most "laughed speech" ($\mathcal{L} \cap \mathcal{S}$) is classified as speech ($\mathcal{S}$)

Confusion Matrix Analysis

	hypothesized as $\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	**2.8**	2.4	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	0.2	**0.5**	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

1. most unvoiced laughter ($\mathcal{L}'_U$) is classified as silence ($\mathcal{N}$)
2. most "laughed speech" ($\mathcal{L} \cap \mathcal{S}$) is classified as speech ($\mathcal{S}$)

Confusions Between $\mathcal{L}$ and $\mathcal{S}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

	$\mathcal{L}$	$\mathcal{S}$
$\mathcal{L}'$	8.9	0.5
$\mathcal{L} \cap \mathcal{S}$	0.2	0.5
$\mathcal{S}$	4.5	79.0

- looking at $\mathcal{L}$ and $\mathcal{S}$ only

Confusions Between $\mathcal{L}$ and $\mathcal{S}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Recall:

	$\mathcal{L}$	$\mathcal{S}$
$\mathcal{L}'$	**94.7**	5.3
$\mathcal{L} \cap \mathcal{S}$	28.6	71.4
$\mathcal{S}$	5.4	**93.6**

1. 94% of speech is hypothesized as speech
2. 95% of laughter (excluding "laughed speech") is hypothesized as laughter

Confusions Between $\mathcal{L}$ and $\mathcal{S}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Recall:

	$\mathcal{L}$	$\mathcal{S}$
$\mathcal{L}'$	**94.7**	5.3
$\mathcal{L} \cap \mathcal{S}$	28.6	71.4
$\mathcal{S}$	5.4	**93.6**

1. 94% of speech is hypothesized as speech
2. 95% of laughter (excluding "laughed speech") is hypothesized as laughter

Confusions Between $\mathcal{L}$ and $\mathcal{S}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Recall:

	$\mathcal{L}$	$\mathcal{S}$
$\mathcal{L}'$	**94.7**	5.3
$\mathcal{L} \cap \mathcal{S}$	28.6	71.4
$\mathcal{S}$	5.4	**93.6**

1. 94% of speech is hypothesized as speech
2. 95% of laughter (excluding "laughed speech") is hypothesized as laughter

Confusions Between $\mathcal{L}$ and $\mathcal{S}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Precision:

	$\mathcal{L}$	$\mathcal{S}$
$\mathcal{L}'$	**65.4**	0.6
$\mathcal{L} \cap \mathcal{S}$	1.5	0.6
$\mathcal{S}$	33.1	**98.8**

1. 99% of hypothesized speech is speech
2. 65% of hypothesized laughter is laughter

Confusions Between $\mathcal{L}$ and $\mathcal{S}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Precision:

	$\mathcal{L}$	$\mathcal{S}$
$\mathcal{L}'$	**65.4**	0.6
$\mathcal{L} \cap \mathcal{S}$	1.5	0.6
$\mathcal{S}$	33.1	**98.8**

1. 99% of hypothesized speech is speech
2. 65% of hypothesized laughter is laughter

Confusions Between $\mathcal{L}$ and $\mathcal{S}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Precision:

	$\mathcal{L}$	$\mathcal{S}$
$\mathcal{L}'$	**65.4**	0.6
$\mathcal{L} \cap \mathcal{S}$	1.5	0.6
$\mathcal{S}$	33.1	**98.8**

1. 99% of hypothesized speech is speech
2. 65% of hypothesized laughter is laughter

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	685.4	22.9
$\mathcal{L}'_U$	2.8	2.4
$\mathcal{L}_V$	3.7	6.7

- looking at $\mathcal{L}$ and $\mathcal{N}$ only

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Recall:

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	**96.8**	3.2
$\mathcal{L}'_U$	53.9	46.2
$\mathcal{L}_V$	35.6	**64.4**

1. 97% of silence is hypothesized as silence
2. 64% of voiced laughter (including "laughed speech") is classified as laughter

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Recall:

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	**96.8**	3.2
$\mathcal{L}'_U$	53.9	46.2
$\mathcal{L}_V$	35.6	**64.4**

1. 97% of silence is hypothesized as silence
2. 64% of voiced laughter (including "laughed speech") is classified as laughter

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Recall:

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	**96.8**	3.2
$\mathcal{L}'_U$	53.9	46.2
$\mathcal{L}_V$	35.6	**64.4**

1. 97% of silence is hypothesized as silence
2. 64% of voiced laughter (including "laughed speech") is classified as laughter

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Precision:

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	**99.1**	71.6
$\mathcal{L}'_U$	0.4	7.5
$\mathcal{L}_V$	0.5	20.9

1. 99% of hypothesized silence is silence
2. 28% of hypothesized laughter is laughter
3. 72% of hypothesized laughter is silence

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Precision:

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	**99.1**	71.6
$\mathcal{L}'_U$	0.4	7.5
$\mathcal{L}_V$	0.5	20.9

1. 99% of hypothesized silence is silence
2. 28% of hypothesized laughter is laughter
3. 72% of hypothesized laughter is silence

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Precision:

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	**99.1**	71.6
$\mathcal{L}'_U$	0.4	7.5
$\mathcal{L}_V$	0.5	20.9

1. 99% of hypothesized silence is silence
2. 28% of hypothesized laughter is laughter
3. 72% of hypothesized laughter is silence

Confusions Between $\mathcal{L}$ and $\mathcal{N}$

	$\mathcal{N}$	$\mathcal{L}$	$\mathcal{S}$	Σ
$\mathcal{N}$	**685.4**	22.9	7.8	716.2
$\mathcal{L}'_U$	2.8	**2.4**	0.2	5.4
$\mathcal{L}'_V$	3.6	**6.5**	0.3	10.4
$\mathcal{L} \cap \mathcal{S}$	0.1	**0.2**	0.5	0.8
$\mathcal{S}$	11.0	4.5	**79.0**	94.4
Σ	702.9	36.6	87.8	827.2

Precision:

	$\mathcal{N}$	$\mathcal{L}$
$\mathcal{N}$	**99.1**	71.6
$\mathcal{L}'_U$	0.4	7.5
$\mathcal{L}_V$	0.5	20.9

1. 99% of hypothesized silence is silence
2. 28% of hypothesized laughter is laughter
3. 72% of hypothesized laughter is silence

Conclusions

1. baseline system for multiparticipant 3-way VAD
   - no pre-segmentation assumed
   - all laughter considered
2. { laughter vs silence } harder than { laughter vs speech }
3. speech/laughter contrastive constraints helpful
   - maximum allowed degree of overlap
   - minimum state duration
4. current performance is a function of
   1. proportion of laughter present
   2. participant novelty

Conclusions

1. baseline system for multiparticipant 3-way VAD
   - no pre-segmentation assumed
   - all laughter considered
2. { laughter vs silence } harder than { laughter vs speech }
3. speech/laughter contrastive constraints helpful
   - maximum allowed degree of overlap
   - minimum state duration
4. current performance is a function of
   1. proportion of laughter present
   2. participant novelty

Conclusions

1. baseline system for multiparticipant 3-way VAD
   - no pre-segmentation assumed
   - all laughter considered
2. { laughter vs silence } harder than { laughter vs speech }
3. speech/laughter contrastive constraints helpful
   - maximum allowed degree of overlap
   - minimum state duration
4. current performance is a function of
   1. proportion of laughter present
   2. participant novelty

Conclusions

1. baseline system for multiparticipant 3-way VAD
   - no pre-segmentation assumed
   - all laughter considered
2. { laughter vs silence } harder than { laughter vs speech }
3. speech/laughter contrastive constraints helpful
   - maximum allowed degree of overlap
   - minimum state duration
4. current performance is a function of
   1. proportion of laughter present
   2. participant novelty

Conclusions

1. baseline system for multiparticipant 3-way VAD
    - no pre-segmentation assumed
    - all laughter considered
2. { laughter vs silence } harder than { laughter vs speech }
3. speech/laughter contrastive constraints helpful
    - maximum allowed degree of overlap
    - minimum state duration
4. current performance is a function of
    1. proportion of laughter present
    2. participant novelty

Possible Future Work

1. model voiced and unvoiced laughter ($\mathcal{L}_V$ and $\mathcal{L}_U$) separately
   - different acoustics
   - different overlap contexts (Laskowski & Burger, 2007c)
   - different semantics
2. characterize laughter by instrumentality to high level tasks
   - which laughter signals different emotional valence
   - which laughter signals involvement hotspots
   - Q: Does instrumentality correspond to how clear and unambiguous laughter is? cf. (Truong & van Leeuwen, 2007a)
3. multi-pass, multi-resolution laughter detection
   - pass 1: large frame size (0.1s), small context (0.1s)
   - pass 2: small frame size (0.01s), large context (1.0s) (Knox & Mirghafori, 2007)

Possible Future Work

1. model voiced and unvoiced laughter ($\mathcal{L}_V$ and $\mathcal{L}_U$) separately
   - different acoustics
   - different overlap contexts (Laskowski & Burger, 2007c)
   - different semantics
2. characterize laughter by instrumentality to high level tasks
   - which laughter signals different emotional valence
   - which laughter signals involvement hotspots
   - Q: Does instrumentality correspond to how clear and unambiguous laughter is? cf. (Truong & van Leeuwen, 2007a)
3. multi-pass, multi-resolution laughter detection
   - pass 1: large frame size (0.1s), small context (0.1s)
   - pass 2: small frame size (0.01s), large context (1.0s) (Knox & Mirghafori, 2007)

Possible Future Work

1. model voiced and unvoiced laughter ($\mathcal{L}_V$ and $\mathcal{L}_U$) separately
   - different acoustics
   - different overlap contexts (Laskowski & Burger, 2007c)
   - different semantics
2. characterize laughter by instrumentality to high level tasks
   - which laughter signals different emotional valence
   - which laughter signals involvement hotspots
   - Q: Does instrumentality correspond to how clear and unambiguous laughter is? cf. (Truong & van Leeuwen, 2007a)
3. multi-pass, multi-resolution laughter detection
   - pass 1: large frame size (0.1s), small context (0.1s)
   - pass 2: small frame size (0.01s), large context (1.0s) (Knox & Mirghafori, 2007)

Possible Future Work

1. model voiced and unvoiced laughter ($\mathcal{L}_V$ and $\mathcal{L}_U$) separately
   - different acoustics
   - different overlap contexts (Laskowski & Burger, 2007c)
   - different semantics
2. characterize laughter by instrumentality to high level tasks
   - which laughter signals different emotional valence
   - which laughter signals involvement hotspots
   - Q: Does instrumentality correspond to how clear and unambiguous laughter is? cf. (Truong & van Leeuwen, 2007a)
3. multi-pass, multi-resolution laughter detection
   - pass 1: large frame size (0.1s), small context (0.1s)
   - pass 2: small frame size (0.01s), large context (1.0s) (Knox & Mirghafori, 2007)

Thanks for attending ...

Also, thanks to

- Susi Burger, help with $\mathcal{L}$ segmentation & classification
- Liz Shriberg, access to ICSI MRDA Corpus
- Khiet Truong and Mary Knox, discussion of own work