| Introduction | Segmentation | Recognition | Data o | Experiments 0000000000 | Conclusions O |
|--------------|--------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| | | | | | |

Simultaneous Multispeaker Segmentation for Automatic Meeting Recognition

Kornel Laskowski^{1,2}, Christian Fügen² & Tanja Schultz^{1,2}

¹interACT, Carnegie Mellon University, USA ²interACT, Universität Karlsruhe, Germany

September 6, 2007

| Introduction ●○○○ | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Goal | | | | | |



Why?

- syntactic continuity \rightarrow can leverage language modeling
- speaker homogeneity → can leverage speaker adaptation

| Introduction ●○○○ | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Goal | | | | | |



Why?

- syntactic continuity \rightarrow can leverage language modeling
- speaker homogeneity \rightarrow can leverage speaker adaptation

・同・・モー・ ・モー

| Introduction ●○○○ | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Goal | | | | | |



Why?

- \bullet syntactic continuity \rightarrow can leverage language modeling
- speaker homogeneity → can leverage speaker adaptation

A (B) + A (B) + A (B) +

| Introduction ●○○○ | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Goal | | | | | |



Why?

- \bullet syntactic continuity \rightarrow can leverage language modeling
- $\bullet\,$ speaker homogeneity $\rightarrow\,$ can leverage speaker adaptation

| Introduction ○●○○ | Segmentation | Recognition 00 | Data ⊙ | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| Main Pro | blem: Cros | stalk | | | |



participants share an acoustic space

- everybody's speech appears on everybody's microphone
- leads to high insertion rates (Pfau et al, ASRU01)

| Introduction ○●○○ | Segmentation | Recognition 00 | Data ⊙ | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| Main Pro | blem: Cros | stalk | | | |



• participants share an acoustic space

- everybody's speech appears on everybody's microphone
- leads to high insertion rates (Pfau et al, ASRU01)

| Introduction ○●○○ | Segmentation | Recognition 00 | Data ⊙ | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| Main Pro | blem: Cross | stalk | | | |



- participants share an acoustic space
- everybody's speech appears on everybody's microphone
- leads to high insertion rates (Pfau et al, ASRU01)

| Introduction ○●○○ | Segmentation | Recognition | Data o | Experiments 00000000000 | Conclusions O |
|----------------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| Main Pro | blem: Cross | stalk | | | |



- participants share an acoustic space
- everybody's speech appears on everybody's microphone
- leads to high insertion rates (Pfau et al, ASRU01)

| Introduction | Segmentation | Recognition 00 | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| A Note o | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- ignoring it can lead to WERs > 100%
- overlap

 our approach: use statistics of the occurrence of overlap to account for crosstalk (Laskowski & Schultz, MLMI07)

イロン 不同と 不同と 不同と

| Introduction | Segmentation | Recognition 00 | Data o | Experiments 00000000000 | Conclusions O |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| A Note (| on Terms | | | | |

• a descriptor of the observation on a single channel

- farfield speech appearing on (nearfield) channel
- relatively frequent
- ignoring it can lead to WERs > 100%
- overlap

 our approach: use statistics of the occurrence of overlap to account for crosstalk (Laskowski & Schultz, MLMI07)

イロト イヨト イヨト

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| A Note o | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- ignoring it can lead to WERs > 100%
- overlap

 our approach: use statistics of the occurrence of overlap to account for crosstalk (Laskowski & Schultz, MLMI07)

イロト イヨト イヨト

| Introduction ○○●○ | Segmentation | Recognition 00 | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| A Note of | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- ignoring it can lead to WERs > 100%
- overlap
 - a descriptor of the state of all participants.
 - more than one participant vocalizing simultaneously.
 - o relatively infrequent
 - occurs <10% of the time (Cetin & Shriberg, MLMI06).
- our approach: use statistics of the occurrence of **overlap** to account for **crosstalk** (Laskowski & Schultz, MLMI07)

・ロン ・回 と ・ ヨ と ・ ヨ と

| Introduction ○○●○ | Segmentation | Recognition 00 | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| A Note of | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- $\bullet\,$ ignoring it can lead to WERs >100%
- overlap

a descriptor of the state of all participants

- more than one participant vocalizing simultaneously.
- relatively infrequent
- occurs <10% of the time (Çetin & Shriberg, MLMI06).
- our approach: use statistics of the occurrence of overlap to account for crosstalk (Laskowski & Schultz, MLMI07)

イロト イヨト イヨト

| Introduction ○○●○ | Segmentation | Recognition 00 | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| A Note of | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- $\bullet\,$ ignoring it can lead to WERs >100%
- overlap
 - a descriptor of the state of all participants
 - more than one participant vocalizing simultaneously
 - relatively infrequent
 - occurs <10% of the time (Çetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of **overlap** to account for **crosstalk** (Laskowski & Schultz, MLMI07)

ロトス回とスヨトスヨト

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| A Note o | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- $\bullet\,$ ignoring it can lead to WERs >100%
- overlap
 - a descriptor of the state of all participants
 - more than one participant vocalizing simultaneously
 - relatively infrequent
 - occurs <10% of the time (Çetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of overlap to account for crosstalk (Laskowski & Schultz, MLMI07)

A (10) A (10) A (10) A

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| A Note o | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- $\bullet\,$ ignoring it can lead to WERs >100%
- overlap
 - a descriptor of the state of all participants
 - more than one participant vocalizing simultaneously
 - relatively infrequent
 - occurs <10% of the time (Çetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of overlap to account for crosstalk (Laskowski & Schultz, MLMI07)

化间面 化医丙烯医丙

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| A Note o | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- $\bullet\,$ ignoring it can lead to WERs >100%
- overlap
 - a descriptor of the state of all participants
 - more than one participant vocalizing simultaneously
 - relatively infrequent
 - occurs <10% of the time (Çetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of **overlap** to account for **crosstalk** (Laskowski & Schultz, MLMI07)

A (B) + A (B) + A (B) +

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| A Note o | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- $\bullet\,$ ignoring it can lead to WERs >100%
- overlap
 - a descriptor of the state of all participants
 - more than one participant vocalizing simultaneously
 - relatively infrequent
 - $\bullet\,$ occurs ${<}10\%$ of the time (Çetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of overlap to account for crosstalk (Laskowski & Schultz, MLMI07)

マロト イヨト イヨト

| Introduction ○○●○ | Segmentation | Recognition 00 | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| A Note of | on Terms | | | | |

- a descriptor of the observation on a single channel
- farfield speech appearing on (nearfield) channel
- relatively frequent
- $\bullet\,$ ignoring it can lead to WERs >100%
- overlap
 - a descriptor of the state of all participants
 - more than one participant vocalizing simultaneously
 - relatively infrequent
 - $\bullet\,$ occurs ${<}10\%$ of the time (Çetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of **overlap** to account for **crosstalk** (Laskowski & Schultz, MLMI07)

伺い イヨト イヨト

| Introduction ○○○● | Segmentation | Recognition 00 | Data o | Experiments | Conclusions O |
|----------------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |

Elsewhere:

model participants
 independently

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$
$$= \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

- Supervised acoustic models
- Small frame size (≈10 ms)

This work:

model participants jointly

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$

$$\neq \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

unsupervised acoustic models

イロト イヨト イヨト イヨト

Iarge frame size (≈100 ms)

K. Laskowski, C. Fügen, T. Schultz EUSIPCO 2007: Simultaneous Multispeaker Segmentation

| Introduction ○○○● | Segmentation | Recognition | Data ○ | Experiments 00000000000 | Conclusions O |
|----------------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |

Elsewhere:

model participants independently

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$
$$= \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

- Supervised acoustic models
- Small frame size (≈10 ms)

This work:

model participants
 jointly

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$

$$\neq \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

unsupervised acoustic models

(日) (四) (王) (王)

3 large frame size $(\approx 100 \text{ ms})$

K. Laskowski, C. Fügen, T. Schultz EUSIPCO 2007: Simultaneous Multispeaker Segmentation

| Introduction ○○○● | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|----------------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |

Elsewhere:

model participants independently

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$
$$= \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

- supervised acoustic models
- Small frame size (≈10 ms)

This work:

model participants
 jointly

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$

$$\neq \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

unsupervised acoustic models

(日) (四) (王) (王)

Iarge frame size (≈100 ms)

| Introduction ○○○● | Segmentation | Recognition | Data ○ | Experiments 00000000000 | Conclusions O |
|----------------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |

Elsewhere:

model participants independently

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$
$$= \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

- supervised acoustic models
- Small frame size (≈10 ms)

This work:

model participants
 jointly

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_{j} | \mathbf{q}_{t} = \mathbf{S}_{i}\right)$$

$$\neq \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}^{k} = \mathbf{S}_{j}^{k} | \mathbf{q}_{t}^{k} = \mathbf{S}_{i}^{k}\right)$$

unsupervised acoustic models

 large frame size (≈100 ms)

| Introduction | Segmentation ●○○○○○ | Recognition | Data ○ | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| 6 | | | | | |

Segmentation System Architecture

3 pass segmentation decoding of meetings



K. Laskowski, C. Fügen, T. Schultz EUSIPCO 2007: Simultaneous Multispeaker Segmentation

| Introduction | Segmentation ●○○○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Segment | ation Syster | n Architect | ure | | |

- 3 pass segmentation decoding of meetings
 - prior to decoding: transition model (TM) training



▲ □ ► ▲ □ ►

| Introduction | Segmentation ●○○○○○ | Recognition | Data ○ | Experiments 00000000000 | Conclusions |
|--------------|------------------------|--------------------|-----------|----------------------------|-------------|
| | | | | | |
| <u> </u> | | A 1 | | | |

Segmentation System Architecture

- 3 pass segmentation decoding of meetings
 - PASS 1: initial label assignment



A⊒ ▶ ∢ ∃

-

| Introduction | Segmentation ●○○○○○ | Recognition | Data ○ | Experiments 0000000000 | Conclusions O |
|--------------|------------------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| ~ | | | | | |

Segmentation System Architecture

- 3 pass segmentation decoding of meetings
 - acoustic model (AM) training



-

< ∃ >

A (1) > A (1) > A

| Introduction | Segmentation ●○○○○○ | Recognition | Data o | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Segment | ation Syster | n Architect | ure | | |

3 pass segmentation decoding of meetings

• PASS 2: Viterbi decoding



A (1) > A (1) > A

| Introduction | Segmentation ●○○○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Segment | ation Syster | n Architect | ure | | |

- 3 pass segmentation decoding of meetings
 - PASS 3: segmentation smoothing



∃ ≥ ►

A (1) > A (1) > A

-

| Introduction | Segmentation ○●○○○○ | Recognition 00 | Data ○ | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

• in a meeting of K participants, at time t:

- each participant k is in one of two states
- the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \le i \le 2^k$

model the probability of transition from state S_i to S_j

 $P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$ = $P(||\mathbf{q}_{t+1}|| = ||\mathbf{S}_j||, ||\mathbf{q}_{t+1} \cdot \mathbf{q}_t|| = ||\mathbf{S}_j \cdot \mathbf{S}_i|| ||\mathbf{q}_t|| = ||\mathbf{S}_i||)$

where a transition (n_i, o_{ij}, n_j)

 the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

| Introduction | Segmentation | Recognition 00 | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

- in a meeting of K participants, at time t:
 - each participant k is in one of two states
 - the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \leq i \leq 2^k$

model the probability of transition from state S_i to S_j

 $P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$ = $P(\|\mathbf{q}_{t+1}\| = \|\mathbf{S}_j\|, \|\mathbf{q}_{t+1} \cdot \mathbf{q}_t\| = \|\mathbf{S}_j \cdot \mathbf{S}_i\| \| \|\mathbf{q}_t\| = \|\mathbf{S}_i\|)$

- where a transition (n_i, o_{ij}, n_j)
 - $o_{i} = ||S_i||$ is the # of active speakers in S_i
 - $|\mathbf{S}_j|$ is the # of active speakers in \mathbf{S}_j
 - $\circ \circ_{ij} = ||S_i \cdot S_j||$ is the # of active speakers in both S_i and S_j
- the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

- in a meeting of K participants, at time t:
 - each participant k is in one of two states
 - the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \le i \le 2^K$

model the probability of transition from state S_i to S_j

 $P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$ = $P(\|\mathbf{q}_{t+1}\| = \|\mathbf{S}_j\|, \|\mathbf{q}_{t+1} \cdot \mathbf{q}_t\| = \|\mathbf{S}_j \cdot \mathbf{S}_i\|\|\|\mathbf{q}_t\| = \|\mathbf{S}_i\|)$

- where a transition (n_i, o_{ij}, n_j)
 - $n_i = ||S_i||$ is the # of active speakers in S_i
 - $\sigma_j = |S_j|$ is the # of active speakers in S_j
 - $| \bullet \circ o_g = | [\mathsf{S}_i \circ \mathsf{S}_j] |$ is the #-of active speakers in both $|\mathsf{S}_i|$ and $|\mathsf{S}_j|$
- the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

・ロン ・回と ・ヨン・

| Introduction | Segmentation ○●○○○○ | Recognition | Data ○ | Experiments | Conclusions |
|--------------|------------------------|--------------------|-----------|-------------|-------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

- in a meeting of K participants, at time t:
 - each participant k is in one of two states
 - the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \leq i \leq 2^K$
- model the probability of transition from state **S**_i to **S**_j

$$P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$$

= $P(||\mathbf{q}_{t+1}|| = ||\mathbf{S}_j||, ||\mathbf{q}_{t+1} \cdot \mathbf{q}_t|| = ||\mathbf{S}_j \cdot \mathbf{S}_i|| |||\mathbf{q}_t|| = ||\mathbf{S}_i||)$

• where a transition (n_i, o_{ij}, n_j)

- $n_i = ||\mathbf{S}_i||$ is the # of active speakers in \mathbf{S}_i
- $n_j = \|\mathbf{S}_j\|$ is the # of active speakers in \mathbf{S}_j
- $o_{ij} = \|\mathbf{S}_i \cdot \mathbf{S}_j\|$ is the # of active speakers in both \mathbf{S}_i and \mathbf{S}_j
- the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

| Introduction | Segmentation ○●○○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

- in a meeting of K participants, at time t:
 - each participant k is in one of two states
 - the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \le i \le 2^K$
- model the probability of transition from state S_i to S_j

$$P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$$

= $P(||\mathbf{q}_{t+1}|| = ||\mathbf{S}_j||, ||\mathbf{q}_{t+1} \cdot \mathbf{q}_t|| = ||\mathbf{S}_j \cdot \mathbf{S}_i|| |||\mathbf{q}_t|| = ||\mathbf{S}_i||)$

• where a transition
$$(n_i, o_{ij}, n_j)$$

- $n_i = \|\mathbf{S}_i\|$ is the # of active speakers in \mathbf{S}_i
- $n_j = \|\mathbf{S}_j\|$ is the # of active speakers in \mathbf{S}_j
- $o_{ij} = \|\mathbf{S}_i \cdot \mathbf{S}_j\|$ is the # of active speakers in both \mathbf{S}_i and \mathbf{S}_j
- the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

イロト イヨト イヨト

| Introduction | Segmentation ○●○○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

- in a meeting of K participants, at time t:
 - each participant k is in one of two states
 - the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \leq i \leq 2^K$
- model the probability of transition from state S_i to S_j

$$P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$$

= $P(||\mathbf{q}_{t+1}|| = ||\mathbf{S}_j||, ||\mathbf{q}_{t+1} \cdot \mathbf{q}_t|| = ||\mathbf{S}_j \cdot \mathbf{S}_i|| |||\mathbf{q}_t|| = ||\mathbf{S}_i||)$

• where a transition
$$(n_i, o_{ij}, n_j)$$

- $n_i = \|\mathbf{S}_i\|$ is the # of active speakers in \mathbf{S}_i
- $n_j = \|\mathbf{S}_j\|$ is the # of active speakers in \mathbf{S}_j

• $o_{ij} = \|\mathbf{S}_i \cdot \mathbf{S}_j\|$ is the # of active speakers in both \mathbf{S}_i and \mathbf{S}_j

• the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

・ロト ・同ト ・ヨト ・ヨト
| Introduction | Segmentation ○●○○○○ | Recognition | Data ○ | Experiments | Conclusions |
|--------------|------------------------|--------------------|-----------|-------------|-------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

- in a meeting of K participants, at time t:
 - each participant k is in one of two states
 - the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \le i \le 2^K$
- model the probability of transition from state **S**_i to **S**_j

$$P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$$

= $P(||\mathbf{q}_{t+1}|| = ||\mathbf{S}_j||, ||\mathbf{q}_{t+1} \cdot \mathbf{q}_t|| = ||\mathbf{S}_j \cdot \mathbf{S}_i|| |||\mathbf{q}_t|| = ||\mathbf{S}_i||)$

• where a transition (n_i, o_{ij}, n_j)

- $n_i = \|\mathbf{S}_i\|$ is the # of active speakers in \mathbf{S}_i
- $n_j = \|\mathbf{S}_j\|$ is the # of active speakers in \mathbf{S}_j
- $o_{ij} = \|\mathbf{S}_i \cdot \mathbf{S}_j\|$ is the # of active speakers in both \mathbf{S}_i and \mathbf{S}_j
- the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

(本間) (本語) (本語)

| Introduction | Segmentation ○●○○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Transitio | n Model Tra | aining | | | |

- in a meeting of K participants, at time t:
 - each participant k is in one of two states
 - the meeting \mathbf{q}_t is in \mathbf{S}_i , $1 \le i \le 2^K$
- model the probability of transition from state **S**_i to **S**_j

$$P(\mathbf{q}_{t+1} = \mathbf{S}_j | \mathbf{q}_t = \mathbf{S}_i)$$

= $P(||\mathbf{q}_{t+1}|| = ||\mathbf{S}_j||, ||\mathbf{q}_{t+1} \cdot \mathbf{q}_t|| = ||\mathbf{S}_j \cdot \mathbf{S}_i|| |||\mathbf{q}_t|| = ||\mathbf{S}_i||)$

• where a transition (n_i, o_{ij}, n_j)

- $n_i = \|\mathbf{S}_i\|$ is the # of active speakers in \mathbf{S}_i
- $n_j = \|\mathbf{S}_j\|$ is the # of active speakers in \mathbf{S}_j
- $o_{ij} = \|\mathbf{S}_i \cdot \mathbf{S}_j\|$ is the # of active speakers in both \mathbf{S}_i and \mathbf{S}_j
- the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

・ 同下 ・ ヨト ・ ヨト

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|---------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| PASS 1 | : Initial Lab | el Assignme | ent | | |

• at time t, for each participant k, $1 \le k \le K$,

- compute energy $\phi_{kk}(0)$
- compute crosscorrelation $\phi_{jk}\left(au
 ight)$ with every other participant j
- NT-Normalize (Laskowski & Schultz, ICSLP04) φ_{jk}(τ, φ_{ii}(0)
- assign the initial label

$$ilde{\mathbf{q}} \begin{bmatrix} k \end{bmatrix} = egin{cases} \mathcal{V}, & ext{if } \sum_{j
eq k} \log\left(rac{\max_{ au} \phi_{jk}(au)}{\phi_{jj}(0)}
ight) > 0 \ \mathcal{N}, & ext{otherwise} \end{cases}$$

 approximates declaring participant k as vocalizing (V) when the dominant sound source is closer to microphone k than to the centroid of the remaining microphones (Laskowski & Schultz, NAACL07)

イロン イヨン イヨン イヨン

| Introduction | Segmentation ○○●○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| PASS 1 | : Initial Lab | el Assignme | ent | | |

- at time t, for each participant k, $1 \le k \le K$,
 - compute energy $\phi_{kk}(0)$
 - compute crosscorrelation $\phi_{jk}\left(au
 ight)$ with every other participant j
 - NT-Normalize (Laskowski & Schultz, ICSLP04) φ_{ik}(τ) φ_{ii}(0)

• assign the initial label

$$ilde{\mathbf{q}} \left[k
ight] \;\; = \;\; \left\{ egin{array}{cc} \mathcal{V}, & ext{if } \sum_{j
eq k} \log \left(rac{\max_{ au} \phi_{jk}(au)}{\phi_{jj}(0)}
ight) > 0 \ \mathcal{N}, & ext{otherwise} \end{array}
ight.$$

 approximates declaring participant k as vocalizing (V) when the dominant sound source is closer to microphone k than to the centroid of the remaining microphones (Laskowski & Schultz, NAACL07)

イロト イヨト イヨト

| Introduction | Segmentation ○○●○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| PASS 1 | : Initial Lab | el Assignmo | ent | | |

- at time t, for each participant k, $1 \le k \le K$,
 - compute energy $\phi_{kk}(0)$
 - compute crosscorrelation $\phi_{jk}\left(au
 ight)$ with every other participant j
 - **NT-Normalize** (Laskowski & Schultz, ICSLP04) $\frac{\phi_{jk}(\tau)}{\phi_{ii}(0)}$
 - assign the initial label

$$ilde{\mathbf{q}} \begin{bmatrix} k \end{bmatrix} = egin{cases} \mathcal{V}, & ext{if } \sum_{j
eq k} \log\left(rac{\max_{ au} \phi_{jk}(au)}{\phi_{jj}(0)}
ight) > 0 \ \mathcal{N}, & ext{otherwise} \end{cases}$$

 approximates declaring participant k as vocalizing (V) when the dominant sound source is closer to microphone k than to the centroid of the remaining microphones (Laskowski & Schultz, NAACL07)

イロト イヨト イヨト

| Introduction | Segmentation ○○●○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| PASS 1 | Initial Lab | el Assignmo | ent | | |

- at time t, for each participant k, $1 \le k \le K$,
 - compute energy $\phi_{kk}(0)$
 - compute crosscorrelation $\phi_{jk}\left(au
 ight)$ with every other participant j
 - NT-Normalize (Laskowski & Schultz, ICSLP04) $\frac{\phi_{jk}(\tau)}{\phi_{ii}(0)}$

assign the initial label

$$ilde{\mathbf{q}}\left[k
ight] \;\;=\;\; \left\{ egin{array}{cc} \mathcal{V}, & ext{if } \sum_{j
eq k} \log\left(rac{\max_{ au} \phi_{jk}(au)}{\phi_{jj}(0)}
ight) > 0 \ \mathcal{N}, & ext{otherwise} \end{array}
ight.$$

 approximates declaring participant k as vocalizing (V) when the dominant sound source is closer to microphone k than to the centroid of the remaining microphones (Laskowski & Schultz, NAACL07)

イロン イヨン イヨン

| Introduction | Segmentation ○○●○○○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| PASS 1 | Initial Lab | al Assignm | ≏nt | | |

- at time t, for each participant k, $1 \le k \le K$,
 - compute energy $\phi_{kk}(0)$
 - compute crosscorrelation $\phi_{jk}\left(au
 ight)$ with every other participant j
 - **NT-Normalize** (Laskowski & Schultz, ICSLP04) $\frac{\phi_{jk}(\tau)}{\phi_{ii}(0)}$
 - assign the initial label

$$\tilde{\mathbf{q}} \begin{bmatrix} k \end{bmatrix} = \begin{cases} \mathcal{V}, & \text{if } \sum_{j \neq k} \log\left(\frac{\max_{\tau} \phi_{jk}(\tau)}{\phi_{jj}(0)}\right) > 0\\ \mathcal{N}, & \text{otherwise} \end{cases}$$

 approximates declaring participant k as vocalizing (V) when the dominant sound source is closer to microphone k than to the centroid of the remaining microphones (Laskowski & Schultz, NAACL07)

(4月) キョン キョン

| Introduction | Segmentation ○○●○○○ | Recognition | Data o | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| PASS 1 | Initial Lab | el Assignm | ent | | |

- at time t, for each participant k, $1 \le k \le K$,
 - compute energy $\phi_{kk}(0)$
 - compute crosscorrelation $\phi_{jk}\left(au
 ight)$ with every other participant j
 - **NT-Normalize** (Laskowski & Schultz, ICSLP04) $\frac{\phi_{jk}(\tau)}{\phi_{ii}(0)}$
 - assign the initial label

$$\tilde{\mathbf{q}} \begin{bmatrix} k \end{bmatrix} = \begin{cases} \mathcal{V}, & \text{if } \sum_{j \neq k} \log\left(\frac{\max_{\tau} \phi_{jk}(\tau)}{\phi_{jj}(0)}\right) > 0\\ \mathcal{N}, & \text{otherwise} \end{cases}$$

 approximates declaring participant k as vocalizing (V) when the dominant sound source is closer to microphone k than to the centroid of the remaining microphones (Laskowski & Schultz, NAACL07)

| Introduction | Segmentation | Recognition 00 | Data ⊙ | Experiments | Conclusions O |
|--------------|--------------|--------------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Trai | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^K states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^{*K*} states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- 4 回 > - 4 回 > - 4 回 >

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
 for each of 2^K states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- 4 回 2 - 4 三 2 - 4 三 2

| Introduction | Segmentation ○○○●○○ | Recognition | Data ○ | Experiments 0000000000 | Conclusions O |
|--------------|------------------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^{κ} states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- \circ shrinkage towards a global model (λ_G).
- \circ channel-rotated feature vectors (λ_R)
- \circ sample-level synthesized overlap (λ_{s})

向下 イヨト イヨト

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^K states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- ullet shrinkage towards a global model (λ_G) .
- \circ channel-rotated feature vectors (λ_R)
- \circ sample-level synthesized overlap (λ_5)

向下 イヨト イヨト

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^{K} states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- shrinkage towards a global model (λ_G)
- channel-rotated feature vectors (λ_R
- sample-level synthesized overlap (λ_S)

マロト イヨト イヨト

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^{K} states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- shrinkage towards a global model (λ_G)
- channel-rotated feature vectors (λ_R)
- sample-level synthesized overlap (λ_S)

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^{K} states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- shrinkage towards a global model (λ_G)
- channel-rotated feature vectors (λ_R)

• sample-level synthesized overlap (λ_S)

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Acoustic | Model Train | ning | | | |

- one Gaussian model
- with a **full** covariance matrix
- for a 2K-length vector (energy and zero-crossing rate per channel) → may be different for different meetings
- for each of 2^{K} states

 since many states may have too little data, train models (Laskowski & Schultz, ICASSP06) using

- shrinkage towards a global model (λ_G)
- channel-rotated feature vectors (λ_R)
- sample-level synthesized overlap (λ_S)

| Introduction | Segmentation ○○○○●○ | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|------------------------|--------------------|-----------|-------------|------------------|
| | | | | | |

• standard Viterbi decoding using

- frame size and step of 110 ms
- the pre-trained, meeting-independent transition model (TM)
- the meeting-specific acoustic model (AM)
- complexity: $1 \le t \le T$, $1 \le i \le N \equiv 2^{\kappa}$
- assume all participants are silent at time t = 0

• following decoding, the state of each participant k at time t is

$$\mathbf{q}_t \left[k \right]^* = \mathbf{q}_t^* \left[k \right]$$

<回と < 回と < 回

| Introduction | Segmentation | Recognition | Data ○ | Experiments 00000000000 | Conclusions O |
|--------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| | | | | | |

standard Viterbi decoding using

- frame size and step of 110 ms
- the pre-trained, meeting-independent transition model (TM)
- the meeting-specific acoustic model (AM)
- complexity: $1 \le t \le T$, $1 \le i \le N \equiv 2^{K}$
- assume all participants are silent at time t = 0

• following decoding, the state of each participant k at time t is

$$\mathbf{q}_t \left[k \right]^* = \mathbf{q}_t^* \left[k \right]$$

<回と < 回と < 回

| Introduction | Segmentation ○○○○●○ | Recognition | Data o | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| DACCA | | 1. | | | |

standard Viterbi decoding using

- frame size and step of 110 ms
- the pre-trained, meeting-independent transition model (TM)
- the meeting-specific acoustic model (AM)
- complexity: $1 \le t \le T$, $1 \le i \le N \equiv 2^k$
- assume all participants are silent at time t = 0

• following decoding, the state of each participant k at time t is

$$\mathbf{q}_t \left[k \right]^* = \mathbf{q}_t^* \left[k \right]$$

向下 イヨト イヨ

| Introduction | Segmentation ○○○○●○ | Recognition 00 | Data ○ | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| PASS 2 | : Viterbi De | coding | | | |

standard Viterbi decoding using

- frame size and step of 110 ms
- the pre-trained, meeting-independent transition model (TM)
- the meeting-specific acoustic model (AM)
- complexity: 1≤t≤T, 1≤i≤N≡2^k
- assume all participants are silent at time t = 0

• following decoding, the state of each participant k at time t is

$$\mathbf{q}_t \left[k \right]^* = \mathbf{q}_t^* \left[k \right]$$

向下 イヨト イヨト

| Introduction | Segmentation ○○○○●○ | Recognition 00 | Data ○ | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| PASS 2 | : Viterbi De | coding | | | |

standard Viterbi decoding using

- frame size and step of 110 ms
- the pre-trained, meeting-independent transition model (TM)
- the meeting-specific acoustic model (AM)
- complexity: $1 \le t \le T$, $1 \le i \le N \equiv 2^K$
- assume all participants are silent at time t = 0

• following decoding, the state of each participant k at time t is

$$\mathbf{q}_t \left[k \right]^* = \mathbf{q}_t^* \left[k \right]$$

| Introduction | Segmentation ○○○○●○ | Recognition 00 | Data ○ | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| DACCO | | and Provide | | | |

standard Viterbi decoding using

- frame size and step of 110 ms
- the pre-trained, meeting-independent transition model (TM)
- the meeting-specific acoustic model (AM)
- complexity: $1 \le t \le T$, $1 \le i \le N \equiv 2^K$
- assume all participants are silent at time t = 0

• following decoding, the state of each participant k at time t is

$$\mathbf{q}_t \left[k \right]^* = \mathbf{q}_t^* \left[k \right]$$

| Introduction | Segmentation ○○○○●○ | Recognition 00 | Data ○ | Experiments 00000000000 | Conclusions O |
|--------------|------------------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| PASS 2 | : Viterbi De | coding | | | |

- standard Viterbi decoding using
 - frame size and step of 110 ms
 - the pre-trained, meeting-independent transition model (TM)
 - the meeting-specific acoustic model (AM)
 - complexity: $1 \le t \le T$, $1 \le i \le N \equiv 2^K$
 - assume all participants are silent at time t = 0

• following decoding, the state of each participant k at time t is

$$\mathbf{q}_t \left[k \right]^* = \mathbf{q}_t^* \left[k \right]$$

| Introduction | Segmentation ○○○○● | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- \blacksquare bridge ${\mathcal V}$ gaps shorter than 500 ms
- ② eliminate $\mathcal V$ spurts shorter than 200 ms
- \blacksquare pre-pad all ${\mathcal V}$ intervals with 100 ms
- I post-pad all $\mathcal V$ intervals with 300 ms
- \bullet bridge remaining $\mathcal V$ gaps shorter than 400 ms
- lacksquare eliminate remaining ${\mathcal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

| Introduction | Segmentation ○○○○● | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- **()** bridge \mathcal{V} gaps shorter than 500 ms
- ② eliminate ${\mathcal V}$ spurts shorter than 200 ms
- \odot pre-pad all ${\mathcal V}$ intervals with 100 ms
- \blacksquare post-pad all ${\mathcal V}$ intervals with 300 ms
- ullet bridge remaining ${\mathcal V}$ gaps shorter than 400 ms
- lacksquare eliminate remaining ${\mathcal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

| Introduction | Segmentation ○○○○● | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- **()** bridge \mathcal{V} gaps shorter than 500 ms
- 2 eliminate \mathcal{V} spurts shorter than 200 ms
- ${f 0}$ pre-pad all ${\cal V}$ intervals with 100 ms
- 0 post-pad all $\mathcal V$ intervals with 300 ms
- ullet bridge remaining ${\mathcal V}$ gaps shorter than 400 ms
- lacksquare eliminate remaining ${\mathcal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

| Introduction | Segmentation ○○○○● | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- **()** bridge \mathcal{V} gaps shorter than 500 ms
- 2 eliminate \mathcal{V} spurts shorter than 200 ms
- $\textcircled{O} \text{ pre-pad all } \mathcal{V} \text{ intervals with 100 ms}$
- 0 post-pad all $\mathcal V$ intervals with 300 ms
- ${igsinable {igsin} {igsin {\mathfrak o} {igni} { higni} {igni} {igni} {igni} {igni} { higni} {igni} {igni} {igni} {igni} { higni} {igni} { higni} {igni} {igni} {igni} {igni} { higni} { higni} {igni} { higni} { higni} { higni} {igni} { higni} {$
- lace eliminate remaining ${\mathcal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

| Introduction | Segmentation ○○○○● | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- **()** bridge \mathcal{V} gaps shorter than 500 ms
- 2 eliminate \mathcal{V} spurts shorter than 200 ms
- $\textcircled{O} \text{ pre-pad all } \mathcal{V} \text{ intervals with 100 ms}$
- **④** post-pad all \mathcal{V} intervals with 300 ms
- ${f ar o}$ bridge remaining ${\cal V}$ gaps shorter than 400 ms
- ullet eliminate remaining ${\mathcal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

伺い イヨト イヨト

| Introduction | Segmentation ○○○○● | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- **()** bridge \mathcal{V} gaps shorter than 500 ms
- 2 eliminate \mathcal{V} spurts shorter than 200 ms
- **③** pre-pad all \mathcal{V} intervals with 100 ms
- post-pad all $\mathcal V$ intervals with 300 ms
- ${f 0}$ bridge remaining ${\cal V}$ gaps shorter than 400 ms
- \blacksquare eliminate remaining ${\mathcal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

向下 イヨト イヨト

| Introduction | Segmentation ○○○○● | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- **()** bridge \mathcal{V} gaps shorter than 500 ms
- 2 eliminate \mathcal{V} spurts shorter than 200 ms
- $\textcircled{O} \text{ pre-pad all } \mathcal{V} \text{ intervals with 100 ms}$
- **④** post-pad all \mathcal{V} intervals with 300 ms
- ${f 0}$ bridge remaining ${\cal V}$ gaps shorter than 400 ms
- ${f 0}$ eliminate remaining ${\cal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

向下 イヨト イヨト

| Introduction | Segmentation ○○○○● | Recognition | Data ⊙ | Experiments | Conclusions O |
|--------------|-----------------------|-------------|-----------|-------------|------------------|
| | | | | | |
| PASS 3: | Smoothing | | | | |

- **()** bridge \mathcal{V} gaps shorter than 500 ms
- 2 eliminate \mathcal{V} spurts shorter than 200 ms
- $\textcircled{O} \text{ pre-pad all } \mathcal{V} \text{ intervals with 100 ms}$
- **④** post-pad all \mathcal{V} intervals with 300 ms
- ${f 0}$ bridge remaining ${\cal V}$ gaps shorter than 400 ms
- ${f 0}$ eliminate remaining ${\cal V}$ spurts shorter than 800 ms

Determined empirically during NIST RT-06s development (Fügen et al, MLMI2006)

| Introduction | Segmentation | Recognition ●○ | Data ○ | Experiments | Conclusions O |
|--------------|--------------|-------------------|-----------|-------------|------------------|
| | | | | | |

• standard: frame-level miss and false alarm rates

- in our work, these correlate poorly with WER rates
- for meeting recognition, will use the **WER gap** between automatic segmentation and manual reference segmentation
 - development ASR system: L pass, MFCC front-end,
 - out-of-domain LM
 - more realistic ASR system: 3 passes, MECC and MVDR front.
 ends, in-domain LM

| Introduction | Segmentation | Recognition ●○ | Data ○ | Experiments | Conclusions |
|--------------|--------------|-------------------|-----------|-------------|-------------|
| | | | | | |

- standard: frame-level miss and false alarm rates
- in our work, these correlate poorly with WER rates
- for meeting recognition, will use the **WER gap** between automatic segmentation and manual reference segmentation
 - development ASR system: 1 pass, MFCC front-end, out-of-domain LM
 - more realistic ASR system: 3 passes, MECC and MVDR front.
 ends, in-domain LM

| Introduction | Segmentation | Recognition ●○ | Data ○ | Experiments | Conclusions |
|--------------|--------------|-------------------|-----------|-------------|-------------|
| | | | | | |

- standard: frame-level miss and false alarm rates
- in our work, these correlate poorly with WER rates
- for meeting recognition, will use the **WER gap** between automatic segmentation and manual reference segmentation
 - development ASR system: 1 pass, MFCC front-end, out-of-domain LM
 - more realistic ASR system: 3 passes, MFCC and MVDR front ends, in-domain LM

マロト イヨト イヨト

| Introduction | Segmentation | Recognition ●○ | Data ○ | Experiments | Conclusions |
|--------------|--------------|-------------------|-----------|-------------|-------------|
| | | | | | |

- standard: frame-level miss and false alarm rates
- in our work, these correlate poorly with WER rates
- for meeting recognition, will use the **WER gap** between automatic segmentation and manual reference segmentation
 - development ASR system: 1 pass, MFCC front-end, out-of-domain LM
 - more realistic ASR system: 3 passes, MFCC and MVDR front ends, in-domain LM

A (10) N (10)
| Introduction | Segmentation | Recognition ●○ | Data ○ | Experiments | Conclusions |
|--------------|--------------|-------------------|-----------|-------------|-------------|
| | | | | | |

Evaluating Segmentation for Speech Recognition

- standard: frame-level miss and false alarm rates
- in our work, these correlate poorly with WER rates
- for meeting recognition, will use the **WER gap** between automatic segmentation and manual reference segmentation
 - development ASR system: 1 pass, MFCC front-end, out-of-domain LM
 - more realistic ASR system: 3 passes, MFCC and MVDR front ends, in-domain LM

| Introduction | Segmentation | Recognition | Data | Experiments | Conclusions |
|--------------|--------------|-------------|------|-------------|-------------|
| | | 00 - | | | |
| | | | | | |

3-Pass Recognition System Architecture



K. Laskowski, C. Fügen, T. Schultz EUSIPCO 2007: Simultaneous Multispeaker Segmentation

<ロ> (四) (四) (三) (三) (三)

| Introduction | Segmentation | Recognition | Data ● | Experiments | Conclusions O |
|--------------|--------------|-------------|-----------|-------------|------------------|
| | | | | | |

development set

- 10 meeting excerpts
 - 2 × CMU
 - $2 \times ED$
 - 2 × ICSI
 - $2 \times \text{NIST}$
 - $2 \times VT$
- 10 minutes each
- one meeting with an unmic'ed speaker
 → excluded during development

evaluation set

- 9 meeting excerpts
 2 × CMU
 2 × EDI
 2 × NIST
 1 × TNO
 2 × VT
- 10 minutes each

| Introduction | Segmentation | Recognition | Data ● | Experiments | Conclusions |
|--------------|--------------|--------------------|-----------|-------------|-------------|
| | | | | | |

development set

- I0 meeting excerpts
 - $2 \times CMU$
 - $2 \times EDI$
 - $2 \times \text{ICSI}$
 - $2 \times \text{NIST}$
 - $2 \times VT$
- 10 minutes each
- one meeting with an unmic'ed speaker
 → excluded during development

evaluation set

- 9 meeting excerpts
 - 2 × CMU
 - 2 × ED
 - 2 × NIST
 - 1 × TNO
 - 2 × VT
- 10 minutes each

(本間) (本語) (本語)

| Introduction | Segmentation | Recognition | Data ● | Experiments | Conclusions |
|--------------|--------------|--------------------|-----------|-------------|-------------|
| | | | | | |

development set

- I0 meeting excerpts
 - $2 \times CMU$
 - $2 \times EDI$
 - $2 \times \text{ICSI}$
 - $2 \times \text{NIST}$
 - $2 \times VT$
- 10 minutes each
- one meeting with an unmic'ed speaker
 → excluded during development

evaluation set

- 9 meeting excerpts
 - 2 × CMU
 - 2 × EDI
 - $2 \times \text{NIST}$
 - 1 × TNO
 - $2 \times VT$
- 10 minutes each

(本間) (本語) (本語)

| Introduction | Segmentation | Recognition | Data ● | Experiments | Conclusions O |
|--------------|--------------|-------------|-----------|-------------|------------------|
| | | | | | |

development set

- I0 meeting excerpts
 - $2 \times CMU$
 - $2 \times EDI$
 - $2 \times \text{ICSI}$
 - $2 \times \text{NIST}$
 - $2 \times VT$
- 10 minutes each
- one meeting with an unmic'ed speaker
 → excluded during development

evaluation set

- 9 meeting excerpts
 - $2 \times CMU$
 - 2 × EDI
 - $2 \times \text{NIST}$
 - $1 \times \text{TNO}$
 - $2 \times VT$

10 minutes each

| Introduction | Segmentation | Recognition | Data ● | Experiments | Conclusions O |
|--------------|--------------|-------------|-----------|-------------|------------------|
| | | | | | |

development set

- I0 meeting excerpts
 - $2 \times CMU$
 - $2 \times EDI$
 - $2 \times \text{ICSI}$
 - $2 \times \text{NIST}$
 - $2 \times VT$
- 10 minutes each
- one meeting with an unmic'ed speaker
 → excluded during development

evaluation set

- 9 meeting excerpts
 - $2 \times CMU$
 - 2 × EDI
 - $2 \times \text{NIST}$
 - $1 \times \text{TNO}$
 - $2 \times VT$
- 10 minutes each

| Introduction | Segmentation | Recognition | Data ○ | Experiments ●○○○○○○○○○ | Conclusions |
|--------------|--------------|--------------------|-----------|---------------------------|-------------|
| | | | | | |
| Baseline | | | | | |

segmentation system as described

- one pass ASR system, MFCC only, out-of-domain LM
- development data (one meeting excluded)

・ 同 ト ・ ヨ ト ・ ヨ ト

| Introduction | Segmentation | Recognition | Data ○ | Experiments ●○○○○○○○○○ | Conclusions |
|--------------|--------------|--------------------|-----------|---------------------------|-------------|
| | | | | | |
| Baseline | | | | | |

- segmentation system as described
- one pass ASR system, MFCC only, out-of-domain LM
- development data (one meeting excluded)

| Introduction | Segmentation | Recognition | Data ○ | Experiments ●○○○○○○○○○ | Conclusions |
|--------------|--------------|--------------------|-----------|---------------------------|-------------|
| | | | | | |
| Baseline | | | | | |

- segmentation system as described
- one pass ASR system, MFCC only, out-of-domain LM
- development data (one meeting excluded)

| Introduction | Segmentation | Recognition | Data ○ | Experiments ●○○○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Baseline | | | | | |

- segmentation system as described
- one pass ASR system, MFCC only, out-of-domain LM
- development data (one meeting excluded)



| Introduction | Segmentation | Recognition | Data ○ | Experiments ○●○○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Modificat | ion 1 · Flim | inate Zero | Crossing | v Rate (no70 | R) |

• remove the ZCR feature

• reduces feature vector length from 2K to K features

• retune acoustic model training parameters (λ_G , λ_R , λ_S)

- 4 回 2 - 4 □ 2 - 4 □

| Introduction | Segmentation | Recognition | Data o | Experiments ○●○○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Modificat | ion 1: Elim | inate Zero | Crossing | g Rate (noZO | CR) |

- remove the ZCR feature
 - reduces feature vector length from 2K to K features
- retune acoustic model training parameters (λ_G , λ_R , λ_S)

・ 同下 ・ ヨト ・ ヨト

| Introduction | Segmentation | Recognition | Data o | Experiments ○●○○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Modifica | tion 1: Elim | inate Zero | Crossing | g Rate (noZO | CR) |

- remove the ZCR feature
 - reduces feature vector length from 2K to K features
- retune acoustic model training parameters (λ_G , λ_R , λ_S)

A⊒ ▶ ∢ ∃

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○●○○○○○○○○ | Conclusions O |
|--------------|--------------------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Modifica | tion 1 [.] Flim | inate Zero | Crossing | Rate (no70 | ^R) |

- remove the ZCR feature
 - reduces feature vector length from 2K to K features
- retune acoustic model training parameters (λ_G , λ_R , λ_S)



| Introduction | Segmentation | Recognition 00 | Data o | Experiments ○○●○○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Modifica | tion 2: Redi | uce the Fra | me Size | (F.100) | |

- reduce the frame size and frame step from 110ms to 100ms
- retune the smoothing pass parameters

- 4 回 2 - 4 三 2 - 4 三 2

| Introduction | Segmentation | Recognition 00 | Data ○ | Experiments ○○●○○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Modificati | on 2: Redi | ice the Fra | me Size | (F.100) | |

- reduce the frame size and frame step from 110ms to 100ms
- retune the smoothing pass parameters

- 4 ⊒ >

| Introduction | Segmentation | Recognition 00 | Data ○ | Experiments ○○●○○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Modifica | tion 2: Redu | ice the Fra | me Size | (F.100) | |

- reduce the frame size and frame step from 110ms to 100ms
- retune the smoothing pass parameters



| Modification 3: Select Data for the Silence Model (ILA.0) | Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○●○○○○○○ | Conclusions O |
|---|--------------|---------------|--------------------|-----------|---------------------------|------------------|
| Modification 3: Select Data for the Silence Model (ILA.0) | | | | | | |
| | Modifica | tion 3: Selec | t Data for | the Sile | ence Model (| ILA.0) |

• OBSERVATION: ILA has high precision, but lower recall

• SOLUTION: use only (quieter) half of the frames classified as silence for training the silence model

→ ∃ →

∃ >

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○●○○○○○○○ | Conclusions O |
|--------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| Modificati | on 3: Select | : Data for | the Sile | nce Model (I | LA.0) |

- OBSERVATION: ILA has high precision, but lower recall
- SOLUTION: use only (quieter) half of the frames classified as silence for training the silence model

| Introduction | Segmentation | Recognition | Data o | Experiments ○○○●○○○○○○ | Conclusions O |
|--------------|--------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Modificat | ion 3: Selec | t Data for | the Sile | nce Model (| ILA.0) |

- OBSERVATION: ILA has high precision, but lower recall
- SOLUTION: use only (quieter) half of the frames classified as silence for training the silence model



| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|---------------|--------------------|------------|--------------------|------------------|
| | | | | | |
| Modificat | tion 4: Disti | ribute EDC |) Mass (| MULT) | |
| | | | | | |
| • OI | BSERVATION: | EDO transitic | ons have m | ultiple target sta | ates |
| | | | | | |
| | | | | | |

• SOLUTION: distribute the probability mass among same-EDO next states with uniform probability

・回 と く ヨ と く ヨ と

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|---------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Modifica | tion 4: Disti | ribute EDO | Mass (| MULT) | |
| | | | | | |

OBSERVATION: EDO transitions have multiple target states
 eg. (2,1,1): {A, B} → {A} and {A, B} → {B}
 SOLUTION: distribute the probability mass among same-EDO

 SOLUTION: distribute the probability mass among same-EDO next states with uniform probability

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions O |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Modificatio | on 4: Distri | ibute EDO | Mass (| MULT) | |

- OBSERVATION: EDO transitions have multiple target states
 eg. (2,1,1): {A, B} → {A} and {A, B} → {B}
- SOLUTION: distribute the probability mass among same-EDO next states with uniform probability

| Introduction | Segmentation | Recognition 00 | Data ○ | Experiments ○○○●○○○○○○ | Conclusions O |
|--------------|--------------|--------------------------|-----------|---------------------------|------------------|
| | | | | | |
| Modificatio | on 4: Distr | ribute EDO | Mass (| MULT) | |

- OBSERVATION: EDO transitions have multiple target states
 eg. (2,1,1): {A, B} → {A} and {A, B} → {B}
- SOLUTION: distribute the probability mass among same-EDO next states with uniform probability



| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○●○○○○○ | Conclusions |
|--------------|--------------|--------------------|-----------|---------------------------|-------------|
| | | | | | |
| Modificat | ion 5: Limi | t Simultane | eous Vo | calization (O | V.2) |

• explicitly eliminate all states with more than 2 participants vocalizing at once

-

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○●○○○○○ | Conclusions O |
|--------------|--------------|--------------------|-----------|---------------------------|------------------|
| | | | | | |
| Modificat | tion 5: Limi | t Simultane | eous Vo | calization (O | V.2) |

• explicitly eliminate all states with more than 2 participants vocalizing at once



| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions |
|--------------|--------------|-------------|-----------|-------------|-------------|
| | | | | | |

- initially, excluded one meeting from the devset
- on the whole devset, WERs are higher all around
- but each modification still leads to a reduction
- overall relative error reduction of 82%



| Introduction | Segmentation | Recognition | Data | Experiments | Conclusions |
|--------------|--------------|-------------|------|-------------|-------------|
| | | | | 00000000000 | |
| | | | | | |

- initially, excluded one meeting from the devset
- on the whole devset, WERs are higher all around
- but each modification still leads to a reduction
- overall relative error reduction of 82%



| Introduction | Segmentation | Recognition | Data | Experiments | Conclusions |
|--------------|--------------|-------------|------|-------------|-------------|
| | | | | 00000000000 | |
| | | | | | |

- initially, excluded one meeting from the devset
- on the whole devset, WERs are higher all around
- but each modification still leads to a reduction
- overall relative error reduction of 82%



| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○●○○○○ | Conclusions |
|--------------|--------------|--------------------|-----------|---------------------------|-------------|
| | | | | | |

- initially, excluded one meeting from the devset
- on the whole devset, WERs are higher all around
- but each modification still leads to a reduction
- overall relative error reduction of 82%



| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○●○○○○ | Conclusions |
|--------------|--------------|-------------|-----------|---------------------------|-------------|
| | | | | | |

- initially, excluded one meeting from the devset
- on the whole devset, WERs are higher all around
- but each modification still leads to a reduction
- overall relative error reduction of 82%



| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○●○○○○ | Conclusions |
|--------------|--------------|--------------------|-----------|---------------------------|-------------|
| | | | | | |

- initially, excluded one meeting from the devset
- on the whole devset, WERs are higher all around
- but each modification still leads to a reduction
- overall relative error reduction of 82%



| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○●○○○ | Conclusions |
|--------------|--------------|-------------|-----------|---------------------------|-------------|
| | | | | | |

Results: Development LM \rightarrow Evaluation LM



 when moving to in-domain LM, absolute WER reduction due to improved segmentation is smaller by approximately 10%rel

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○●○○○ | Conclusions O |
|--------------|--------------|--------------------|------------------|---------------------------|------------------|
| | | | | | |

Results: Development LM \rightarrow Evaluation LM



 when moving to in-domain LM, absolute WER reduction due to improved segmentation is smaller by approximately 10%rel

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○●○○ | Conclusions |
|--------------|--------------|--------------------|-----------|--------------------------|-------------|
| | | | | | |

Results: MFCC.1 \rightarrow CNC.1



 when moving to higher complexity front end, WER reduction due to improved segmentation is smaller by approximately 10%rel

< 🗇 > < 🗆 >

4 王
| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions |
|--------------|--------------|-------------|-----------|-------------|-------------|
| | | | | | |

Results: MFCC.1 \rightarrow CNC.1



 when moving to higher complexity front end, WER reduction due to improved segmentation is smaller by approximately 10%rel

| Introduction | Segmentation | Recognition | Data | Experiments | Conclusions |
|--------------|--------------|-------------|------|-------------|-------------|
| | | | | 0000000000 | |
| | | | | | |

Results: CNC.1 \rightarrow CNC.2 \rightarrow CNC.3



 when moving across ASR passes, relative WER reduction due to improved segmentation is approximately constant

▲ 御 ▶ → ミ ▶

< E

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○○●○ | Conclusions |
|--------------|--------------|-------------|-----------|--------------------------|-------------|
| | | | | | |

Results: CNC.1 \rightarrow CNC.2 \rightarrow CNC.3



 when moving across ASR passes, relative WER reduction due to improved segmentation is approximately constant

▲ □ ► < □ ►</p>

< E

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○○●○ | Conclusions |
|--------------|--------------|-------------|-----------|--------------------------|-------------|
| | | | | | |

Results: $CNC.1 \rightarrow CNC.2 \rightarrow CNC.3$



 when moving across ASR passes, relative WER reduction due to improved segmentation is approximately constant

A (1) > (1)

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○○○● | Conclusions O |
|--------------|--------------|-------------|-----------|--------------------------|------------------|
| | | | | | |



• eval set was overall more difficult

- 27% WER versus 25.3% WER in CNC.3
- trends noted for the dev set almost identical
- except MFCC.1 \rightarrow CNC.1 WER on manual segmentation did not improve

イロト イヨト イヨト

-

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○○○● | Conclusions O |
|--------------|--------------|-------------|-----------|--------------------------|------------------|
| | | | | | |



- eval set was overall more difficult
- 27% WER versus 25.3% WER in CNC.3
- trends noted for the dev set almost identical
- except MFCC.1 \rightarrow CNC.1 WER on manual segmentation did not improve

A (10) A (10) A (10) A

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○○○● | Conclusions O |
|--------------|--------------|-------------|-----------|--------------------------|------------------|
| | | | | | |



- eval set was overall more difficult
- 27% WER versus 25.3% WER in CNC.3
- trends noted for the dev set almost identical
- except MFCC.1 \rightarrow CNC.1 WER on manual segmentation did not improve

イロト イボト イヨト イヨト

| Introduction | Segmentation | Recognition | Data ○ | Experiments ○○○○○○○○● | Conclusions O |
|--------------|--------------|-------------|-----------|--------------------------|------------------|
| | | | | | |



- eval set was overall more difficult
- 27% WER versus 25.3% WER in CNC.3
- trends noted for the dev set almost identical
- except MFCC.1 \rightarrow CNC.1 WER on manual segmentation did not improve

| Introduction | Segmentation | Recognition | Data ○ | Experiments | Conclusions • |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Conclusi | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants.
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data.
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:

| Introduction | Segmentation | Recognition 00 | Data ○ | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusi | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:

マロト マラト マラ

| Introduction | Segmentation | Recognition 00 | Data ○ | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusi | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:

| Introduction | Segmentation | Recognition | Data o | Experiments | Conclusions • |
|--------------|--------------|--------------------|-----------|-------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:

| Introduction | Segmentation | Recognition | Data o | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:

| Introduction | Segmentation | Recognition | Data o | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes

• to unseen evaluation data

• WER error rates with automatic segmentation are currently higher than with manual segmentation by:

3.0% absolute on the eval set.

similar to results published by others

| Introduction | Segmentation | Recognition | Data o | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:
 - 3.3% absolute on the dev set
 - 3.0% absolute on the eval set.
 - similar to results published by others

| Introduction | Segmentation | Recognition | Data o | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:
 - 3.3% absolute on the dev set
 - 3.0% absolute on the eval set
 - similar to results published by others

| Introduction | Segmentation | Recognition 00 | Data o | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:
 - 3.3% absolute on the dev set
 - 3.0% absolute on the eval set
 - similar to results published by others

| Introduction | Segmentation | Recognition 00 | Data o | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:
 - 3.3% absolute on the dev set
 - 3.0% absolute on the eval set
 - similar to results published by others

| Introduction | Segmentation | Recognition 00 | Data o | Experiments 00000000000 | Conclusions • |
|--------------|--------------|--------------------------|-----------|----------------------------|------------------|
| | | | | | |
| Conclusio | ons | | | | |

- modeling overlap is an effective means of suppressing crosstalk in meetings
- the proposed system improvements appear to generalize
 - to scenarios with uninstrumented participants
 - to improved language models in ASR
 - to more complex front ends in ASR
 - across ASR adaptation passes
 - to unseen evaluation data
- WER error rates with automatic segmentation are currently higher than with manual segmentation by:
 - 3.3% absolute on the dev set
 - 3.0% absolute on the eval set
 - similar to results published by others