Simultaneous Multispeaker Segmentation for Automatic Meeting Recognition

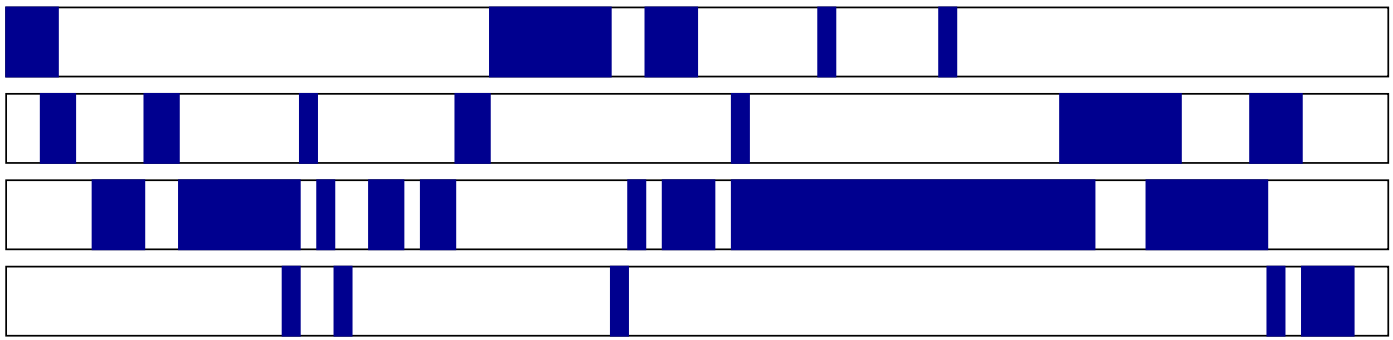
Kornel Laskowski[1,2], Christian Fügen[2] & Tanja Schultz[1,2]

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

September 6, 2007

Goal

Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.

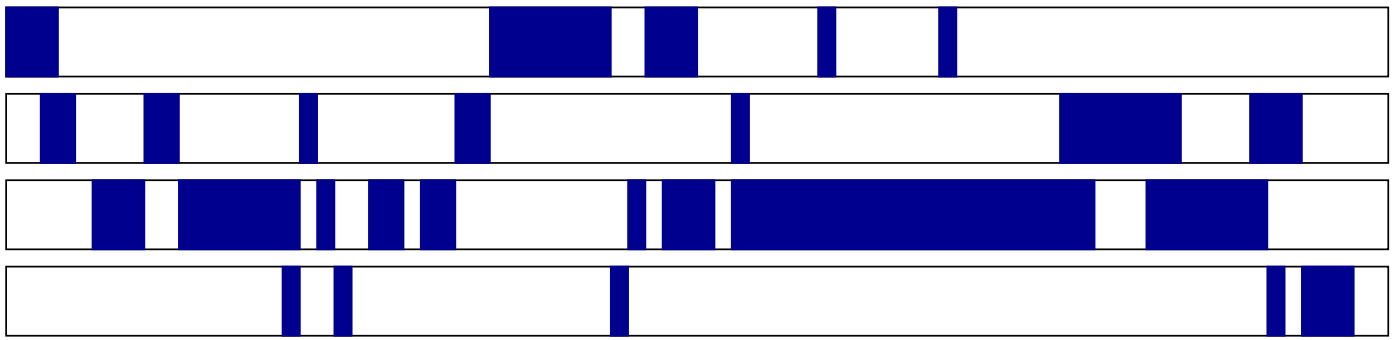


Why?

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

Goal

Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.

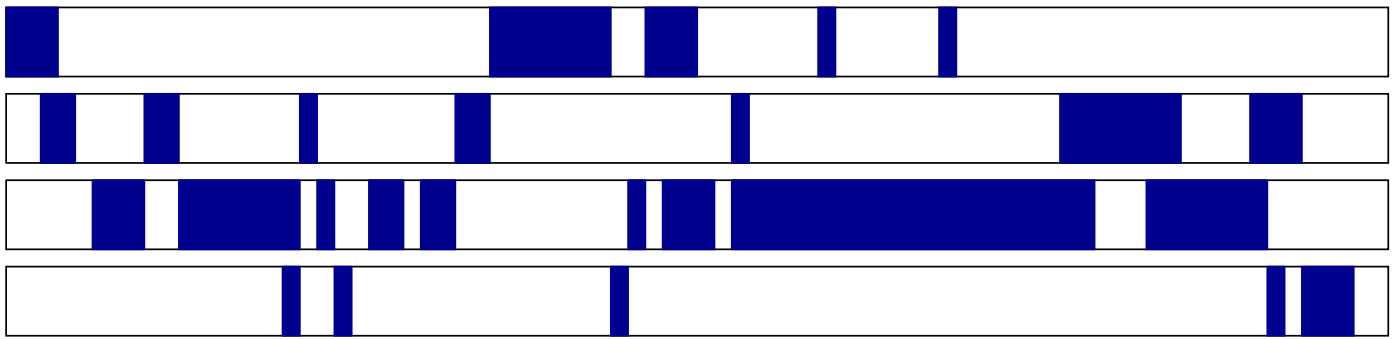


Why?

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

Goal

Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.

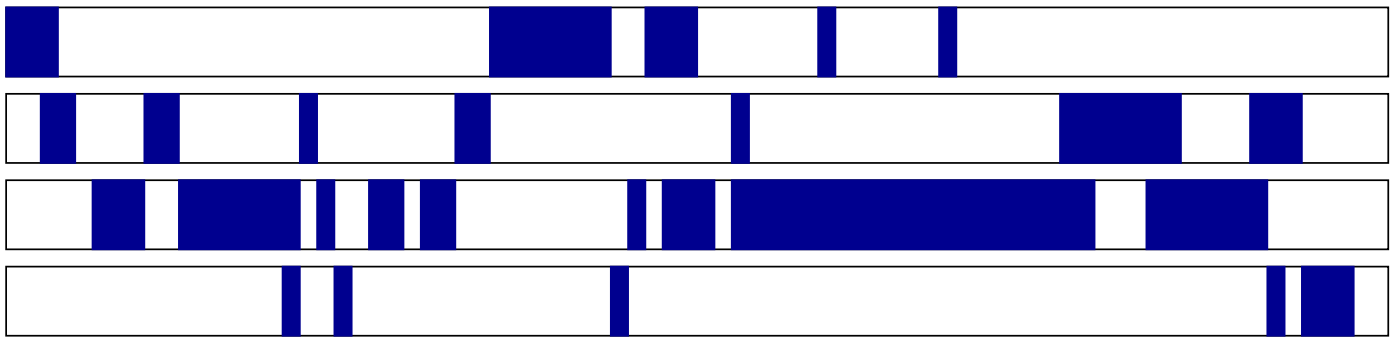


Why?

- syntactic continuity $\longrightarrow$ can leverage language modeling
- speaker homogeneity $\longrightarrow$ can leverage speaker adaptation

Goal

Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.

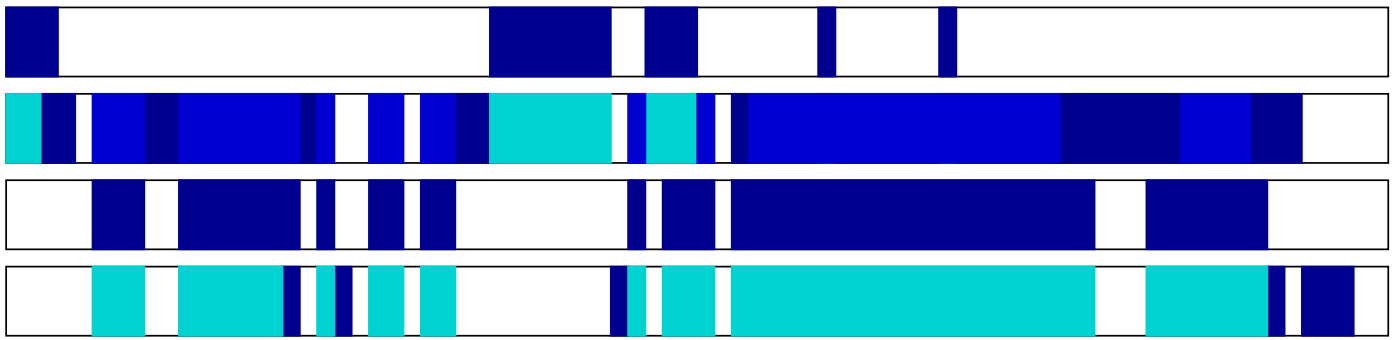


Why?

- syntactic continuity $\longrightarrow$ can leverage language modeling
- speaker homogeneity $\longrightarrow$ can leverage speaker adaptation

Main Problem: Crosstalk

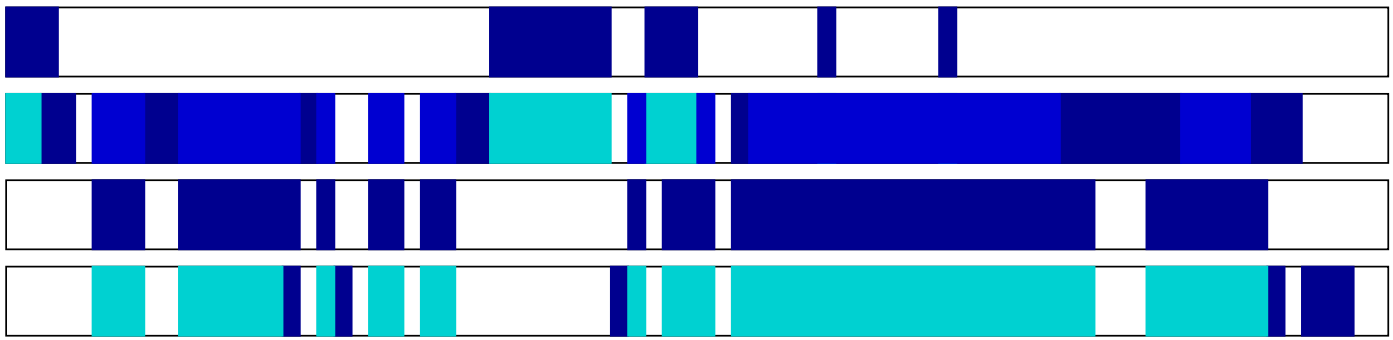
Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.



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

Main Problem: Crosstalk

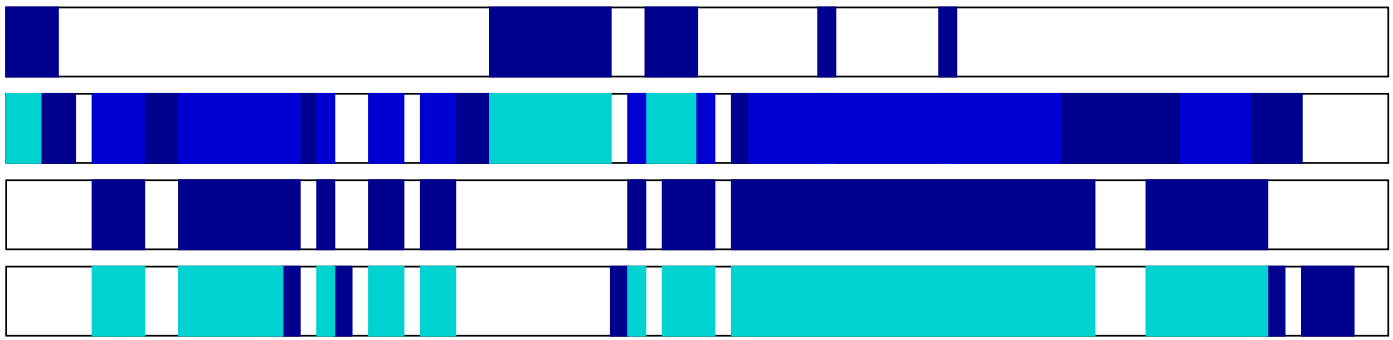
Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.



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

Main Problem: Crosstalk

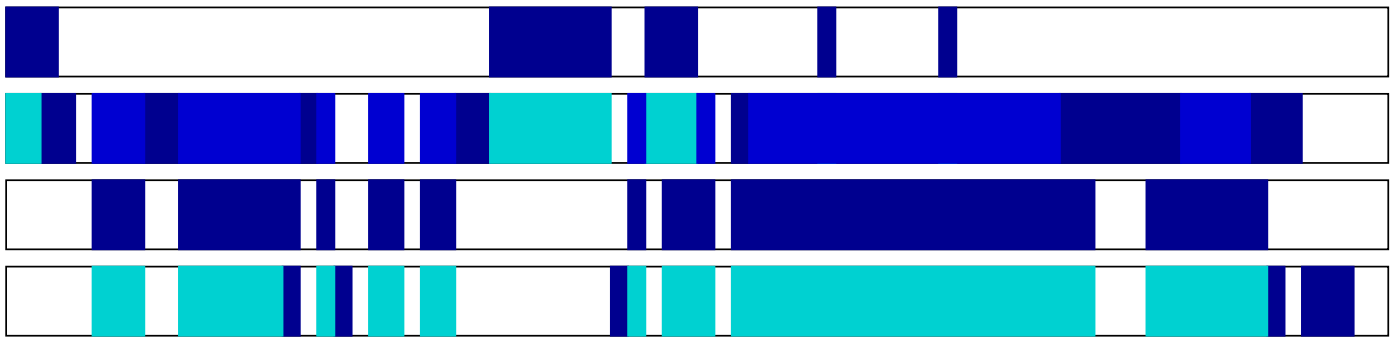
Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.



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

Main Problem: Crosstalk

Automatic speech/non-speech segmentation in close-talk microphone recordings of meetings.



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

A Note on Terms

- **crosstalk**
  - 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)

A Note on Terms

- **crosstalk**
  - 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)

A Note on Terms

- **crosstalk**
  - 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)

A Note on Terms

- **crosstalk**
  - 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
  - relatively infrequent
  - occurs $\approx 10\%$ of the time (Cetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of **overlap** to account for **crosstalk** (Laskowski & Schultz, MLMI07)

A Note on Terms

- **crosstalk**
  - 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
  - relatively infrequent
  - occurs $\approx 10\%$ of the time (Cetin & Shriberg, MLMI06)
- our approach: use statistics of the occurrence of **overlap** to account for **crosstalk** (Laskowski & Schultz, MLMI07)

A Note on Terms

- **crosstalk**
  - 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
  - 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 Note on Terms

- **crosstalk**
  - 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
  - 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 Note on Terms

- **crosstalk**
  - 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
  - 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 Note on Terms

- **crosstalk**
  - 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
  - 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 Note on Terms

- **crosstalk**
  - 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
  - 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 Note on Terms

- **crosstalk**
  - 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
  - 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)

How to Address Crosstalk

Elsewhere:

1. model participants **independently**

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

2. **supervised** acoustic models
3. **small** frame size (≈10 ms)

This work:

1. model participants **jointly**

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

2. **unsupervised** acoustic models
3. **large** frame size (≈100 ms)

How to Address Crosstalk

Elsewhere:

1. model participants **independently**

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

2. **supervised** acoustic models
3. **small** frame size (≈10 ms)

This work:

1. model participants **jointly**

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

2. **unsupervised** acoustic models
3. **large** frame size (≈100 ms)

How to Address Crosstalk

Elsewhere:

1. model participants **independently**

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

2. **supervised** acoustic models
3. **small** frame size (≈10 ms)

This work:

1. model participants **jointly**

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

2. **unsupervised** acoustic models
3. **large** frame size (≈100 ms)

How to Address Crosstalk

Elsewhere:

1. model participants **independently**

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

2. **supervised** acoustic models
3. **small** frame size (≈10 ms)

This work:

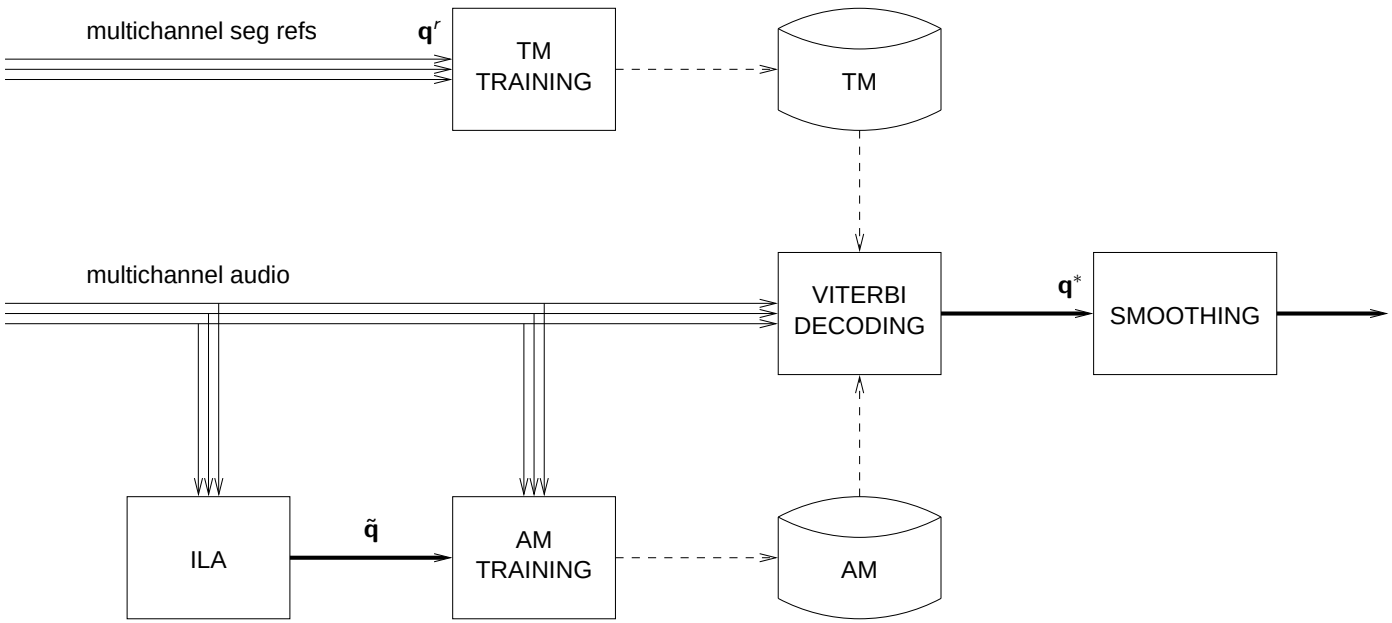
1. model participants **jointly**

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

2. **unsupervised** acoustic models
3. **large** frame size (≈100 ms)

Segmentation System Architecture

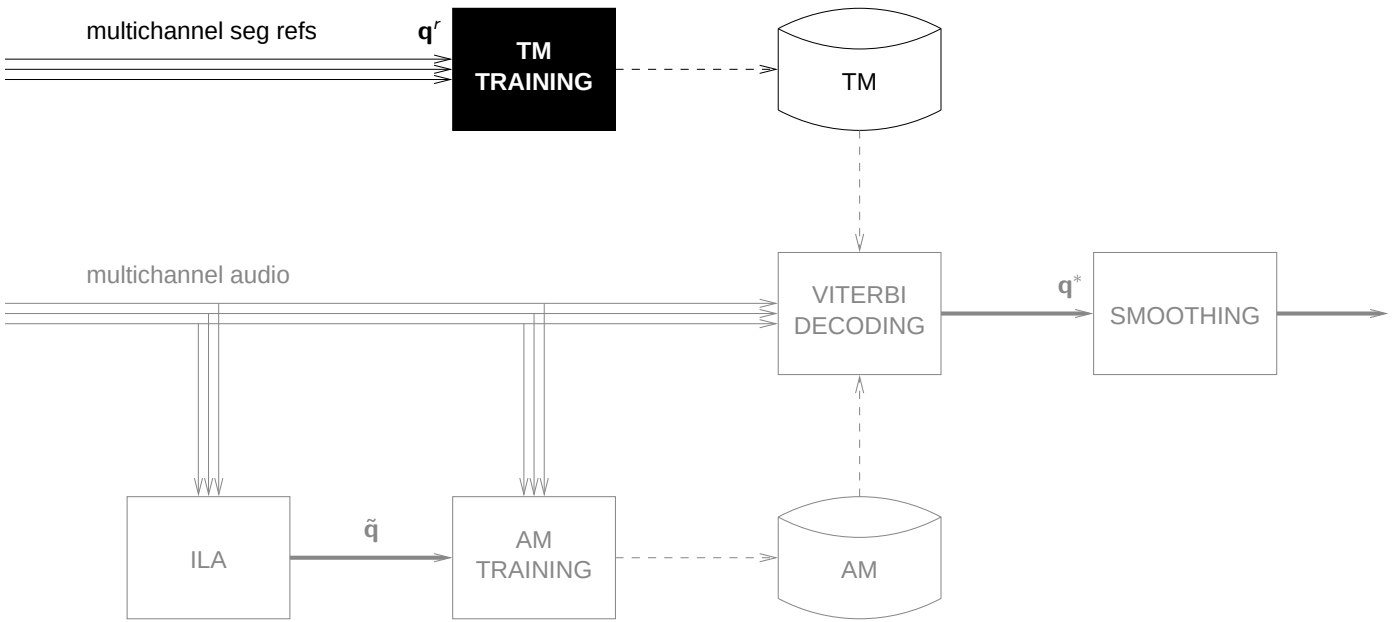
3 pass segmentation decoding of meetings



Segmentation System Architecture

3 pass segmentation decoding of meetings

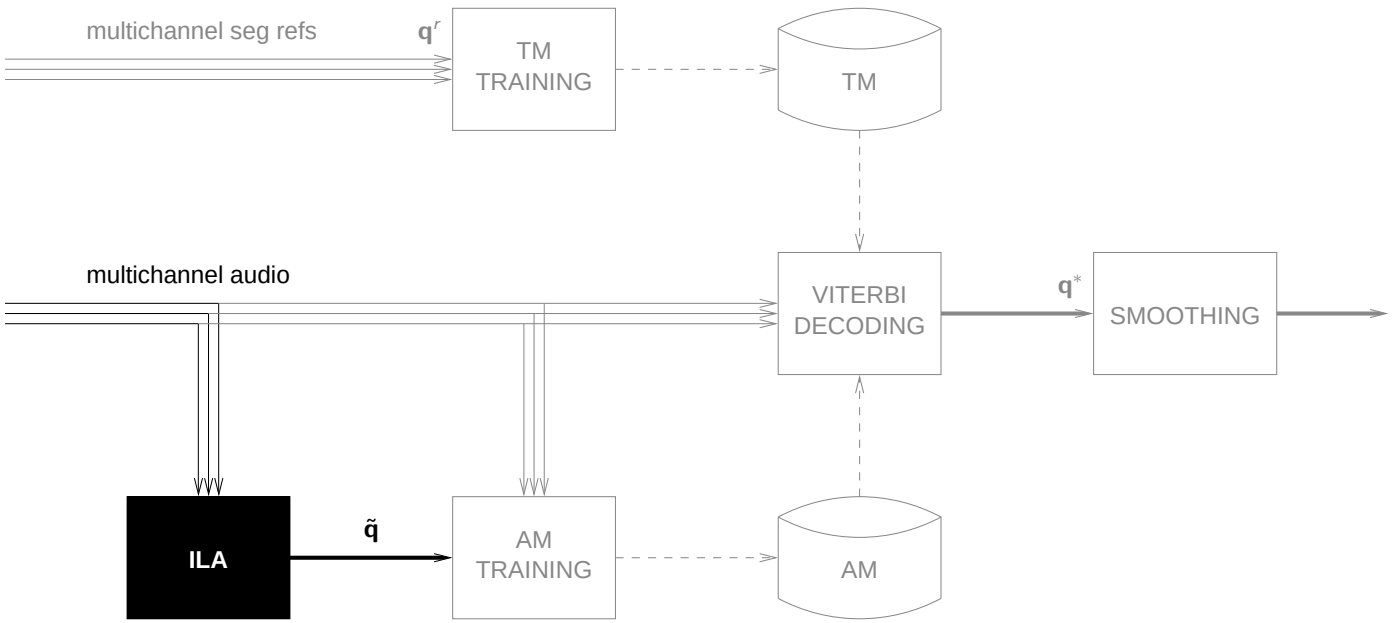
- prior to decoding: transition model (TM) training



Segmentation System Architecture

3 pass segmentation decoding of meetings

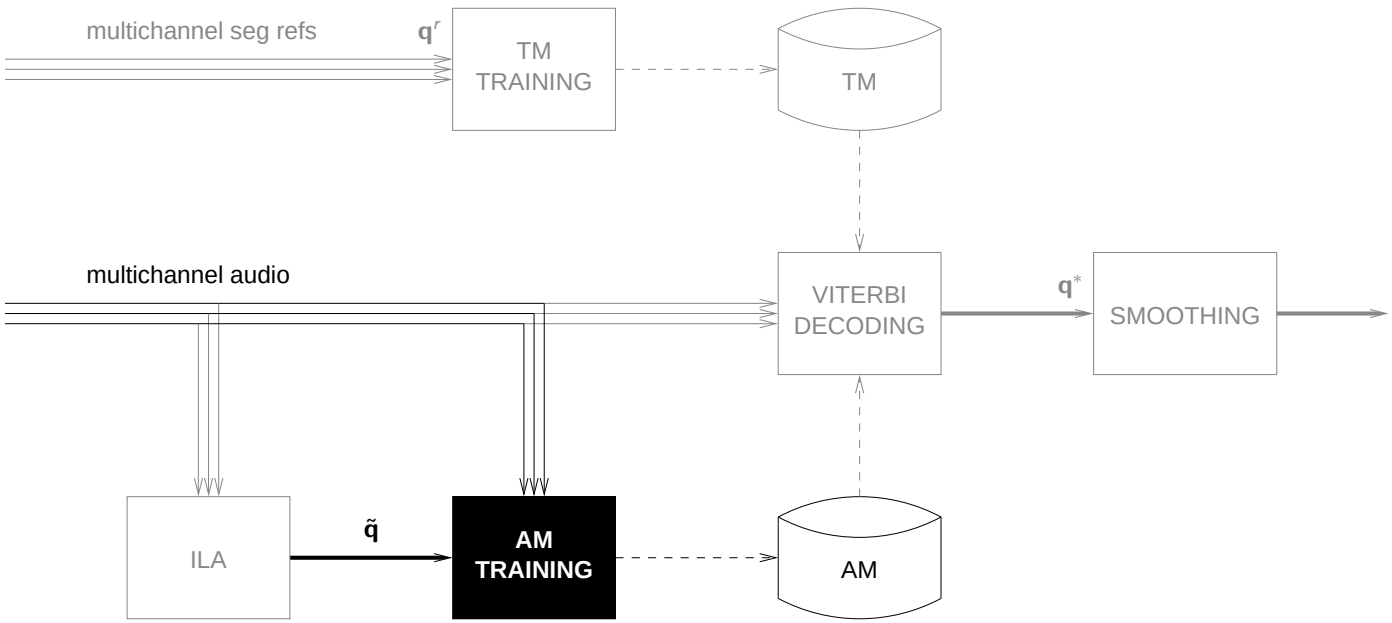
- **PASS 1**: initial label assignment



Segmentation System Architecture

3 pass segmentation decoding of meetings

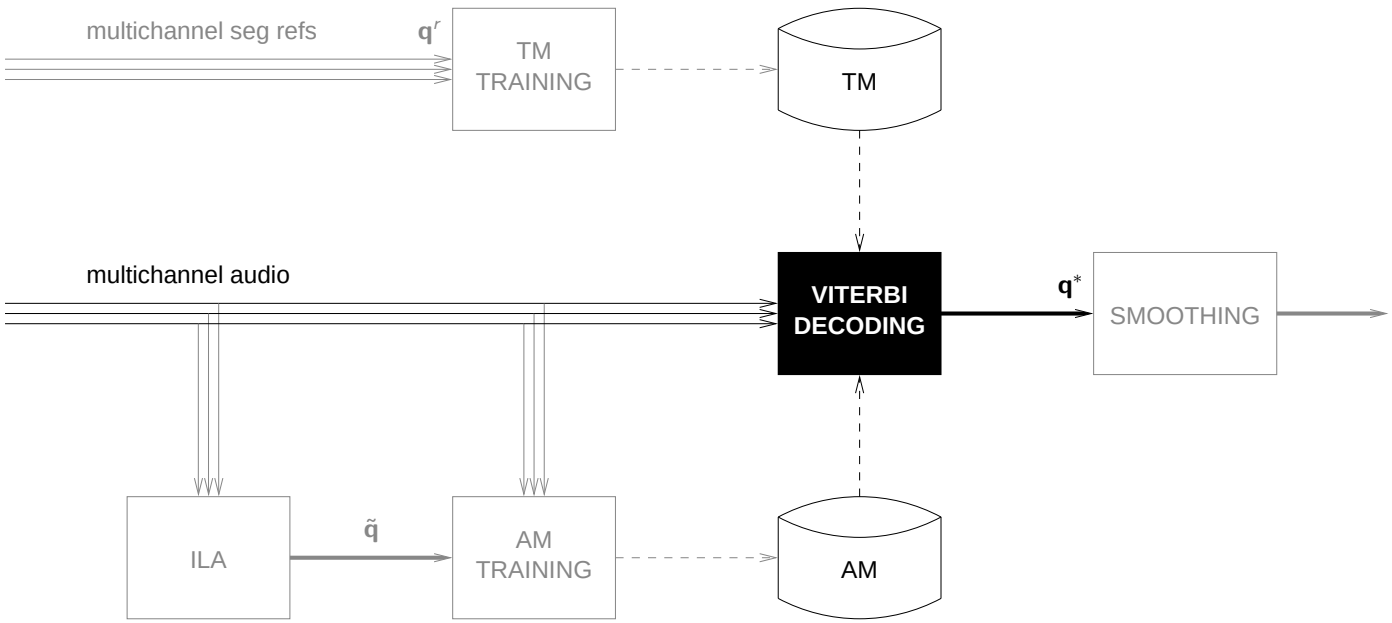
- acoustic model (AM) training



Segmentation System Architecture

3 pass segmentation decoding of meetings

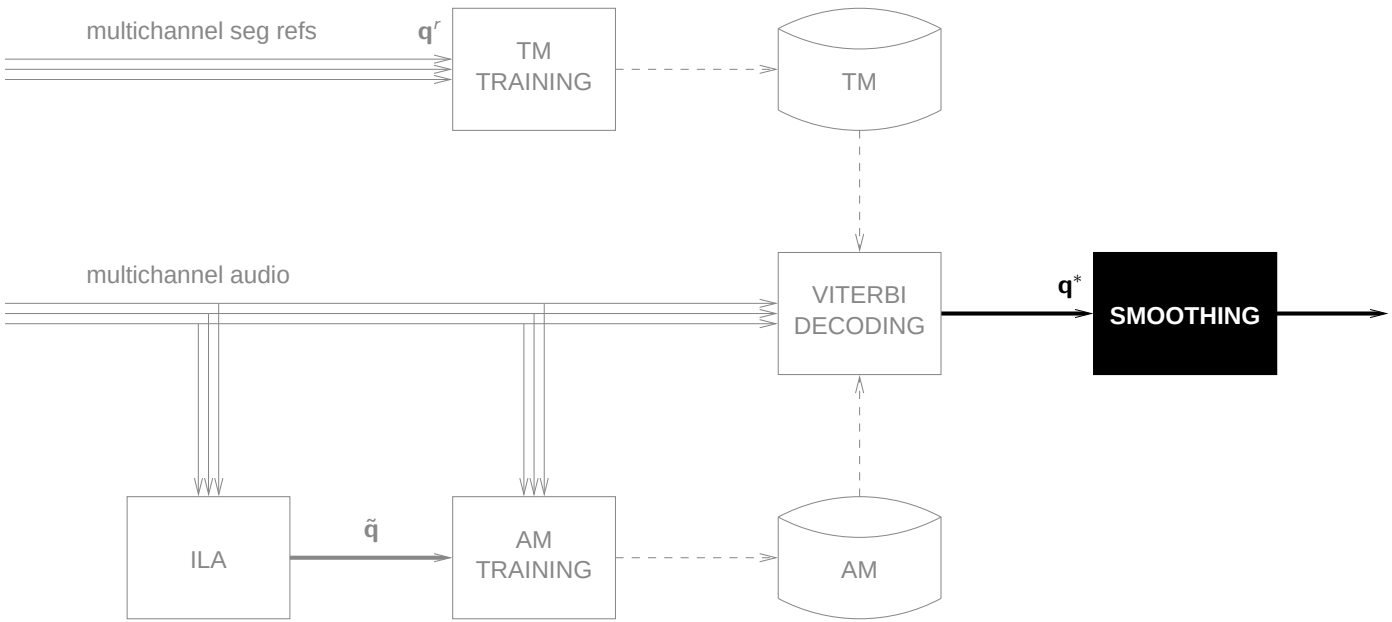
- **PASS 2:** Viterbi decoding



Segmentation System Architecture

3 pass segmentation decoding of meetings

- **PASS 3:** segmentation smoothing



Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- where a transition (n_i, o_{ij}, n_j)
- the Extended Degree of Overlap (EDO) model (Laskowski & Schultz, MLMI07)

Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- 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)

Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- 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)

Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- 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)

Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- 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)

Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- 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)

Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- 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)

Transition Model Training

- 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 $\mathbf{S}_i$ to $\mathbf{S}_j$

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

- 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)

PASS 1: Initial Label Assignment

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

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

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

PASS 1: Initial Label Assignment

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

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

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

PASS 1: Initial Label Assignment

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

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

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

PASS 1: Initial Label Assignment

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

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

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

PASS 1: Initial Label Assignment

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

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

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

PASS 1: Initial Label Assignment

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

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

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

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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_2)
  - sample-level synthesized overlap (λ_3)

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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)

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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)

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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)

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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)

Acoustic Model Training

- use the test audio and the initial labels from ILA to train
  - **one** Gaussian model
  - with a **full** covariance matrix
  - for a $2K$-length vector (energy and zero-crossing rate per channel) $\longrightarrow$ 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)

PASS 2: Viterbi Decoding

- 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: $1 \leq t \leq T$, $1 \leq i \leq N = 2^K$
  - assume all participants are silent at time $t = 0$
- following decoding, the state of each participant k at time t is

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

PASS 2: Viterbi Decoding

- 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: $1 \leq t \leq T$, $1 \leq i \leq N = 2^K$
  - assume all participants are silent at time $t = 0$
- following decoding, the state of each participant k at time t is

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

PASS 2: Viterbi Decoding

- 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: $1 \leq t \leq T$, $1 \leq i \leq N \equiv 2^K$
  - assume all participants are silent at time $t = 0$
- following decoding, the state of each participant k at time t is

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

PASS 2: Viterbi Decoding

- 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: $1 \leq t \leq T$, $1 \leq i \leq N \equiv 2^K$
  - assume all participants are silent at time $t = 0$
- following decoding, the state of each participant k at time t is

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

PASS 2: Viterbi Decoding

- 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: $1 \leq t \leq T$, $1 \leq i \leq N \equiv 2^K$
  - assume all participants are silent at time $t = 0$
- following decoding, the state of each participant k at time t is

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

PASS 2: Viterbi Decoding

- 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: $1 \leq t \leq T$, $1 \leq i \leq N \equiv 2^K$
  - assume all participants are silent at time $t = 0$
- following decoding, the state of each participant k at time t is

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

PASS 2: Viterbi Decoding

- 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: $1 \leq t \leq T$, $1 \leq i \leq N \equiv 2^K$
  - assume all participants are silent at time $t = 0$
- following decoding, the state of each participant k at time t is

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

PASS 3: Smoothing

For the Viterbi segmentation of each participant k:

1. 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
4. post-pad all $\mathcal{V}$ intervals with 300 ms
5. bridge remaining $\mathcal{V}$ gaps shorter than 400 ms
6. eliminate remaining $\mathcal{V}$ spurts shorter than 800 ms

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

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
  - a development ASR system: 1 pass, MFCC, non-adapt, out-of-domain LM
  - a more realistic ASR system: 3 passes, MFCC, fMLLR, MVDR, fast ends, in-domain LM

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 front-end, VTLN, fMLLR, CMLLR, in-domain LM

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

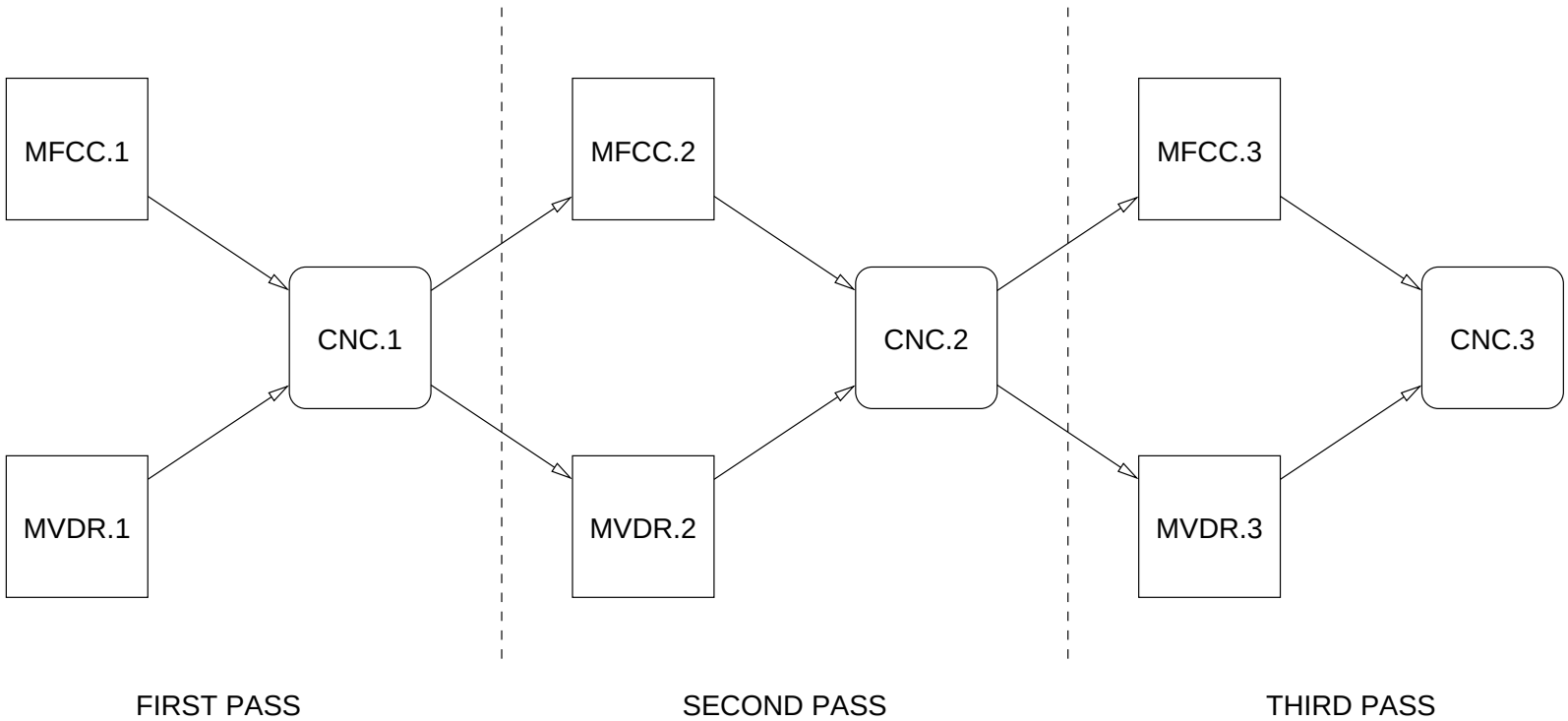
Evaluating Segmentation for Speech Recognition

- 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

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

3-Pass Recognition System Architecture



NIST Rich Transcription Meeting Recognition Corpora

development set

- 10 meeting excerpts
  - 2 × CMU
  - 2 × EDI
  - 2 × ICSI
  - 2 × NIST
  - 2 × 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

NIST Rich Transcription Meeting Recognition Corpora

development set

- 10 meeting excerpts
  - 2 × CMU
  - 2 × EDI
  - 2 × ICSI
  - 2 × NIST
  - 2 × 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

NIST Rich Transcription Meeting Recognition Corpora

development set

- 10 meeting excerpts
  - 2 × CMU
  - 2 × EDI
  - 2 × ICSI
  - 2 × NIST
  - 2 × 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

NIST Rich Transcription Meeting Recognition Corpora

development set

- 10 meeting excerpts
  - 2 × CMU
  - 2 × EDI
  - 2 × ICSI
  - 2 × NIST
  - 2 × 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

NIST Rich Transcription Meeting Recognition Corpora

development set

- 10 meeting excerpts
  - 2 × CMU
  - 2 × EDI
  - 2 × ICSI
  - 2 × NIST
  - 2 × 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

Baseline

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

Baseline

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

Baseline

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

Baseline

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

Modification 1: Eliminate Zero Crossing Rate (noZCR)

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

Modification 1: Eliminate Zero Crossing Rate (noZCR)

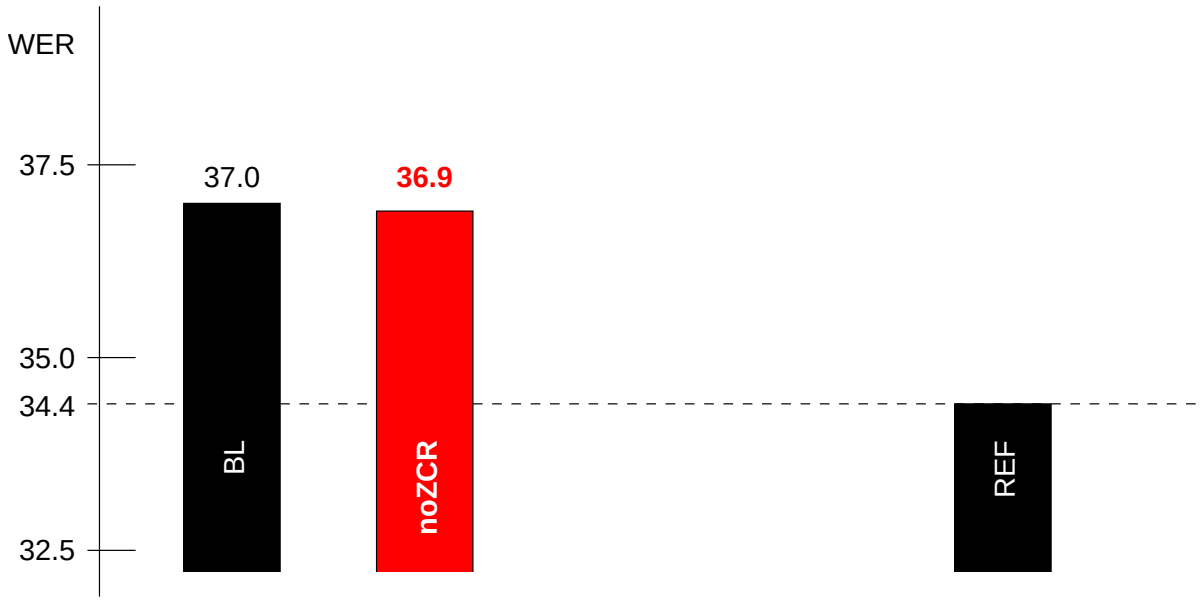
- remove the ZCR feature
  - reduces feature vector length from $2K$ to K features
- retune acoustic model training parameters (λ_G, λ_R, λ_S)

Modification 1: Eliminate Zero Crossing Rate (noZCR)

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

Modification 1: Eliminate Zero Crossing Rate (noZCR)

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



Modification 2: Reduce the Frame Size (F.100)

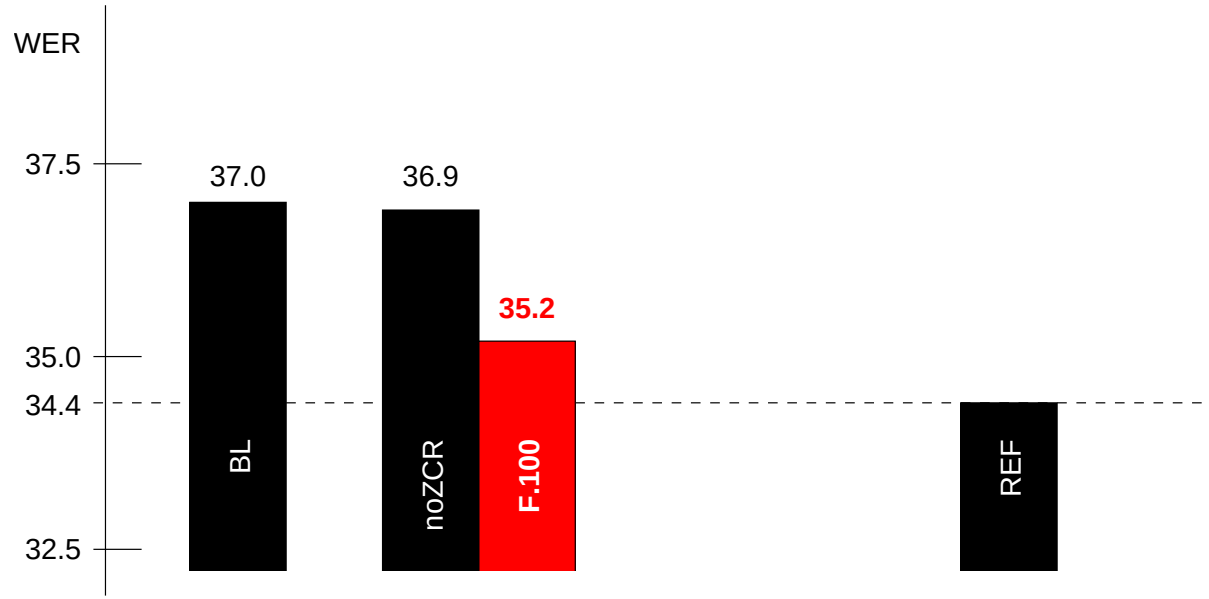
- reduce the frame size and frame step from 110ms to 100ms
- retune the smoothing pass parameters

Modification 2: Reduce the Frame Size (F.100)

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

Modification 2: Reduce the Frame 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)

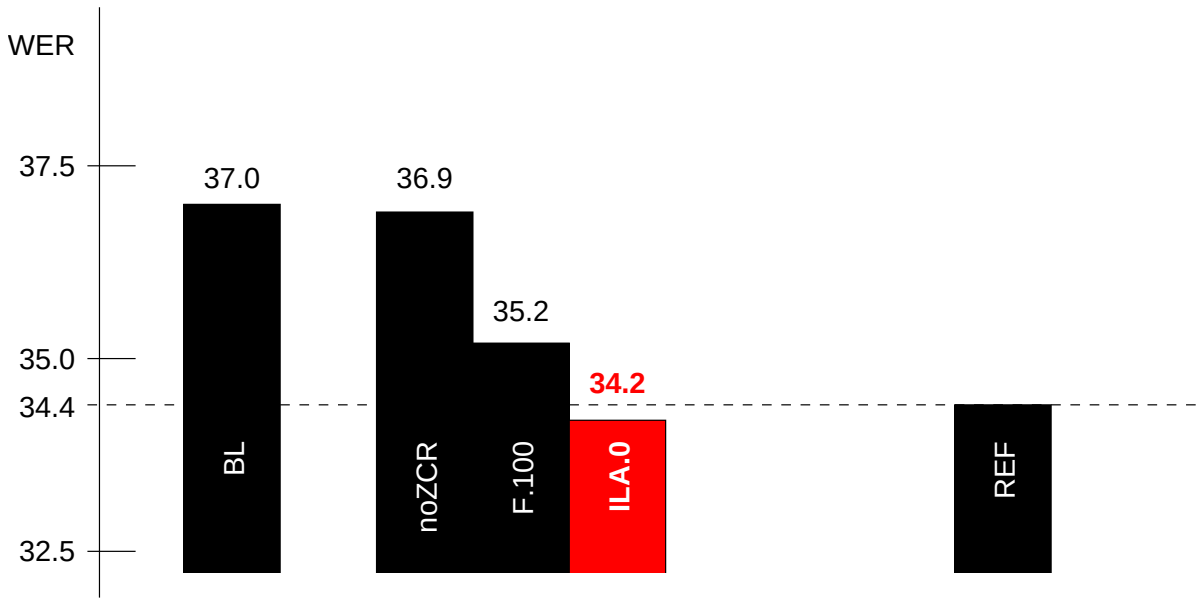
- OBSERVATION: ILA has high precision, but lower recall
- SOLUTION: use only (quieter) half of the frames classified as silence for training the silence model

Modification 3: Select Data for the Silence 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

Modification 3: Select Data for the Silence 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



Modification 4: Distribute EDO Mass (MULT)

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

Modification 4: Distribute EDO Mass (MULT)

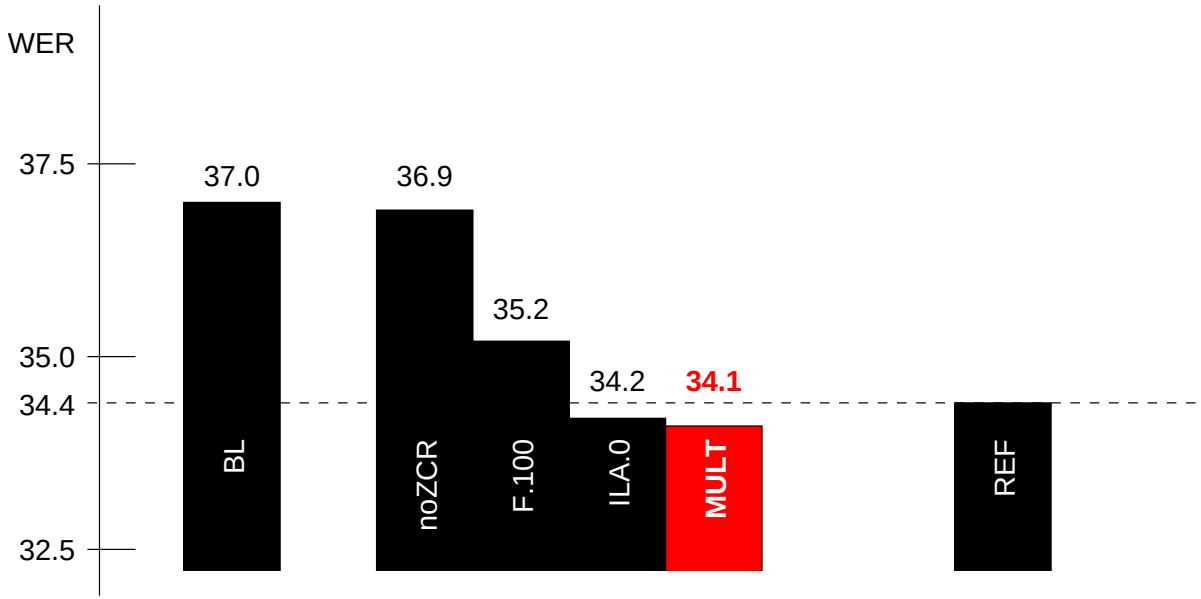
- OBSERVATION: EDO transitions have multiple target states
  - eg. (2,1,1): $\{A, B\} \rightarrow \{A\}$ **and** $\{A, B\} \rightarrow \{B\}$
- SOLUTION: distribute the probability mass among same-EDO next states with uniform probability

Modification 4: Distribute EDO Mass (MULT)

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

Modification 4: Distribute EDO Mass (MULT)

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

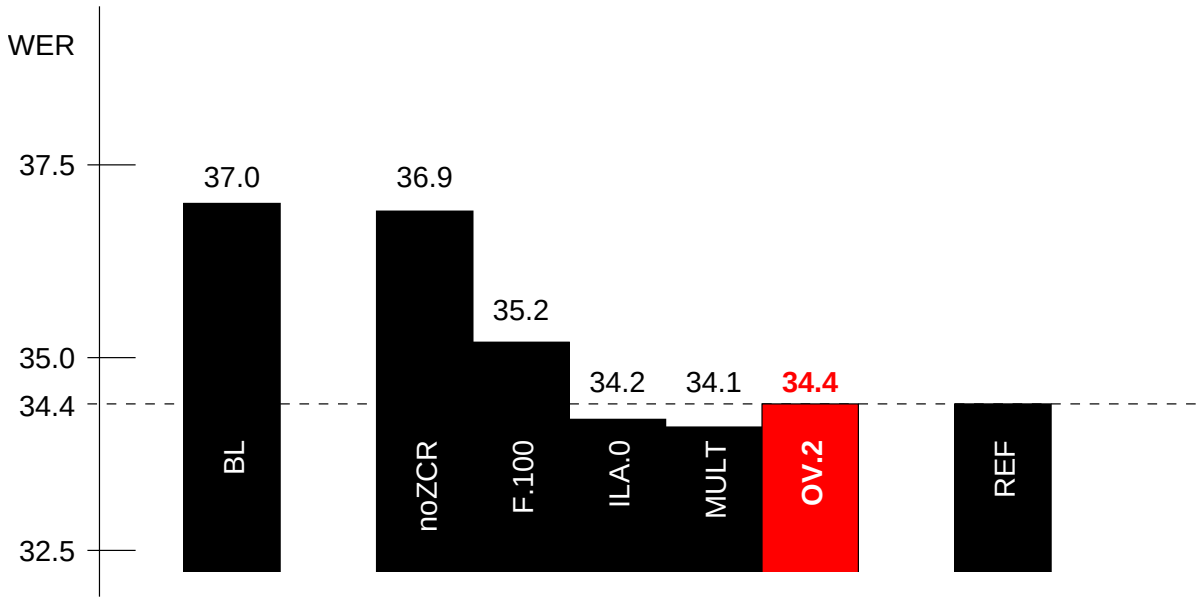


Modification 5: Limit Simultaneous Vocalization (OV.2)

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

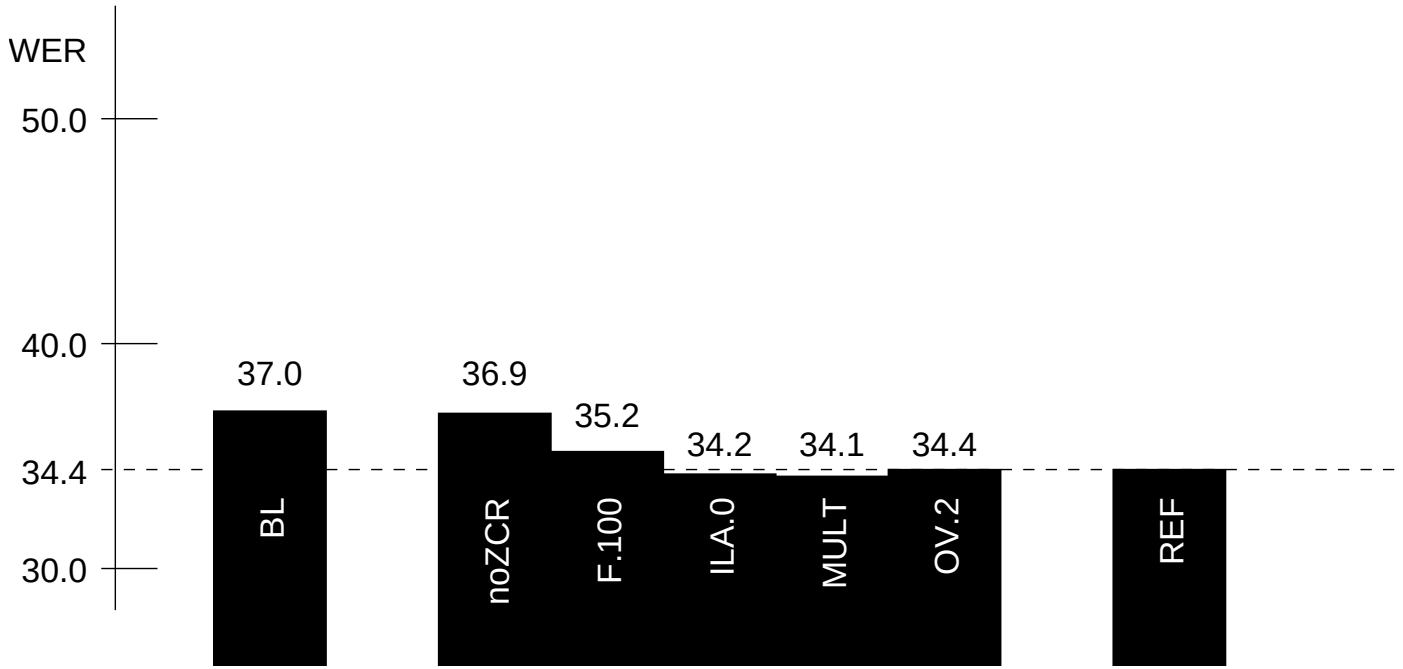
Modification 5: Limit Simultaneous Vocalization (OV.2)

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



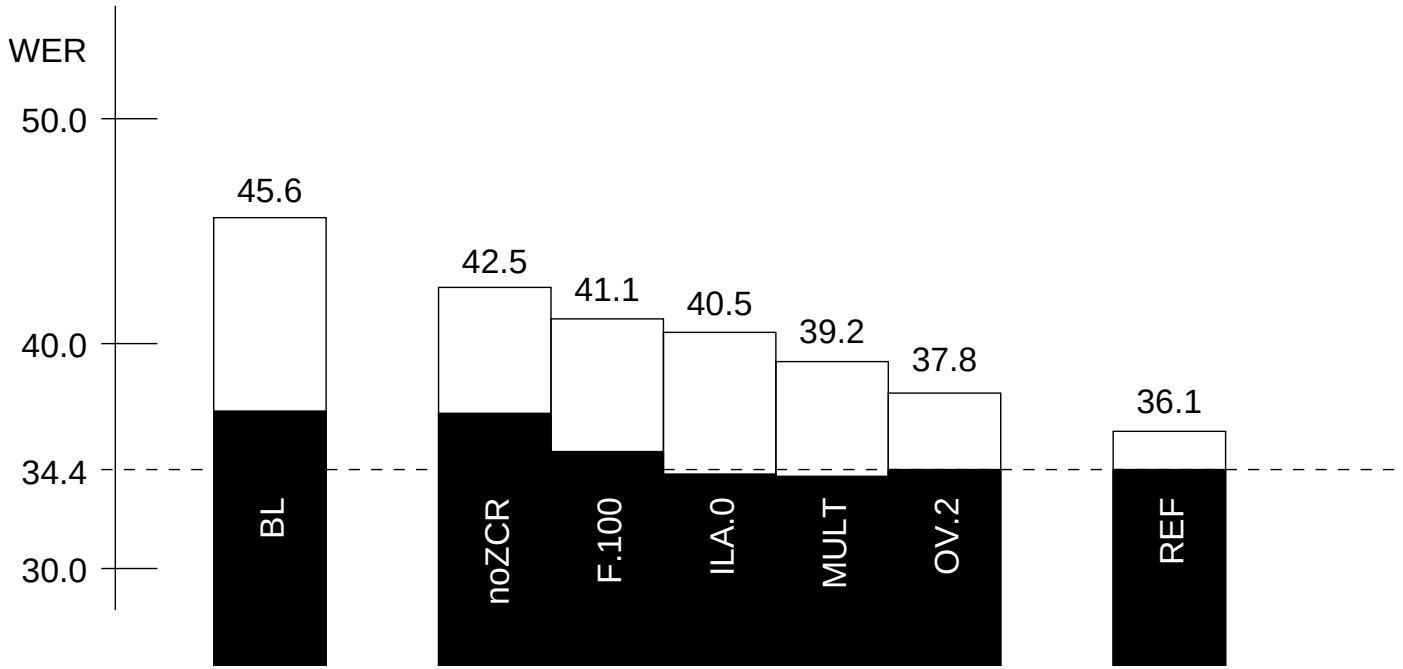
Results Using the Development ASR System

- 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%



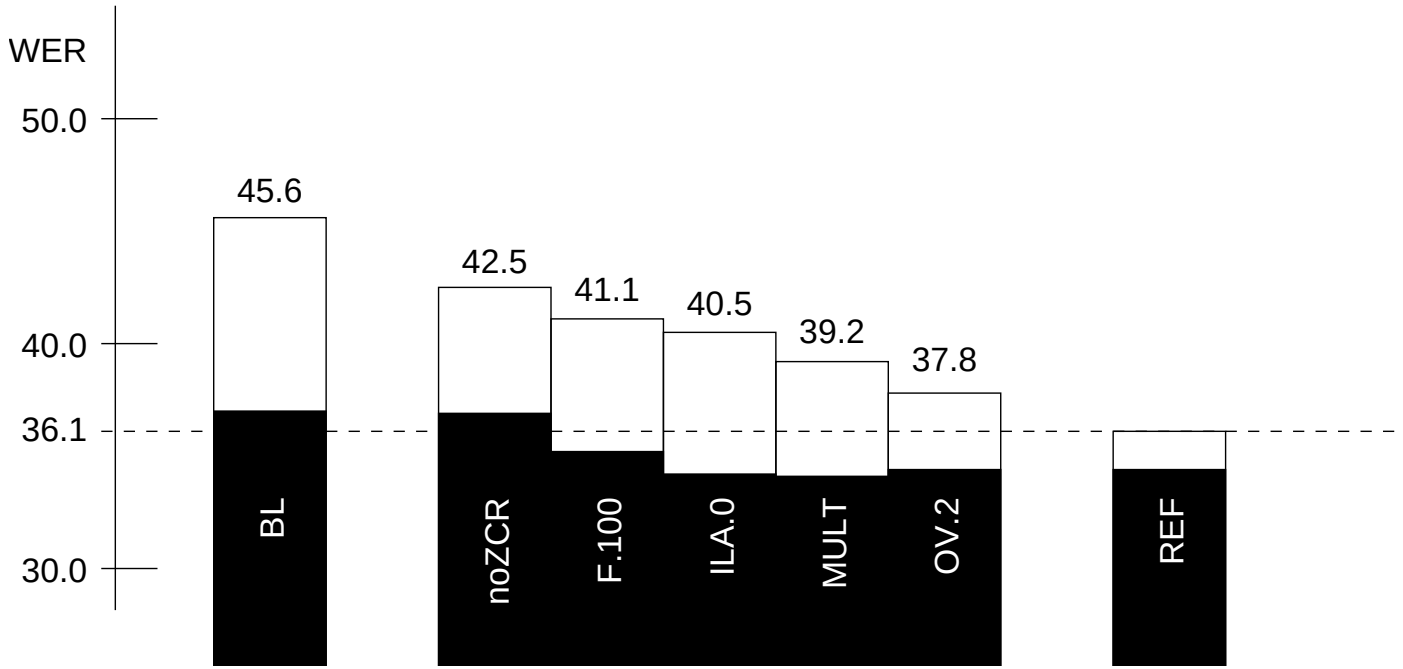
Results Using the Development ASR System

- 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%



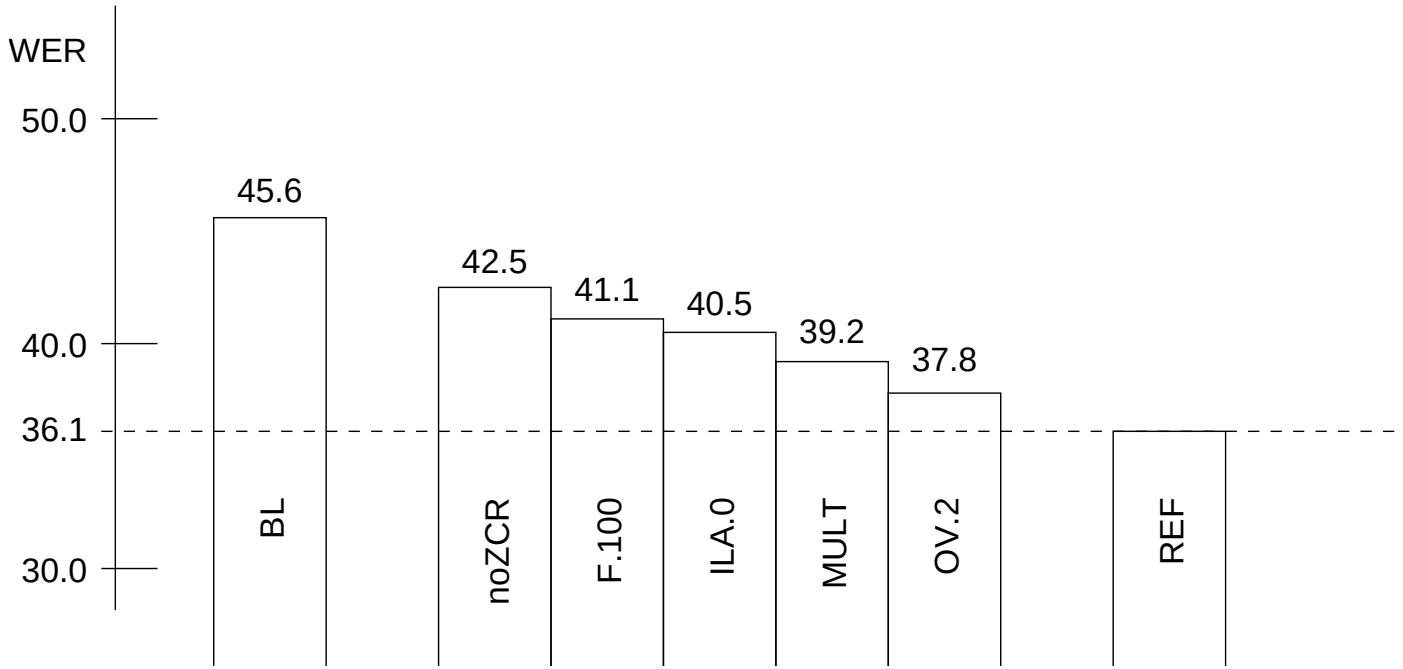
Results Using the Development ASR System

- 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%



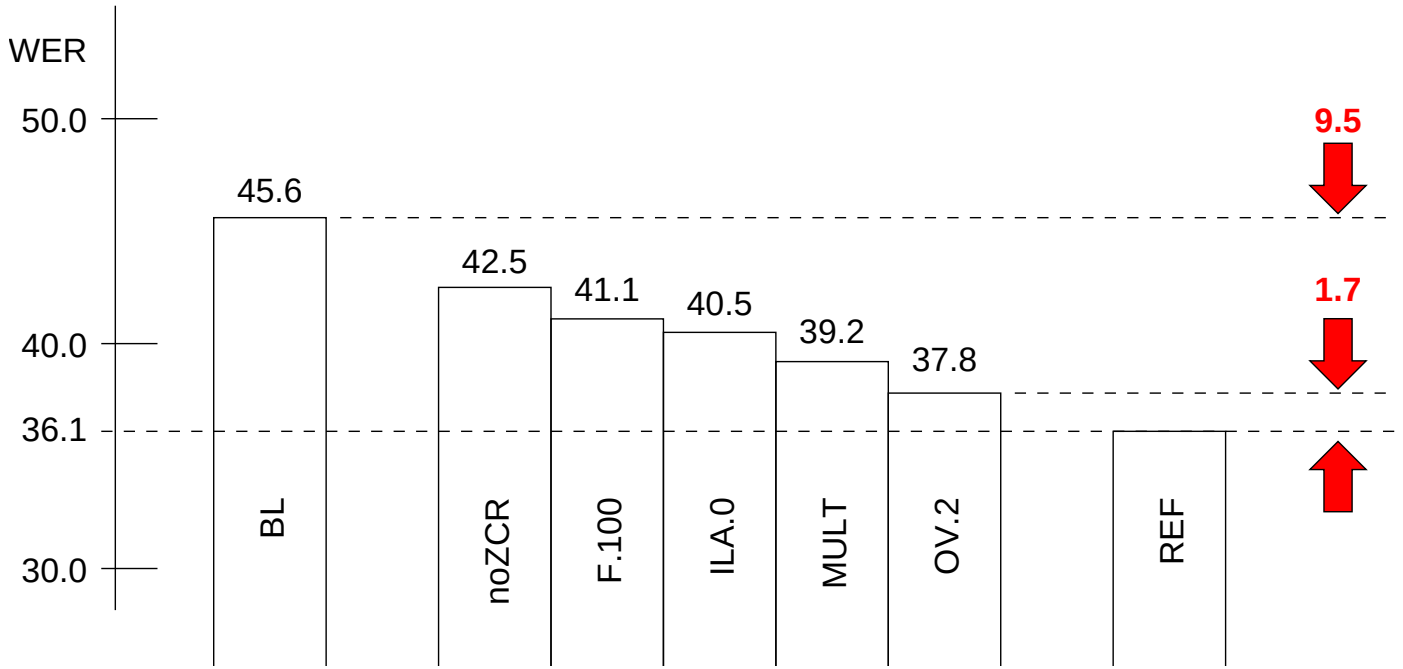
Results Using the Development ASR System

- 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%



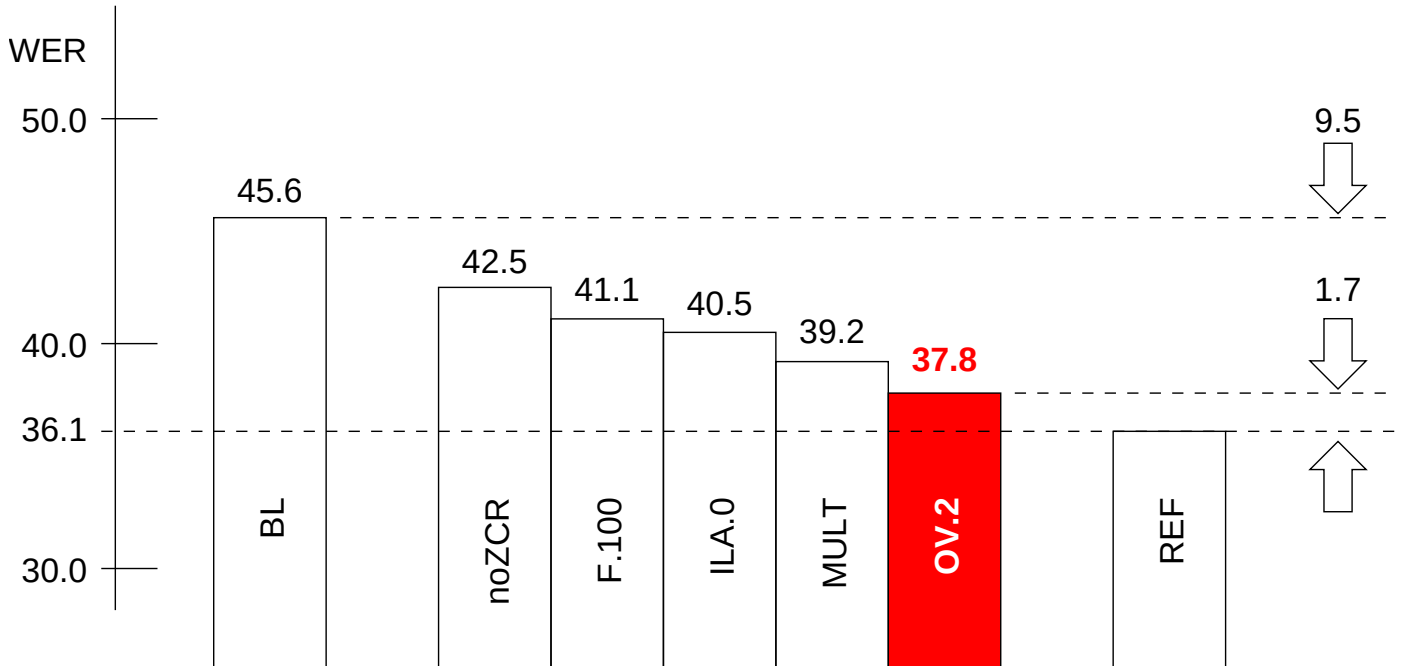
Results Using the Development ASR System

- 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%

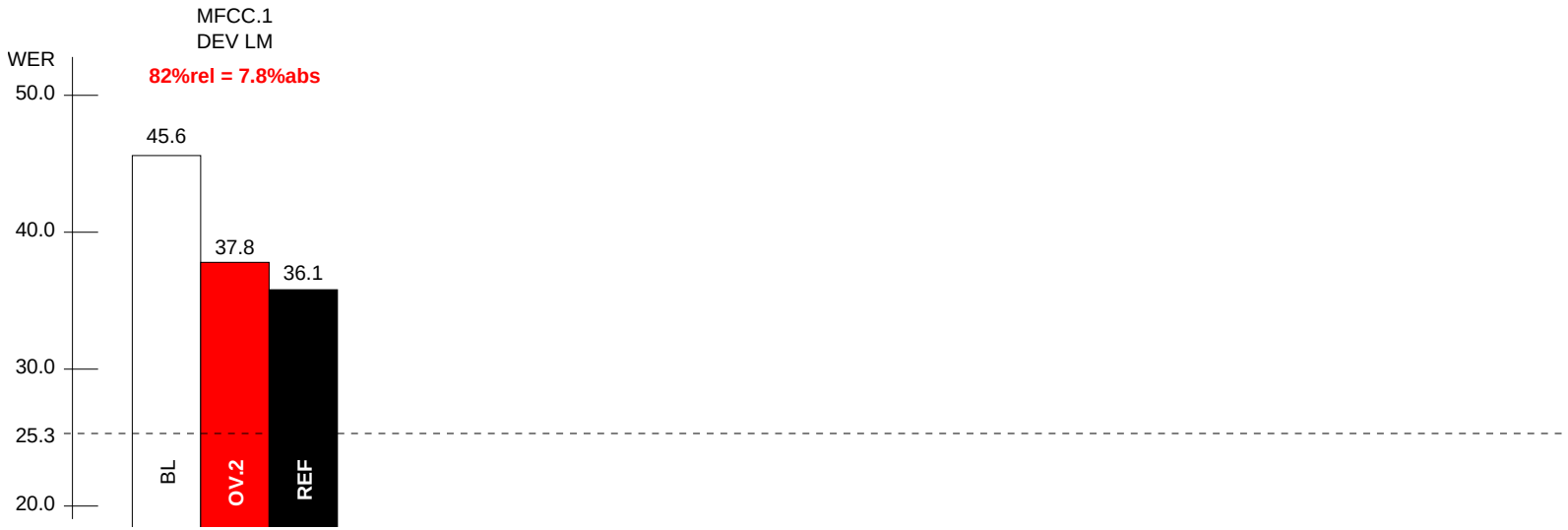


Results Using the Development ASR System

- 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%

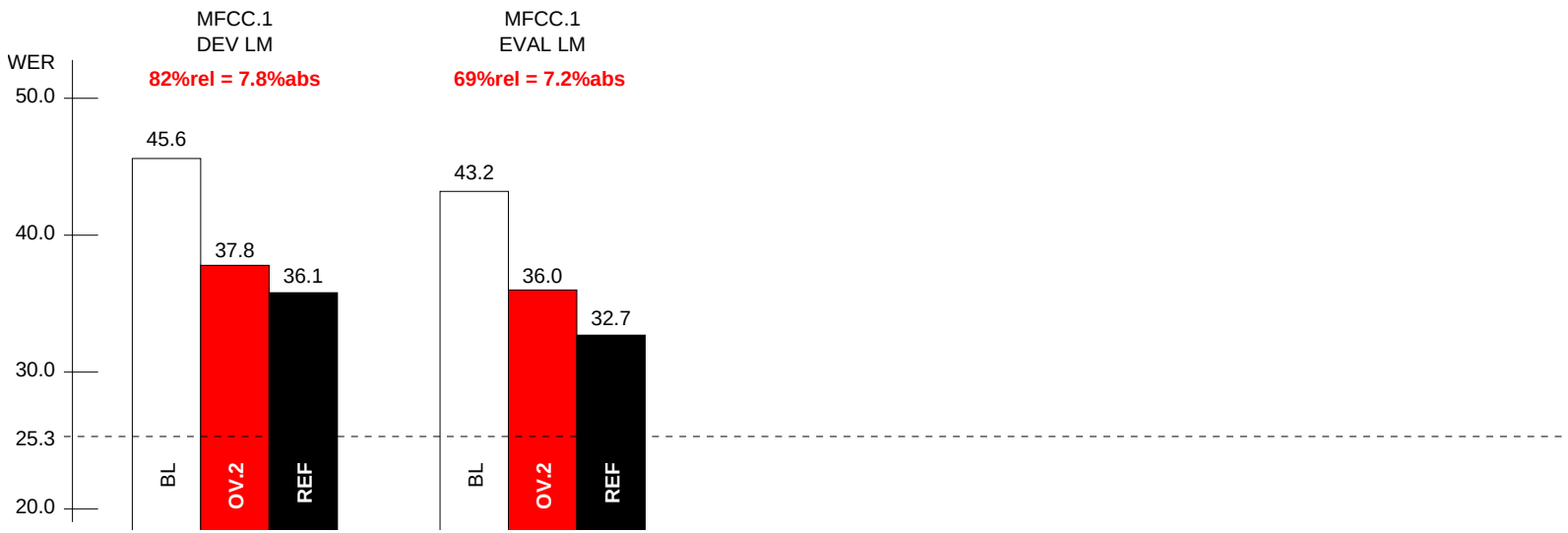


Results: Development LM → Evaluation LM



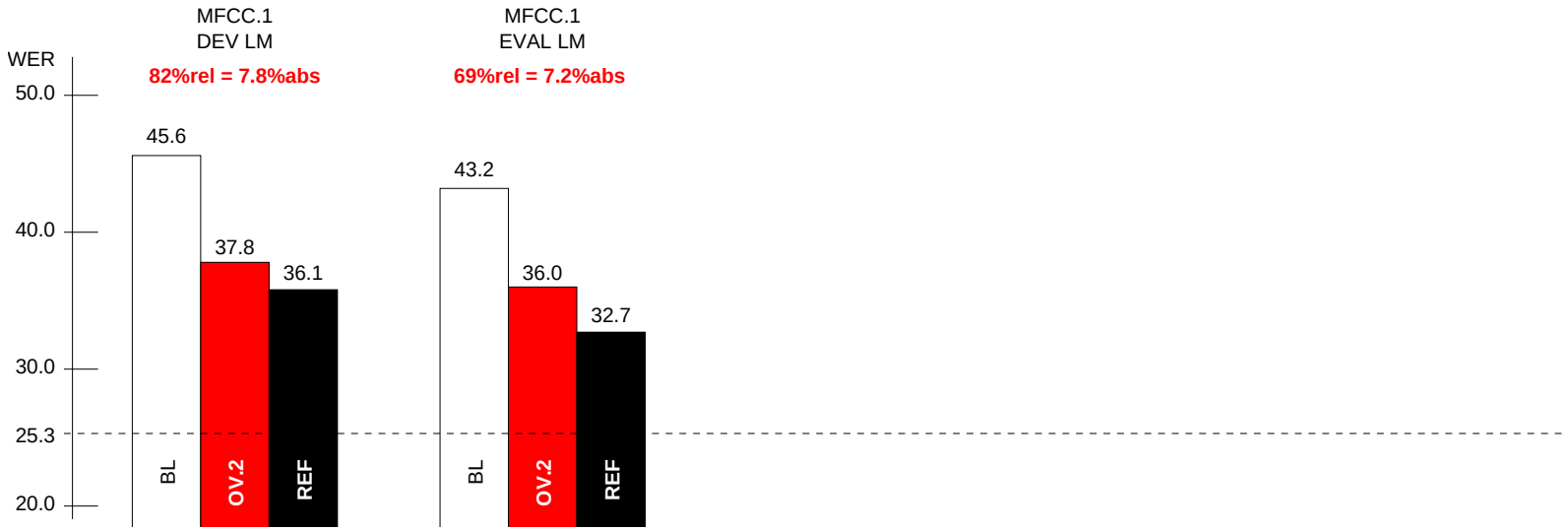
- when moving to in-domain LM, absolute WER reduction due to improved segmentation is smaller by approximately 10%rel

Results: Development LM → Evaluation LM



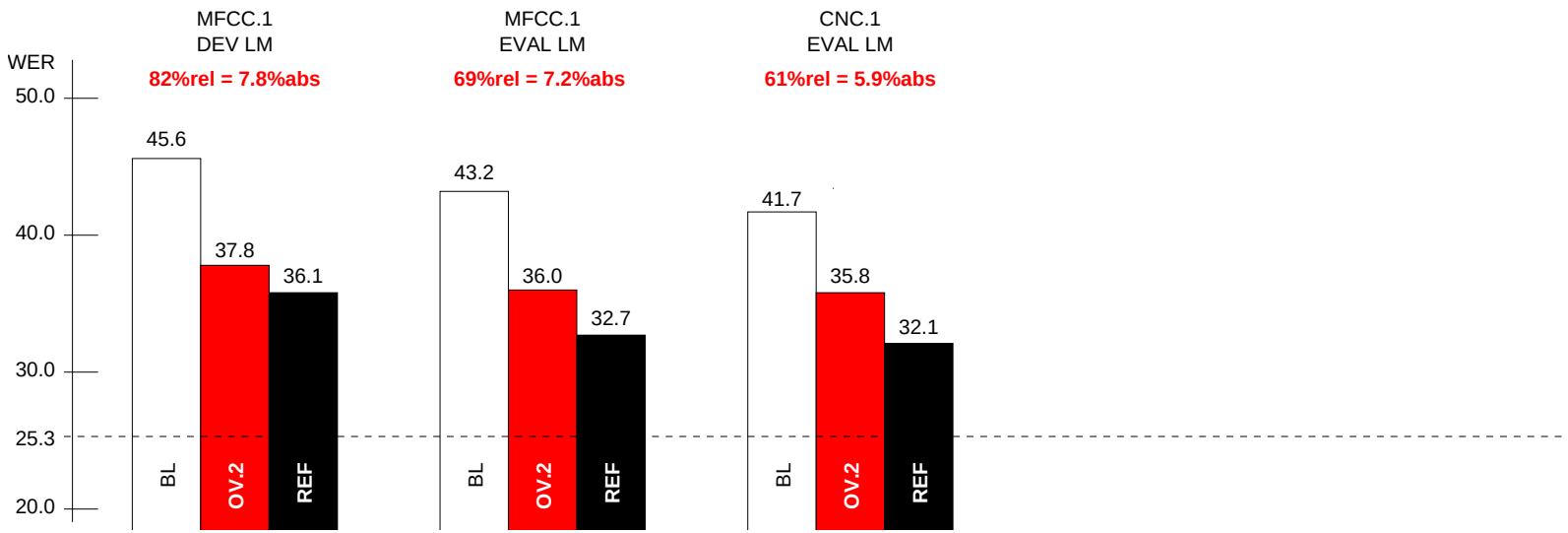
- when moving to in-domain LM, absolute WER reduction due to improved segmentation is smaller by approximately 10%rel

Results: MFCC.1 → CNC.1



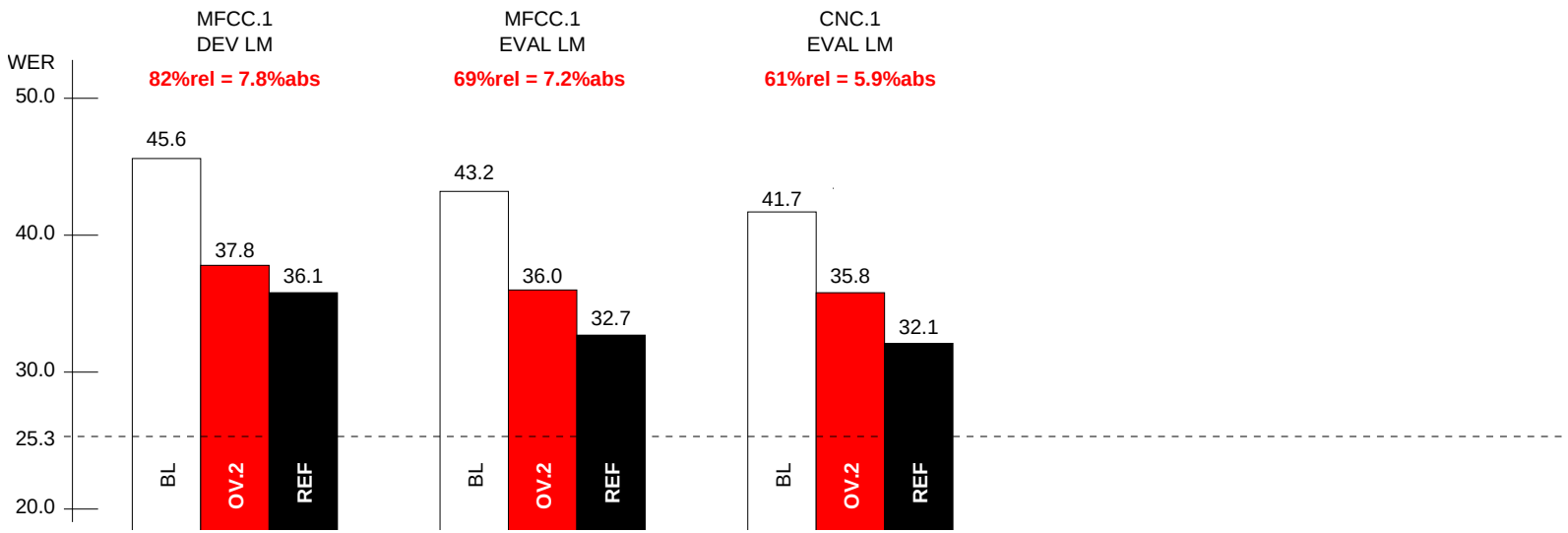
- when moving to higher complexity front end, WER reduction due to improved segmentation is smaller by approximately 10%rel

Results: MFCC.1 → CNC.1



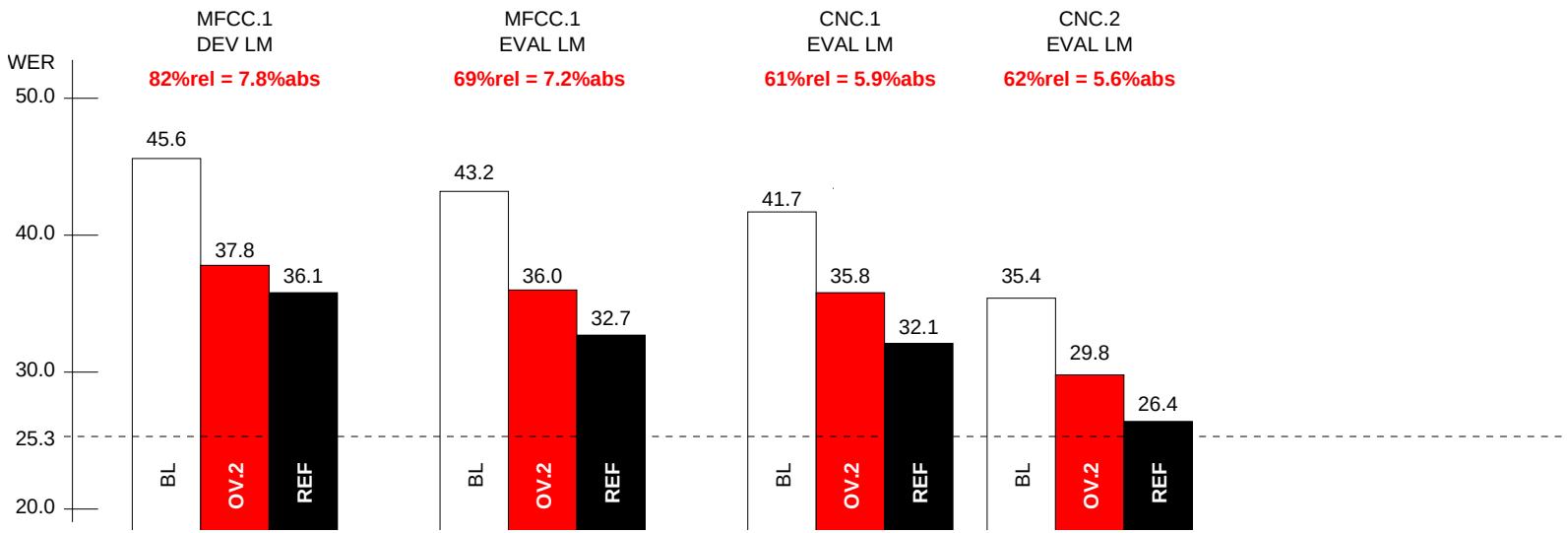
- when moving to higher complexity front end, WER reduction due to improved segmentation is smaller by approximately 10%rel

Results: CNC.1 → CNC.2 → CNC.3



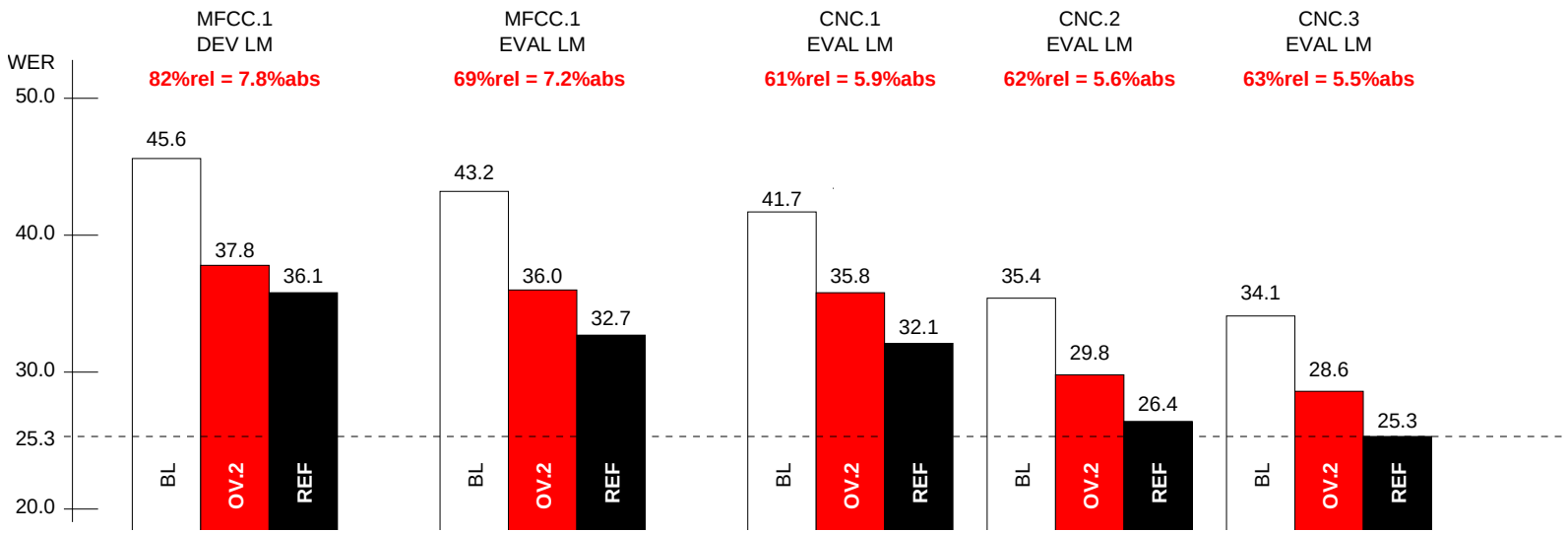
- when moving across ASR passes, relative WER reduction due to improved segmentation is approximately constant

Results: CNC.1 → CNC.2 → CNC.3



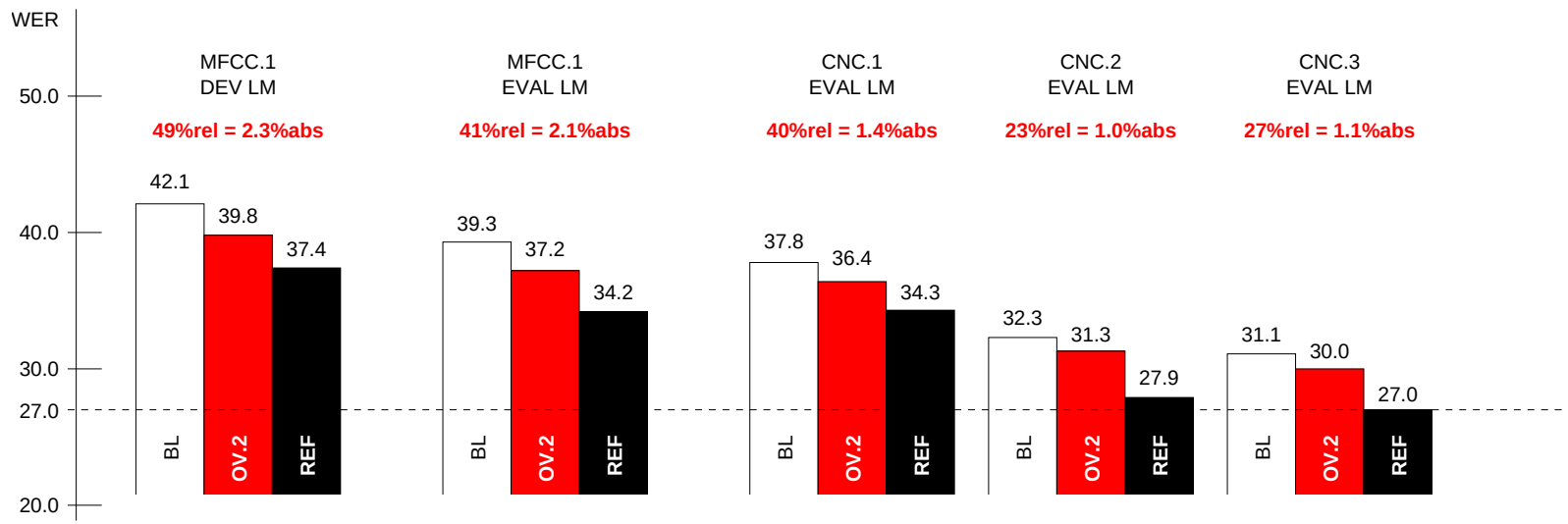
- when moving across ASR passes, relative WER reduction due to improved segmentation is approximately constant

Results: CNC.1 → CNC.2 → CNC.3



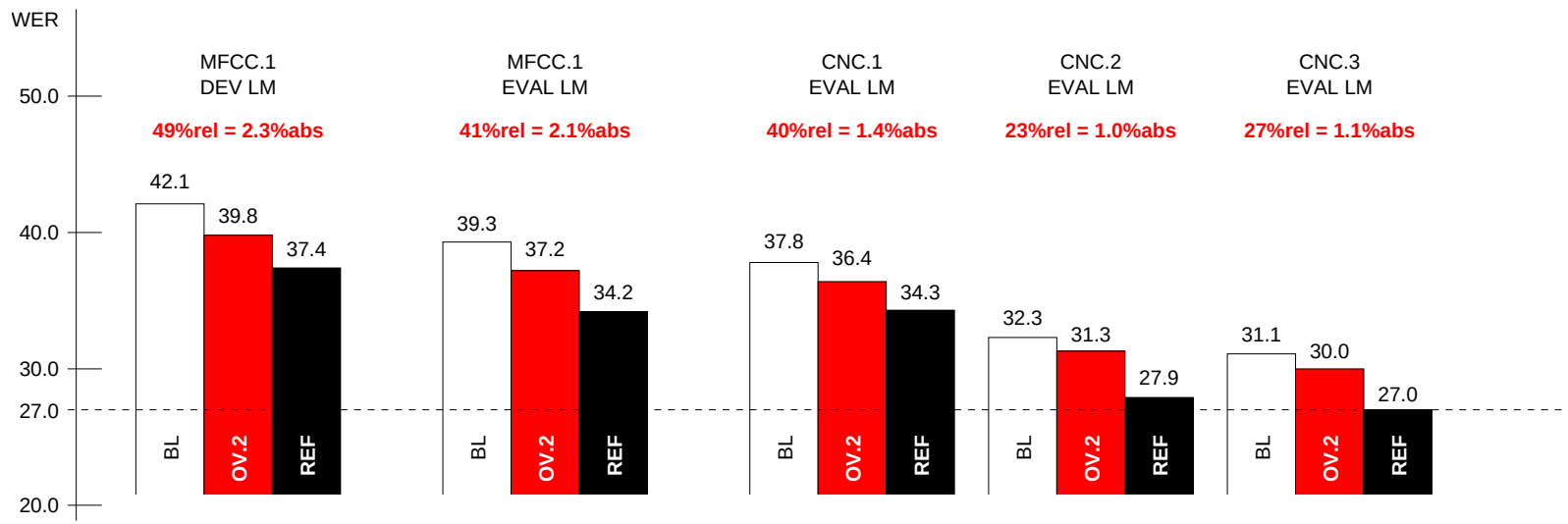
- when moving across ASR passes, relative WER reduction due to improved segmentation is approximately constant

Results: Evaluation Data



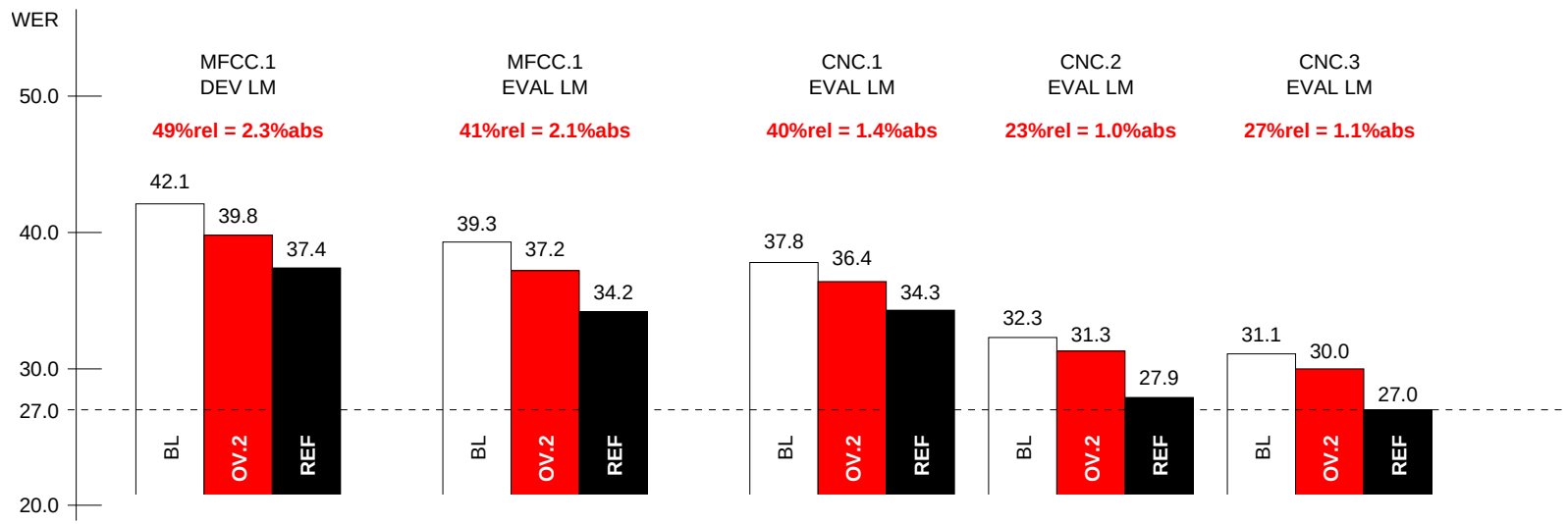
- eval set was overall more difficult
- 27% WER versus 25.3% WER in CNC.3
- trends noted for the dev set almost identical
- except MFCC.1 → CNC.1 WER on manual segmentation did not improve

Results: Evaluation Data



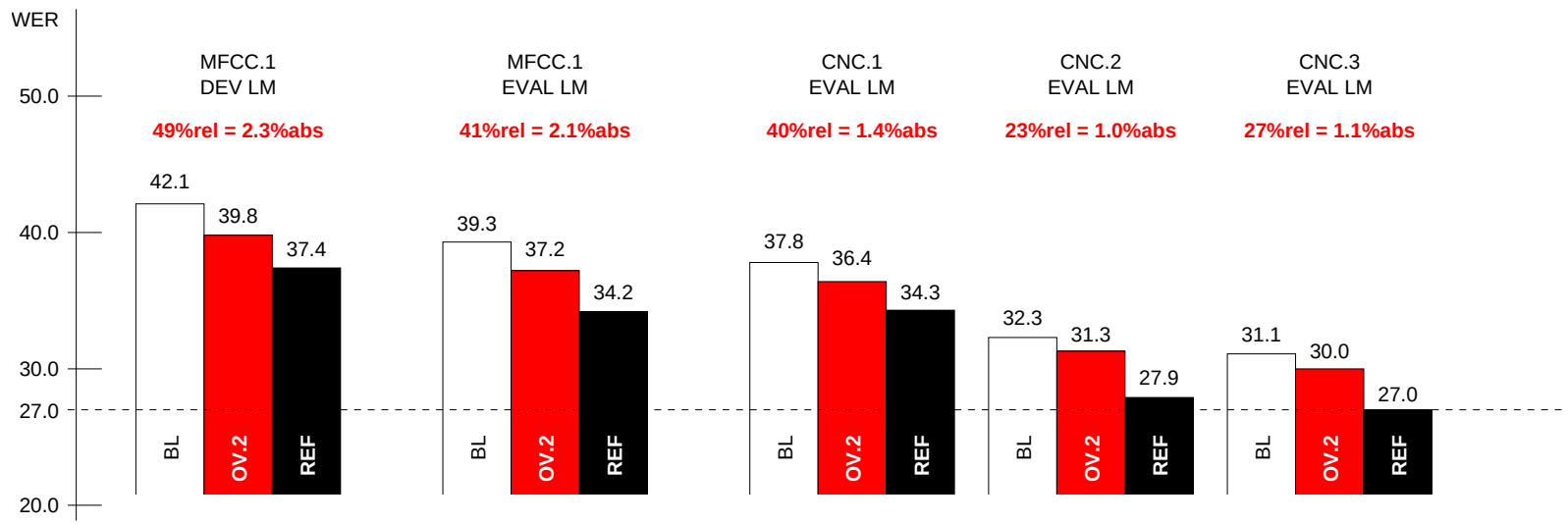
- eval set was overall more difficult
- 27% WER versus 25.3% WER in CNC.3
- trends noted for the dev set almost identical
- except MFCC.1 → CNC.1 WER on manual segmentation did not improve

Results: Evaluation Data



- 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 → CNC.1 WER on manual segmentation did not improve

Results: Evaluation Data



- 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

Conclusions

- 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:

Conclusions

- 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:

Conclusions

- 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:

Conclusions

- 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:

Conclusions

- 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:

Conclusions

- 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:
  - 5.2% absolute on the dev set
  - 3.7% absolute on the eval set
  - similar to results published by others

Conclusions

- 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
  - 3.7% absolute on the eval set
  - similar to results published by others

Conclusions

- 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

Conclusions

- 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

Conclusions

- 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

Conclusions

- 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