# Modeling Vocal Interaction for Segmentation in Meeting Recognition

Kornel Laskowski and Tanja Schultz

interACT, Carnegie Mellon University, Pittsburgh PA, USA,
{kornel|tanja}@cs.cmu.edu

**Abstract.** Automatic segmentation is an important technology for both automatic speech recognition and automatic speech understanding. In meetings, participants typically vocalize for only a fraction of the recorded time, but standard vocal activity detection algorithms for close-talk microphones in meetings continue to treat participants independently. In this work we present a multispeaker segmentation system which models a particular aspect of human-human communication, that of vocal interaction or the interdependence between participants' on-off speech patterns. We describe our vocal interaction model, its training, and its use during vocal activity decoding. Our experiments show that this approach almost completely eliminates the problem of crosstalk, and word error rates on our development set are lower than those obtained with human-generatated reference segmentation. We also observe significant performance improvements on unseen data.

## 1 Introduction

Vocal activity detection (VAD) is an important technology for any application with an automatic speech recognition (ASR) front end. In meetings, participants typically vocalize for only a fraction of the recorded time. Their temporally contiguous contributions should be identified prior to speech recognition in order to associate recognized output with specific speakers (who said what) and to leverage speaker adaptation schemes. Segmentation into such contributions is primarily informed by vocal activity detection on a frame-by-frame basis.

This work focuses on VAD for meetings in which each participant is instrumented with a close-talk microphone, a task which remains challenging primarily due to crosstalk from other participants (regardless of whether the latter have their own microphones). State-of-the-art meeting VAD systems which attempt to account for crosstalk rely on Viterbi decoding in a binary speech/non-speech space [12], assuming independence among participants. They employ traditional Mel-ceptral features as used by ASR, with Gaussian mixture models [1] or multilayer perceptrons [6]. Increasingly, such systems are integrating new features, designed specifically for discriminating between nearfield and farfield speech, or speaker overlap and no-overlap situations [14]. Research in this field is being fueled in large part by the Rich Transcription (RT) Meeting Recognition evaluations organized by NIST[1]. Generally reported ASR word error rates (WERs) on NIST RT corpora are still at least 2-3% absolute higher with automatically generated segments than with manual segmentation [1], a difference which is significant in the context of overall transcription system performance.

This paper describes an automatic segmentation system which is an extension to the segmentation component in our NIST RT-06s Speech-to-Text submission system in the individual head-mounted microphone (IMH) condition for conference meetings [8]. Both segmentation systems implement a fundamentally different approach from those used in other state-of-the-art transcription systems, in three main ways. First, we have chosen to address the crosstalk problem by explicitly modeling the correlation between all channels. This results in a feature vector whose length is a function of the number of meeting participants, which may vary from test meeting to test meeting. Because a variable feature vector length precludes the direct use of exclusively supervised acoustic models, we have proposed an unsupervised joint-participant acoustic modeling approach [10]. Second, we employ a model of multi-participant vocal interaction, which allows us to explicitly model the fact that starting to speak while other participants are speaking is dispreferred to starting in silence. Finally, as a consequence of our fully-connected, ergodic hidden Markov model architecture, state duration cannot be modeled directly. Our analysis window size, an order of magnitude larger than that in other state-of-the-art systems, is a trade-off between the desired endpoint granularity and minimum expected talkspurt duration.

Following a description of the new system in Sections 2, 3 and particularly 4, we compare the system to our NIST RT-06s segmentation system. We show that our final segmentation system outperforms manual segmentation on our development set, effectively treats uninstrumented participants, and leads to WERs only 2.2% absolute higher on unseen data than with manual segmentation.

## 2 Computational Framework

The VAD system we use as our baseline was introduced in [10]. Rather than detecting the 2-state speech ($\mathcal{V}$) vs. non-speech ($\mathcal{N}$) activity of each partipant independently, the baseline implements a Viterbi search for the best path through a $2^K$-state vocal interaction space, where $K$ is the number of participants. Our state vector, $\mathbf{q}_t$, formed by concatenating the concurrent binary vocal activity states $\mathbf{q}_t[k]$, $1 \geq k \geq K$, of all participants, is allowed to evolve freely over the vocal interaction space hypercube, under stochastic transition constraints imposed by a fully-connected, ergodic hidden Markov model (eHMM). Once the best vocal interaction state path $\mathbf{q}^*$ is found, we index out the corresponding best vocal activity state path $\mathbf{q}^*[k]$ for each participant $k$. The underlying motivation for this approach is that it allows us to model the constraints that participants exert on one another; it is generally accepted that participants are more likely to begin vocalizing in silence than when someone else is already vocalizing [4].

[1] http://www.nist.gov/speech/tests/rt/

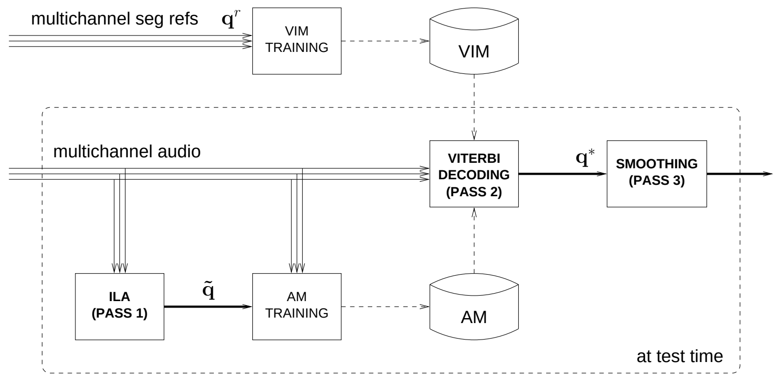



**Fig. 1.** Segmentation system architecture.

The architecture of the proposed segmentation system is depicted in Figure 1. Tasks associated with its operation, shown as rectangles in the figure, include:

1. VIM TRAINING: training of a conversation-independent vocal interaction model (Section 4);
2. **PASS 1**: initial label assignment (ILA) for the test audio (Section 3.1);
3. AM TRAINING: training of conversation-specific acoustic models (Section 3.2) using the *test* audio and the labels from (2);
4. **PASS 2**: simultaneous Viterbi decoding of all participant channels, using the vocal interaction model from (1) and the acoustic models from (3); and
5. **PASS 3**: smoothing VAD output to produce a segmentation suitable for ASR.

Space constraints prohibit a comprehensive description of each task or component. We only briefly describe the multiparticipant IHM acoustic model in the following section. In Section 4, we detail the structure of the proposed vocal interaction model, and outline its training and use during decoding.

## 3 Unsupervised Multispeaker IHM Acoustic Modeling

### 3.1 Initial Label Assignment

We perform an unsupervised initial assignment of state labels to multichannel frames of audio using the heuristic

$$\tilde{\mathbf{q}}\left[k\right] = \begin{cases} \mathcal{V}, \text{ if } \displaystyle\sum_{j \neq k} \log \left( \frac{\max_\tau \phi_{jk}\left(\tau\right)}{\phi_{jj}\left(0\right)} \right) > 0 \\ \mathcal{N}, \text{otherwise} \end{cases} , \quad (1)$$

where $\phi_{jk}\left(\tau\right)$ is the crosscorrelation between IHM channels $j$ and $k$ at lag $\tau$, and $\tilde{\mathbf{q}}\left[k\right]$ is the initial label assigned to the frame in question. We have shown, in [11], that under certain assumptions the criterion in Equation 1 is equivalent to declaring a participant as vocalizing when the distance between the location of the dominant sound source and that participant's microphone is smaller than the geometric mean of the distances from the source to each of the remaining microphones.

### 3.2 Acoustic Model Training

The initial label assignment described in Equation 1 produces a partitioning of the multichannel test audio. The labeled frames are used to train a single, full-covariance Gaussian for each of the $2^K$ states in our search space, over a feature space of $2K$ features: a log-energy and a normalized zero-crossing rate for each IHM channel. Features are computed using 110 ms non-overlapping windows, following signal preemphasis $\left(1 - z^{-1}\right)$.

For certain participants, and especially for frames in which more than one participant vocalizes, the ILA may identify too few frames in the test meeting for standard acoustic model training. To address this problem, we have proposed and evaluated two methods: feature space rotation, and sample-level overlap synthesis. Due to space constraints, we refer the reader to [10] for a description. We only mention here that the methods are controlled by three parameters, $\{\lambda_G, \lambda_R, \lambda_S\}$, whose magnitudes empirically appear to depend on the number of features per channel and on the overall test meeting duration.

## 4 Vocal Interaction Modeling

The role of the vocal interaction model during decoding is to provide estimates of $P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \mid \mathbf{q}_t = \mathbf{S}_i\right)$, the probability of transitioning to a state $\mathbf{S}_j$ at time $t+1$ from a state $\mathbf{S}_i$ at time $t$. The complete description of the conversation, when modeled as a first-order Markov process, is an $N \times N$ matrix, where $N \equiv 2^K$. When participants are assumed to behave independently of one another, this probability reduces to $\prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}[k] = \mathbf{S}_j[k] \mid \mathbf{q}_t[k] = \mathbf{S}_i[k]\right)$. As a result, a participant-independent description consists of a $2 \times 2$ matrix.

In this work, we have chosen to not assume that participants behave independently. Descriptive studies of conversation [13] and of meetings [4], as well as computational models in various fields [2][5], have unequivocally demonstrated that an assumption of independence is patently false. To our knowledge, however, suitable models of multiparty vocal interaction have not been designed for or applied to the task of detecting vocal activity for automatic speech recognition in meetings. A main difficulty is the need to collapse the $2^K \times 2^K$ transition probability matrix in a conversation-independent and participant-independent manner, such that model parameters learned in one conversation will generalize to unseen conversations, even when the participants are different, and/or when the number of participants in the train meetings does not match the number of participants in the test meeting.

### 4.1 Model Structure

To address this issue, we have proposed the following model of vocal interaction:

$$
\begin{aligned}
&P\,(\,\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\,) \;\; = \\
&\quad P\,(\,\|\mathbf{q}_{t+1}\| = n_j\,,\, \|\mathbf{q}_{t+1}\cdot\mathbf{q}_t\| = o_{ij} \,|\, \|\mathbf{q}_t\| = n_i\,) \;\; \times \\
&\quad P\,(\,\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \|\mathbf{q}_{t+1}\| = n_j\,,\, \|\mathbf{q}_{t+1}\cdot\mathbf{q}_t\| = o_{ij}\,,\, \|\mathbf{q}_t\| = n_i\,) \;\; ,
\end{aligned}
\tag{2}
$$

where $\|\mathbf{q}_t\|$ represents the number of participants vocalizing at time $t$, and $\|\mathbf{q}_t \cdot \mathbf{q}_{t+1}\|$ represents the number of participants who were vocalizing at time $t$ and who continue to vocalize at time $t+1$. Equation 2 introduces some additional notational shorthand: $n_i \equiv \|\mathbf{S}_i\|$ and $n_j \equiv \|\mathbf{S}_j\|$ are the number of vocally active participants in states $\mathbf{S}_i$ and $\mathbf{S}_j$, respectively, and $o_{ij} \equiv \|\mathbf{S}_i \cdot \mathbf{S}_j\| \leq \min(n_i, n_j)$ is the number of same participants which are vocally active in both $\mathbf{S}_i$ and $\mathbf{S}_j$.

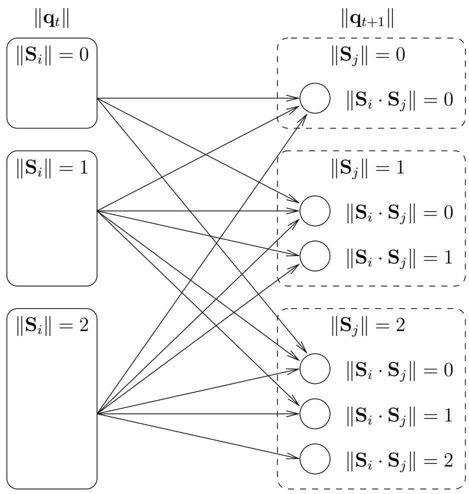



**Fig. 2.** Unique transition probabilities in the EDO model space with at most 2 simultaneously vocalizing participants.

The first factor in Equation 2 represents a time-independent, conversation-independent, and participant-independent model of transition among various degrees of multiparticipant overlap at times $t$ and $t+1$. We refer to this factor as the Extended Degree of Overlap (EDO) model. In particular, we claim that the probability of transition between two specific states is proportional to the probability of transition between the degrees of simultaneous vocalization in each of them. Furthermore, the term $\|\mathbf{q}_t \cdot \mathbf{q}_{t+1}\|$ accounts for participant state continuity; it allows the probability of the transition $\{A, B\} \longrightarrow \{A, C\}$ to differ from that of $\{A, B\} \longrightarrow \{C, D\}$, which agrees with intuition. Figure 2 shows the total number of unique transitions in the EDO space; for reasons of figure readability, we limit the maximum degree of participant overlap to 2.

The second factor in Equation 2 accounts for the multiplicity of specific next $\mathbf{S}_j$ states that are licensed by a particular EDO state transition $(n_i, o_{ij}, n_j)$. We illustrate this in Figure 3. As an example, the transitions $\{A\} \longrightarrow \{A, B\}$ and $\{A\} \longrightarrow \{A, C\}$ are both of $(n_i = 1, o_{ij} = 1, n_j = 2)$ EDO transition type, and they must divide the EDO transition mass between them (for $K = 3$ participants; for $K > 3$ participants, there are additional next state candidates). Because we are constructing a participant-independent model, we assume a uniform distribution over such candidate next states,

$$P\left( \mathbf{q}_{t+1} = \mathbf{S}_j \mid \|\mathbf{q}_{t+1}\| = n_j \,,\, \|\mathbf{q}_{t+1} \cdot \mathbf{q}_t\| = o_{ij} \,,\, \|\mathbf{q}_t\| = n_i \right) = \left( \frac{n_i!}{o_{ij}!\,(n_i - o_{ij})!} \cdot \frac{(K - n_i)!}{(n_j - o_{ij})!\,(K - n_i - n_j + o_{ij})!} \right)^{-1} \quad , \tag{3}$$

where $K$ is the number of participants in the test meeting. Equation 3 ensures that the conditional probabilities in Equation 2, for $1 \leq j \leq N$, sum to one.

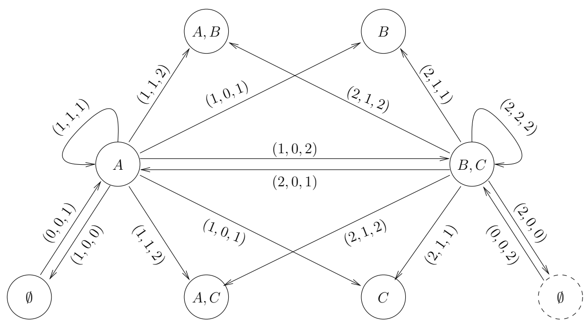


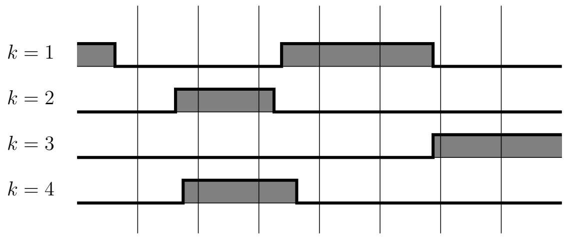

**Fig. 3.** The 7-state $\mathbf{S}_i$ space for a 3-participant conversation, showing the mapping of $(n_i, o_{ij}, n_j)$ transition probabilities from the EDO space. The all-silent state $\emptyset$ is duplicated for readability; we also show the transitions from only one single one-participant state ($\{A\}$), and from only one single two-participant state. The single three-participant state is not shown.

### 4.2 Training the EDO Model

To train the EDO model, we use the multi-participant utterance-level segmentation (`.mar`) from the ISL Meeting Corpus [3], where the number of meetings is $R = 18$. As in [10], the references are first discretized into a time-sequence of states $\mathbf{q}_t^r$; we illustrate this process in Figure 4. The model parameters are then estimated by accumulating bigram counts from the observed time-sequence, according to

$$P\left(\|\mathbf{q}_{t+1}\| = n_j\,,\, \|\mathbf{q}_t \cdot \mathbf{q}_{t+1}\| = o_{ij} \mid \|\mathbf{q}_t\| = n_i\right) \;=\; \frac{\displaystyle\sum_{\substack{r=1\\n_i+n_j-o_{ij}<K}}^{R}\sum_{t=1}^{T_r-1} \delta\left(\|\mathbf{q}_t^r\|, n_i\right)\delta\left(\|\mathbf{q}_t^r \cdot \mathbf{q}_{t+1}^r\|, o_{ij}\right)\delta\left(\|\mathbf{q}_{t+1}^r\|, n_j\right)}{\displaystyle\sum_{\substack{r=1\\n_i+n_j-o_{ij}<K}}^{R}\sum_{t=1}^{T_r-1} \delta\left(\|\mathbf{q}_t^r\|, n_i\right)}\,, \tag{4}$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta, and $r$ indexes training meetings. $K$ is the number of participants in the test meeting, and is given by the number of IHM channels to segment; its appearance in Equation 4 is due to the fact that the EDO model must be recompiled each time $K$ changes. This is because transitions may occur in the training material which are not possible in a particular test meeting: for example, a transition of type $(n_i = 2, o_{ij} = 0, n_j = 2)$, such as $\{A, B\} \longrightarrow \{C, D\}$, is not possible for a test meeting of $K = 3$ participants.




$$\mathbf{q}_t^r = \begin{array}{cccccccc} \mathcal{V} & \mathcal{N} & \mathcal{N} & \mathcal{V} & \mathcal{V} & \mathcal{V} & \mathcal{N} & \mathcal{N} \\ \mathcal{N} & \mathcal{N} & \mathcal{V} & \mathcal{N} & \mathcal{N} & \mathcal{N} & \mathcal{N} & \mathcal{N} \\ \mathcal{N} & \mathcal{N} & \mathcal{N} & \mathcal{N} & \mathcal{N} & \mathcal{N} & \mathcal{V} & \mathcal{V} \\ \mathcal{N} & \mathcal{N} & \mathcal{V} & \mathcal{V} & \mathcal{N} & \mathcal{N} & \mathcal{N} & \mathcal{N} \end{array}$$

**Fig. 4.** Assignment of discrete multi-participant values for consecutive frames of $\mathbf{q}^r$ from a reference segmentation. A frame is assigned a value $\mathcal{V}$ for participant $k$ if $k$ vocalizes for at least 50% of the frame duration; otherwise, $\mathcal{N}$ is assigned.

The estimation of the first term of Equation 2, as proposed in Equation 4, is participant-independent since, at each time $t$, only the *number* of currently vocalizing participants is inspected, rather than specific participants (as indexed by $k$ in $\mathbf{q}_t$). However, because the amount of vocalization overlap may vary across meetings or conversation types, the model is biased towards the interactions patterns observed in the training data; we have addressed this issue by training on a sizable corpus of meetings. There may be significant scope for selecting training material based on anticipated vocal interaction style. The model is also independent of the number of participants in each training or test meeting.

Studies of overlap occurrence in meetings [4] do not report a strong correlation with participant number; the increased potential for overlap due to larger group size appears to not be realized in general [7].

## 5 Experiments

We assess the performance of our algorithms by directly comparing the WERs as was done in [1][6]. WERs were produced using our multi-pass NIST RT-06s Speech-to-Text submission system [8]; however, in the current work, we show only the first-pass MFCC front end WERs, obtained with our RT-06s development language model. We note that an optimistic aim of an automatic segmenter is to produce WERs achievable with human-produced reference segmentation.

The data used in the described experiments consist of two datasets from the NIST RT-05s and RT-06s evaluations. Our development set, `rt05s_eval*` (referred to as *confDEV* in [8]), is the complete `rt05s_eval` set less one anomalous meeting (with a participant on speakerphone). We use the complete `rt06s_eval` as held out unseen data for final evaluation purposes.

The baseline segmenter in these experiments is that used in our NIST RT-06s submission, which differs from the current system in 4 ways. In this section we evaluate these four modifications, and present experiments which explore the impact that vocal interaction modeling has on ASR performance.

### 5.1 Elimination of Zero-Crossing Rate (ZCR)

The first delta from our RT-06s submission is the elimination of the zero crossing rate feature, whose implementation contained an error and which, following correction, was shown not to affect WERs. Since this modification reduces the feature vector size from $2K$ to $K$, we have also retuned the acoustic model factors $\{\lambda_G, \lambda_R, \lambda_S\}$ on the development set. The negligible effect of this change to the WER, alongside the performance of the RT06s baseline, is shown in Table 1.

### 5.2 Frame step/size reduction (F.100)

In a second experiment, we reduced the frame size and step from 0.110s to 0.100s. Since these parameters affect the smoothing pass, we have also modified the latter to consist of: (1) bridging gaps shorter than 0.45s; (2) eliminating spurts shorter than 0.25s; and (3) prepadding and postpadding all segments with 0.15s and 0.2s, respectively. The original smoothing consisted of 5 postprocessing passes: (1) bridging gaps shorter than 0.5s; (2) eliminating spurts shorter than 0.2s; (3) prepadding and postpadding all segments with 0.1s and 0.3s, respectively; (4) bridging remaining gaps shorter than 0.4s; and (5) eliminating remaining spurts shorter than 0.8s. As in the first experiment, these parameters were tuned to minimize WER on our development set. Table 1 shows that these two changes reduce substitutions and deletions on the development set, without increasing insertions.

| Segmentation | sub | del | ins | WER | WER' |
|---|---|---|---|---|---|
| RT06s baseline | 22.5 | 11.9 | 4.8 | 39.2 | 37.0 |
| – ZCR | 21.1 | 13.7 | 4.0 | 38.8 | 36.9 |
| + F.100 | 20.7 | 12.8 | 4.0 | 37.4 | 35.2 |
| + ILA.0 | 21.2 | 10.8 | 4.6 | 36.6 | 34.2 |
| + MULT | 21.1 | 11.1 | 4.3 | **36.5** | **34.1** |
| maxOV.4 | 21.1 | 11.1 | 4.3 | 36.5 | 34.1 |
| maxOV.3 | 21.1 | 11.2 | 4.3 | 36.5 | 34.1 |
| maxOV.2 | 21.0 | 11.5 | 4.3 | 36.8 | 34.4 |
| MIP | 21.3 | 11.5 | 4.4 | 37.2 | 34.9 |
| manual refs | 24.4 | 8.3 | 4.8 | **37.5** | **34.4** |

**Table 1.** First-pass ASR substitutions (sub), insertions (ins), deletions (del), and overall WER before rescoring, and overall WER after rescoring in the first pass (WER'). Detailed ASR errors prior to rescoring are shown because they correlate with frame-level miss and false alarm rates (not shown) better than do post-rescoring errors. Results are for our development set `rt05s_eval*`; best automatic and manual performance shown in bold.

### 5.3 Data selection for training the all-silent state (ILA.0)

A third reduction in the `rt05s_eval*` set WER was achieved by noting that the ILA algorithm is characterized by high precision but significantly lower recall [9]. This suggests that a large number of frames identified by the ILA as silence may in fact be missed vocal activity. To test this hypothesis, we chose to use only 50% of the ILA-identified silence frames for training the all-silent state model $\mathbf{S}_0$. These are selected by picking the bottom two quartiles in terms of average per-channel log-energy, over all channels. As Table 1 shows, this leads to a significant reduction in deletions, and produces an overall WER which is lower than that produced using manual segmentation.

### 5.4 Sharing probability mass among candidate next states (MULT)

The last delta between our RT-06s submission segmenter and the current system is the implementation of Equation 3. In the baseline system, this factor was ignored in Equation 2. This resulted in more frequent insertions, since the probability of transitioning to states with a high degree of overlap was not normalized by their multiplicity. This modification reduces the WER further below that obtained with manual segmentation.

### 5.5 Robustness and Generalization

In total, the four modifications described above and shown in Table 1 reduce the WER in the first pass from 37.0% to 34.1%, which surpasses ASR performance achieved with manual segmentation.

| Segm. | AMI1 | AMI2 | CMU1 | CMU2 | ICSI1 | ICSI2 | NIST1 | NIST2 | VT1 | VT2 | all |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RT06s | 33.7 | 47.4 | 36.8 | 37.8 | 34.5 | 27.6 | 119.8 | 37.9 | 37.7 | 40.8 | 45.6 |
| – ZCR | 33.8 | 38.8 | 37.6 | 34.5 | 43.5 | 27.1 | 91.1 | 40.9 | 34.5 | 41.9 | 42.5 |
| + F.100 | 33.6 | 36.3 | 33.1 | 34.0 | 42.3 | 27.1 | 91.7 | 39.5 | 33.7 | 38.7 | 41.1 |
| + ILA.0 | 34.0 | 36.6 | 32.9 | 33.9 | 34.4 | 27.0 | 94.8 | 37.7 | 34.5 | 38.4 | 40.5 |
| + MULT | 33.3 | 35.7 | 33.3 | 33.5 | 33.0 | 27.2 | 83.1 | 38.3 | 34.0 | 40.4 | 39.2 |
| maxOV.4 | 33.3 | 35.7 | 33.3 | 33.5 | 32.9 | 27.2 | 84.0 | 38.3 | 34.0 | 40.4 | 39.3 |
| maxOV.3 | 33.3 | 35.8 | 33.3 | 33.5 | 33.0 | 27.3 | 81.0 | 38.3 | 34.0 | 40.4 | 39.0 |
| maxOV.2 | 33.5 | 36.1 | 34.1 | 33.8 | 33.6 | 27.8 | 66.4 | 38.7 | 34.0 | 39.8 | 37.8 |
| MIP | 33.6 | 36.5 | 34.8 | 33.6 | 35.2 | 26.9 | 69.3 | 38.8 | 36.0 | 40.5 | 38.5 |
| manual | 34.7 | 39.3 | 32.9 | 31.3 | 25.8 | 25.3 | 51.2 | 44.0 | 34.3 | 44.8 | 36.1 |

**Table 2.** First-pass WERs after rescoring, for individual meetings in `rt05s_eval`.

In Table 2, we show the performance of our segmentation system individually for each meeting in `rt05s_eval`. As mentioned above, the `rt05s_eval` set is identical to our development set, plus the meeting identified as NIST1. As can be seen, the performance of the final system exceeds that of the baseline for every meeting except NIST2. For five meetings (AMI1, AMI2, NIST2, VT1 and VT2), performance is better with automatic than with human-generated segmentation.

| Segm. | CMU1 | CMU2 | EDI1 | EDI2 | NIST1 | NIST2 | TNO1 | VT1 | VT2 | all |
|---|---|---|---|---|---|---|---|---|---|---|
| RT06s | 36.9 | 45.1 | 31.6 | 33.3 | 48.1 | 51.8 | 42.9 | 47.8 | 39.4 | 42.1 |
| – ZCR | 37.1 | 45.2 | 35.9 | 41.1 | 43.1 | 49.5 | 46.9 | 45.2 | 37.2 | 42.6 |
| + F.100 | 36.1 | 45.5 | 36.3 | 35.8 | 43.8 | 49.7 | 46.6 | 44.3 | 36.0 | 41.8 |
| + ILA.0 | 55.0 | 42.6 | 34.6 | 35.3 | 42.8 | 43.5 | 41.2 | 44.7 | 37.0 | 42.5 |
| + MULT | 36.5 | 42.9 | 35.2 | 46.0 | 40.9 | 43.6 | 40.6 | 43.7 | 36.4 | 40.8 |
| maxOV.4 | 36.5 | 42.9 | 35.0 | 35.6 | 40.9 | 43.6 | 40.8 | 43.6 | 36.0 | 39.6 |
| maxOV.3 | 36.5 | 42.9 | 35.0 | 35.6 | 40.8 | 43.6 | 40.8 | 43.8 | 36.0 | 39.6 |
| maxOV.2 | 36.6 | 43.1 | 35.5 | 35.6 | 41.0 | 43.8 | 40.9 | 43.4 | 36.3 | 39.8 |
| MIP | 36.8 | 43.4 | 35.4 | 36.1 | 41.7 | 43.6 | 40.9 | 44.3 | 37.6 | 40.1 |
| manual | 37.2 | 40.0 | 34.7 | 32.2 | 39.7 | 35.6 | 41.7 | 39.3 | 33.9 | 37.4 |

**Table 3.** First-pass WERs after rescoring, for individual meetings in `rt06s_eval`.

We show a similar analysis in Table 3 for the `rt06s_eval` set. Cumulatively, our post-evaluation modifications improve performance on all but the two EDI meetings. These two meetings, together with TNO1, appear to have benefited from the faulty ZCR feature, and WERs for them never fully recover once that feature is eliminated. For two of the meetings, CMU1 and TNO1, WERs with automatic segmentation are lower than those with manual segmentation.

### 5.6 Impact of Modeling Vocal Interaction

Finally, we show results from several experiments in which we explore the impact of modeling vocal interaction on ASR performance. In the first, we limit the state space to states of at most 4 (maxOV.4), at most 3 (maxOV.3), and at most 2 (maxOV.2) simultaneously vocalizing participants. The results on our development set are shown in Table 1; those on the complete `rt05s_eval` and `rt06s_eval` sets are shown in Tables 2 and 3, respectively.

As can be seen, limiting the maximal degree of overlap always leads to more deletion errors, although the effect asymptotes after 4-participant overlap is included. This partly corroborates the observations on overlap in [4], namely that more-than-3-participant overlap is extremely rare. However, we note that for the NIST1 meeting in `rt05s_eval`, which contained a participant without a microphone and suffered from a large number of ASR insertion errors as a result, limiting the maximal degree of overlap effectively reduces the insertions. This effect more than compensates for the slightly increased deletions in the remaining meetings in that set, such that the overall WER is significantly lower.

We also explore the ASR performance which would be achieved with the current segmentation system if the transition model probabilities were provided not by our vocal interaction model but by a model which treats participants in a mutually independent manner, as in other state-of-the-art meeting segmenters [1][6]. In the context of our system, such a model would have the form

$$P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\right) = \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}\left[k\right] = \mathbf{S}_j\left[k\right] \,|\, \mathbf{q}_t\left[k\right] = \mathbf{S}_i\left[k\right]\right) \ . \tag{5}$$

ASR results using this model are given in Tables 1, 2, and 3 as MIP. It shows systematically worse performance; on our development set, the WER difference is 0.8% absolute, while that on the entire `rt05s_eval` is 0.7% absolute. On unseen data, the mutually independent participant model leads to a WER which is 0.5% absolute higher. We note that this is a conservative estimate of the difference; a fair estimate in the context of our system would require acoustic models for all possible overlap states, whereas our acoustic model training procedure typically produces models for at most 4-participant overlap. Furthermore, our full-covariance acoustic models treat participants jointly [10].

## 6 Conclusions

We have described the automatic segmentation system used in our NIST RT-06s Speech-to-Text Evaluation submission, together with several improvements. The system implements a novel approach to segmenting multi-channel, multi-speaker meeting recordings, in particular in its use of multi-participant acoustic and transition models. In its current state, the system outperforms human segmentation in first-pass ASR performance on our development set. The performance on the complete `rt05s_eval` and `rt06s_eval` sets leads to first-pass WERs which are 1.6%-2.2% absolute higher than with human segmentation, comparing favorably with other state-of-the-art systems [1][6].

## 7 Acknowledgments

We would like to thank Mari Ostendorf for her suggestions and help in clarifying and presenting the ideas in this paper. This work was partly supported by the European Union under the integrated project CHIL (IST-506909), Computers in the Human Interaction Loop (`http://chil.server.de`).

## References


1. K. Boakye and A. Stolcke. 2006. Improved speech activity detection using cross-channel features for recognition of multiparty meetings. *Proc. of INTERSPEECH2006*, Pittsburgh PA, USA, pp1962–1965.
2. P. Brady. 1969. A model for generating on-off speech patterns in two-way conversation. *Bell Systems Technical Journal*, **48**:7, pp2445–2472.
3. S. Burger, V. MacLaren, and H. Yu. 2002. The ISL Meeting Corpus: The impact of meeting type on speech style. *Proc. of ICSLP2002*, Denver CO, USA, pp301–304.
4. Ö. Çetin and E. Shriberg. 2006. Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site. *Proc. of MLMI2006 (Springer Lexture Notes in Computer Science **4299**)*, Washington DC, USA, pp200–211.
5. J. Dabbs Jr. and R. Ruback. 1987. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology*, **20**, pp123–169.
6. J. Dines, J. Vepa, and T. Hain. 2006. The segmentation of multi-channel meeting recordings for automatic speech recognition. *Proc. of INTERSPEECH2006*, Pittsburgh PA, USA, pp1213–1216.
7. N. Fay, S. Garrod, and J. Carletta. 2000. Group discussion as interactive dialogue or serial monologue: The influence of group size. *Psychological Science*, **11**(6), pp487–492.
8. C. Fügen, S. Ikbal, F. Kraft, K. Kumatani, K. Laskowski, J. McDonough, M. Ostendorf, S. Stüker, and M. Wölfel. 2006. The ISL RT-06S Speech-to-Text System. *Proc. of MLMI2006 (Springer Lexture Notes in Computer Science **4299**)*, Washington DC, USA, pp407–418.
9. K. Laskowski, Q. Jin, and T. Schultz. 2004. Crosscorrelation-based multispeaker speech activity detection. *Proc. of ISLP2004*, Jeju Island, South Korea, pp973–976.
10. K. Laskowski and T. Schultz. 2006. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. *Proc. of ICASSP2006*, Toulouse, France, **I**, pp993–996.
11. K. Laskowski and T. Schultz. 2007. A geometric interpretation of normalized maximum crosscorrelation for vocal activity detection in meetings. *Proc. of HLT-NAACL, Short Papers*, Rochester NY, USA, pp89–92.
12. T. Pfau, D. Ellis, and A. Stolcke. 2001. Multispeaker speech activity detection for the ICSI Meeting Recorder. *Proc. of ASRU2001*, Madonna di Campiglio, Italy, pp107–110.
13. H. Sacks, E. Schegloff, and G. Jefferson. 1974. A simplest sematics for the organization of turn-taking for conversation. *Language*, **50**:4, pp696–735.
14. S. Wrigley, G. Brown, V. Wan, and S. Renals. 2003. Feature selection for the classification of crosstalk in multi-channel audio. *Proc. of EUROSPEECH2003*, Geneva, Switzerland, pp469–472.