SIMULTANEOUS MULTISPEAKER SEGMENTATION FOR AUTOMATIC MEETING RECOGNITION

Kornel Laskowski[1,2]*, Christian Fügen*[1]*, and Tanja Schultz*[2]

[1]interACT, Universität Karlsruhe
am Fasanengarten 5, 76131 Karlsruhe, Germany

[2]interACT, Carnegie Mellon University
407 South Craig St., Pittsburgh PA 15213, USA

kornel@ira.uka.de, fuegen@ira.uka.de, tanja@cs.cmu.edu

ABSTRACT

Vocal activity detection 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, and standard vocal activity detection algorithms for close-talk microphones have shown to be ineffective. This is primarily due to the problem of crosstalk, in which a participant's speech appears on other participants' microphones, making it hard to attribute detected speech to its correct speaker. We describe an automatic multichannel segmentation system for meeting recognition, which accounts for both the observed acoustics and the inferred vocal activity states of all participants using joint multi-participant models. Our experiments show that this approach almost completely eliminates the crosstalk problem. Recent improvements to the baseline reduce the development set word error rate, achieved by a state-of-the-art multi-pass speech recognition system, by 62% relative to manual 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 VAD on a frame-by-frame basis.

This work focuses on VAD for meetings in which each participant is wearing 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 [9], assuming independence among participants. They employ traditional Mel-ceptral features as used by ASR, with Gaussian mixture models [1] or multi-layer perceptrons [3]. Increasingly, such systems are integrating new features, designed specifically for discriminating between nearfield and farfield speech, or speaker overlap and no-overlap situations [10].

Our approach to meeting segmentation deviates from that of other state-of-the-art segmenters in three main ways. First, we address the crosstalk problem by explicitly modeling the correlation between energy on all channels, which results in a meeting-dependent feature vector length and precludes the use of exclusively supervised acoustic models. Second, we explicitly model the interaction between participants, and perform a single decode in a 2^K-state space rather than K decodes in a 2-state space. Third, for simplicity, we emply a fully-connected hidden Markov model topology, which leads to optimal frame rates which are significantly larger than those employed in other meeting segmenters (ie. [1],[3]).

We assess the performance of our automatic segmentation systems by comparing the subsequent word error rate (WER), obtained with an automatic speech recognition (ASR) system, to WERs achieved by the same ASR system using a manually produced reference segmentation of the same audio. The baseline segmenter evaluated here is the same as that in our NIST RT-06s Speech-to-Text Evaluation[1] submission [4]. We present a description of the segmenter in Section 2.

In contrast to our previous work [7][8], in which a single-pass recongizer was used, we evaluate segmentation performance using a 3-pass ASR system. This system, described in Section 3, is a simplified version of our NIST RT-06s submission [4]. A WER comparison between segmentations over at least three passes is important in order to take into account the influence of speaker and channel adaptation on hypotheses from the first pass. Our experience has been that the WER gap between automatic segmentations shrinks with the number of passes. Our experiments and analysis are presented in Section 4; concluding remarks can be found in Section 5.

2. SEGMENTATION SYSTEM DESCRIPTION

The VAD system we use as our baseline was introduced in [6]. 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 $\mathbf{q}_t[k]$, $1 \le k \le K$, 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

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

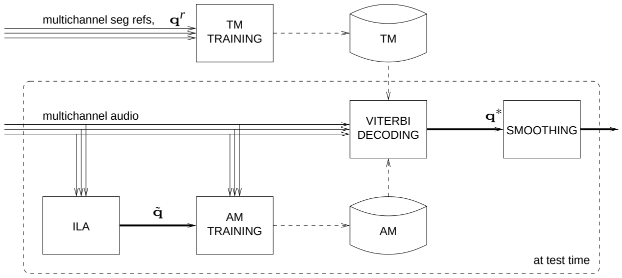

Figure 1: Segmentation system architecture

to model the constraints that participants exert on one another. The system, which employs 110 ms, non-overlapping frames, is depicted in Figure 1.

Tasks associated with running the system include:

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

The remainder of this section is devoted to a component-by-component description of the system.

2.1 TM: Transition Model Training

The role of the transition model during decoding is to provide estimates of $P(\mathbf{q}_{t+1}=\mathbf{S}_j\,|\,\mathbf{q}_t=\mathbf{S}_i)$, 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 a 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 $\prod_{k=1}^{K} P(\mathbf{q}_{t+1}[k]=\mathbf{S}_j[k]\,|\,\mathbf{q}_t[k]=\mathbf{S}_i[k])$.

As in previous work [6], we have chosen to not assume that participants behave independently. A main difficulty in modeling inter-participant dependencies is the need to collapse the 2^{2K} 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.

To address this issue, we have proposed the Extended Degree of Overlap (EDO) model [8], in which

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

where $n_i\equiv\|\mathbf{S}_i\|$ and $n_j\equiv\|\mathbf{S}_j\|$ represent the number of vocally active participants in states $\mathbf{S}_i$ and $\mathbf{S}_j$, respectively, and $o_{ij}\equiv\|\mathbf{S}_i\cdot\mathbf{S}_j\|\leq\min(n_i,n_j)$ represents the number of same participants which are vocally active in both $\mathbf{S}_i$ and $\mathbf{S}_j$. The model represents the probability of transition between specific states as 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 readability, we limit the degree of overlap in the figure to 2.

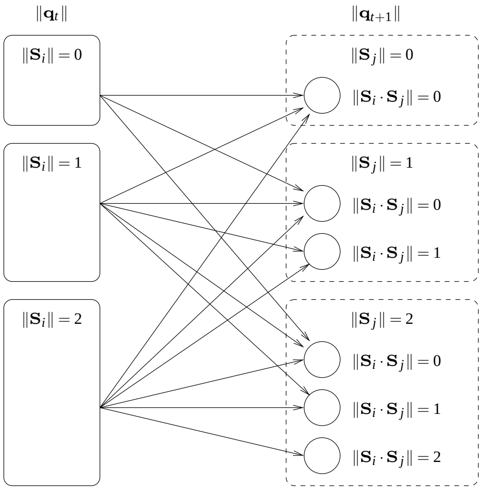
$\|\mathbf{q}_t\|$
$\|\mathbf{q}_{t+1}\|$
$\|\mathbf{S}_i\|=0$
$\|\mathbf{S}_i\|=1$
$\|\mathbf{S}_i\|=2$
$\|\mathbf{S}_j\|=0$
$\|\mathbf{S}_i\cdot\mathbf{S}_j\|=0$
$\|\mathbf{S}_j\|=1$
$\|\mathbf{S}_i\cdot\mathbf{S}_j\|=0$
$\|\mathbf{S}_i\cdot\mathbf{S}_j\|=1$
$\|\mathbf{S}_j\|=2$
$\|\mathbf{S}_i\cdot\mathbf{S}_j\|=0$
$\|\mathbf{S}_i\cdot\mathbf{S}_j\|=1$
$\|\mathbf{S}_i\cdot\mathbf{S}_j\|=2$

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

To train the EDO model, we use the multi-participant utterance-level segmentation (`.mar`) from the ISL Meeting Corpus [2], where the number of meetings is $R=18$. The training procedure is explained in detail in [8].

2.2 Initial Label Assignment

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

$$\tilde{\mathbf{q}}[k] \;=\; \begin{cases} \mathscr{V}, & \text{if } \sum_{j\neq k}\log\left(\frac{\max_\tau \phi_{jk}(\tau)}{\phi_{jj}(0)}\right)>0 \\ \mathscr{N}, & \text{otherwise} \end{cases} \tag{2}$$

where $\phi_{jk}(\tau)$ is the crosscorrelation between IHM channels j and k at lag τ, and $\tilde{\mathbf{q}}[k]$ is the initial label we assign to the frame in question. The crosscorrelation is computed in the spectral domain, using rectangular windows[2]. We perform maximization directly without the use of weighting schemes such as the phase transform (PHAT). In [7], we showed that the ratio $\max_\tau\phi_{jk}(\tau)/\phi_{jj}(0)$, under certain simplifying assumptions of signal propagation and microphone response, approximates the ratio d_j/d_k where d_j and d_k are the distances to microphones j and k, respectively, from a single dominant sound source. Equation 2 therefore declares participant k as vocalizing when the distance between the dominant sound source and microphone k is smaller than the geometric mean of the distances between the dominant sound source and each of the remaining microphones.

2.3 AM: Acoustic Model Training

The initial label assignment described in Equation 2 produces a partitioning of the multichannel test audio. The labeled

[2] In [7], we erroneously specified that we are using Hamming windows. There, as well as here, we use a rectangular window for ILA.

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. These features are computed using 110 ms non-overlapping Hamming windows following signal preemphasis $(1-z^{-1})$.

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 to effectively train acoustic models. 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 [6] 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.

2.4 Viterbi Decoding and Segmentation Smoothing

We perform standard Viterbi decoding to obtain the best path $\mathbf{q}^*$, and then extract a vocal activity estimate $\mathbf{q}_t[k]$ for each participant k. To produce usable segments for ASR, we perform 5 postprocessing passes per participant: (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. These post-processing parameters were tuned on the development set to minimize a first-pass WER.

3. ASR SYSTEM DESCRIPTION

Figure 3 shows the mult-pass ASR system used in our experiments. The system includes the first three passes of our NIST RT-06s Speech-to-Text Evaluation system. Two different semi-continuous acoustic models were used for decoding: one trained with vocal tract length normalization (VTLN) only, and one trained using speaker adaptive training (SAT) with constrained maximum likelihood linear regression (MLLR). In both cases, 16000 distributions over 4000 codebooks were trained, with a maximum of 64 Gaussians per model in a 42-dimensional feature space. In addition to the traditional front-end based on Mel-frequency Cepstral Coefficients (MFCCs), a second front-end based on warped minimum variance distortionless response (MVDR) was used. Combining the outputs of MFCC and MVDR systems, using confusion network combination (CNC), leads to significant cross-adaptation gains in subsequent passes.

We perform incremental adaptation using VTLN and constrained MLLR during decoding in the first pass. Thereafter, the parameters for VTLN, constrained MLLR and model-space MLLR are computed using confidence-annotated hypotheses of the first CNC pass and kept fixed during subsequent decoding passes. VTLN-only acoustic models are used in the first and second passes, while the third pass uses the SAT models. The language model used for all decoding passes was an interpolation of different 4-gram models computed on meeting data, lecture data, conversational telephone speech data and several other sources collected from the web. A more detailed description of the ASR components can be found in [4].

The development of the segmentation systems [8] evaluated in this work was carried out using a single-pass ASR system, which we refer to as MFCC.0. This corresponds to MFCC.1 in Figure 3; however, MFCC.0 relies on a dictionary and language model which were available during the development of [4].

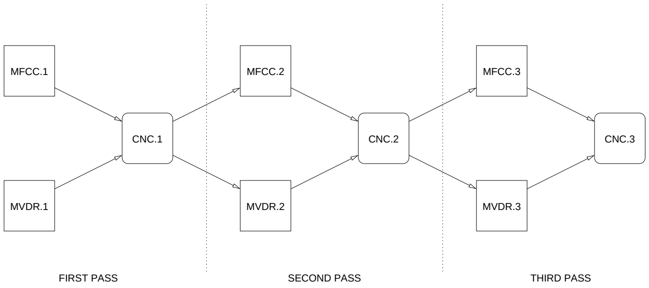

Figure 3: 3-pass ASR system.

4. EXPERIMENTS

We compare the performance of our segmentation algorithms by directly comparing the WERs as was done in [1], [3], [7], and [8]. WERs reported here are obtained using the 3-pass variant of our NIST RT-06s Speech-to-Text submission system, described in the previous section. We note that an optimistic aim of an automatic segmenter is to produce WERs achievable with manual, human-produced segmentation.

4.1 Data

The data used in our experiments consist of two datasets from the NIST RT-06s evaluation. `rt05s_eval` was used for development, and `rt06s_eval` was used for final evaluation; we have retained this separation in the current work. The two sets consist of 10-minute excerpts from several meetings recorded at different sites; the number of participants per meeting varies between 3 and 11.

Segmentation system development was carried out while excluding a single meeting from `rt05s_eval` which contained a participant on speakerphone; this condition was known in advance not to occur in `rt06s_eval`. We refer to the limited development set as `rt05s_eval*` (it was referred to as *confDEV* in [4]). For the purposes of future comparison with other meeting transcription systems, in the current work we include the offending meeting in our development set results and analysis.

4.2 Modifications to the Baseline Segmenter

We briefly describe 5 modifications made to the baseline, after the NIST RT-06s evaluation.

The first modification involved the elimination of the zero crossing rate (ZCR) feature, which 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 parameters $\{\lambda_G, \lambda_R, \lambda_S\}$ on `rt05s_eval*`. The negligible effect of this change to the MFCC.0 WER on `rt05s_eval*`, alongside the performance of the RT06s baseline, is shown in Table 1.

In a second modification (F.100), we reduced the frame size and step from 0.110 s to 0.100 s. 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

Segmentation	rt05s_eval*	rt05s_eval
RT06s baseline	37.0	45.6
– ZCR	36.9	42.5
+ F.100	35.2	41.1
+ ILA.0	34.2	40.5
+ MULT	**34.1**	39.2
+ OV.2	34.4	**37.8**
manual refs	34.4	36.1

Table 1: MFCC.0 WERs on our original rt05s_eval* development set (rt05s_eval less one meeting) and on the complete rt05s_eval development set, for five consecutive modifications to the segmentation baseline. Best-performing automatic segmentations are shown in bold.

postpadding all segments with 0.15s and 0.2s, respectively. As for the first modifications, these parameters were tuned to minimize the MFCC.0 WER on rt05s_eval*.

A third reduction (ILA.0) in the MFCC.0 WER on rt05s_eval* was achieved by noting that the ILA algorithm is characterized by high precision but significantly lower recall [5]. 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 model $\mathbf{S}_0$. These are selected by picking the first two quartiles in terms of average per-channel log-energy, over all channels. Table 1 shows the resulting WER reduction.

The fourth modification consisted of replacing Equation 2 with true probabilities of the form

$$\begin{aligned}
&P\left(\mathbf{q}_{t+1}=\mathbf{S}_j \,|\, \mathbf{q}_t=\mathbf{S}_i\right) = \\
&\quad P\left(\|\mathbf{q}_{t+1}\|=n_j\,,\,\|\mathbf{q}_{t+1}\cdot\mathbf{q}_t\|=o_{ij}\,|\,\|\mathbf{q}_t\|=n_i\right) \;\times \\
&\quad P\left(\mathbf{q}_{t+1}\,|\,\|\mathbf{q}_{t+1}\|=n_j\,,\,\|\mathbf{q}_{t+1}\cdot\mathbf{q}_t\|=o_{ij}\,,\,\|\mathbf{q}_t\|=n_i\right),
\end{aligned} \tag{3}$$

where only the first factor is supplied by the EDO model. Given an EDO model transition, we assume the distribution over next states licensed by that transition type to be uniform, ie.

$$\begin{aligned}
&P\left(\mathbf{q}_{t+1}\,|\,\|\mathbf{q}_{t+1}\|=n_j\,,\,\|\mathbf{q}_{t+1}\cdot\mathbf{q}_t\|=o_{ij}\,,\,\|\mathbf{q}_t\|=n_i\right) = \\
&\quad \frac{n_i!}{o_{ij}!\,(n_i-o_{ij})!}\cdot\frac{(K-n_i)!}{(n_j-o_{ij})!\,(K-n_i-n_j+o_{ij})!}\;.
\end{aligned} \tag{4}$$

where K is the number of participants in the test meeting. This ensures that the transition probabilities given by Equation 3 sum to unity. The MFCC.0 WER of the system with this modification is shown as MULT in Table 1.

Finally, we have simplified the search space by considering states with only zero, one or two simultaneously vocalizing participants (OV.2). This entails redistributing transition probability mass among the remaining $\|\mathbf{S}\|\leq 2$ states and ignoring, during the training of the EDO model, all N-grams involving $\|\mathbf{S}\| > 2$ states. Table 1 shows the impact of OV.2 on the MFCC.0 WER; although it leads to a negligible WER increase for rt05s_eval*, it yields a very large improvement in the segmentation of the excluded meeting with a speakerphone participant. We have therefore chosen OV.2 for validation of our segmentation approach using the 3-pass recognizer on the rt06s_eval dataset.

4.3 Generalization to a Multi-pass Recognizer

In Table 2, we show the performance of our segmentation system individually for each meeting in rt05s_eval, using the 3-pass ASR system described in Section 3. WERs for each of the 3 CNC passes are shown, together with WERs using the single-pass MFCC.0 ASR system for comparison.

We note first of all that the OV.2 segmentation, for all meetings combined, represents an 82% reduction in WER relative to manual segmentation, over the baseline segmenter when the single-pass ASR system is used. However, for each pass of the 3-pass ASR system, this figure is smaller, at 62%. In absolute terms, OV.2 performance using the single-pass recognizer is 1.7% absolute worse than manual segmentation, but with the multi-pass recognizer the same difference is 3.3-3.7%. OV.2 WERs, relative to manual segmentation WERs, represent a degradation of 12-13% in each pass.

Looking at each rt05s_eval meeting individually, it is apparent that OV.2 outperforms manual segmentation for a majority of meetings with the single-pass recognizer with which it was developed. MFCC.0 WERs on AMI1, AMI2, NIST2, VT1, and VT2 are lower that with manual segmentation by as much as 5.3% absolute. However, already in the first pass with the multi-pass recognizer, only AMI1, VT1, and VT2 have lower WERs for OV.2. By the third pass, only the AMI2 WER is the same with OV.2 segmentation as with manual segmentation; for all other meetings, manual segmentation yields the lowest WERs.

The situation is similar for unseen data in rt06s_eval, presented in Table 3. Although OV.2 outperforms manual segmentation for CMU1 and TNO1 with the single-pass recognizer, by the second pass of the 3-pass recognizer manual segmentation achieves WERs which are lower by 0.8-5.3% absolute than OV.2. Overall for rt06s_eval, OV.2 represents a 49% reduction of the MFCC.0 WER gap between manual segmentation and the baseline; for CNC.1, CNC.2 and CNC.3, the same reduction is 40%, 22%, 27%, respectively.

5. CONCLUSIONS

We have described the automatic segmentation system used in our NIST RT-06s Speech-to-Text Evaluation submission, together with several modifications. Although the improved OV.2 segmenter was developed by minimizing the WER obtained with a single-pass, development ASR system, the benefits generalize when a state-of-the-art 3-pass recognizer is used. With the exception of two meetings in the development set, CMU2 and ICSI2, and two meetings in the evaluation set, EDI1 and EDI2, third pass WERs obtained using OV.2 are lower than those with the baseline segmentation. When all meetings are considered, OV.2 reduces the rt05s_eval WER difference between the baseline and manual segmentation by 62%, and the unseen rt06s_eval WER difference by 27%. The current gap between manual and automatic segmentation is 3.3% for the rt05s_eval development set and 3.0% for the rt06s_eval evaluation set. These numbers are comparable to those reported elsewhere (ie. [1]), in spite of significant differences in segmentation system design.

A breakdown of CNC.3 ASR errors by type for rt06s_eval reveals that the number of substitutions, deletions, and insertions using OV.2 is 13.9%, 13.3%, and 2.9%, respectively. The same breakdown for manual segmentation yields 15.0%, 9.1%, and 2.9%, respectively. That OV.2 in-

Segm.	ASR	AMI1	AMI2	CMU1	CMU2	ICSI1	ICSI2	NIST1	NIST2	VT1	VT2	all
baseline	MFCC.0	33.7	47.4	36.8	37.8	34.5	27.6	119.8	37.9	37.7	40.8	45.6
OV.2	MFCC.0	33.5	36.1	34.1	33.8	33.6	27.8	66.4	38.7	34.0	39.8	37.8
manual	MFCC.0	34.7	39.3	32.9	31.3	25.8	25.3	51.2	44.0	34.3	44.8	36.1
baseline	CNC.1	28.4	43.9	34.8	34.7	31.3	26.2	108.9	34.3	33.6	38.8	41.7
OV.2	CNC.1	29.5	33.3	33.3	32.9	30.9	27.0	61.6	42.5	30.0	37.0	35.8
manual	CNC.1	33.8	31.6	31.2	31.0	22.0	23.4	45.4	35.6	30.1	38.7	32.1
baseline	CNC.2	24.6	30.4	27.3	28.1	29.0	21.6	100.0	30.1	27.7	33.8	35.4
OV.2	CNC.2	24.8	25.7	27.9	27.6	27.6	22.7	52.2	31.0	26.5	32.3	29.8
manual	CNC.2	24.7	26.8	25.0	25.1	20.0	20.4	37.4	27.7	26.7	31.3	26.4
baseline	CNC.3	23.7	27.0	27.0	26.9	28.2	20.8	96.5	29.5	26.5	33.3	34.1
OV.2	CNC.3	23.4	24.0	27.0	27.1	26.5	21.9	49.4	29.3	25.6	32.2	28.6
manual	CNC.3	22.3	24.0	24.8	24.8	19.9	20.2	35.9	26.5	25.3	30.3	25.3

Table 2: WERs for the baseline, OV.2 and manual segmentations using a single-pass development ASR system (MFCC.0) and a 3-pass system (cf. Figure 3), for the individual meetings in `rt05s_eval`.

Segm.	ASR	CMU1	CMU2	EDI1	EDI2	NIST1	NIST2	TNO1	VT1	VT2	all
baseline	MFCC.0	36.9	45.1	31.6	33.3	48.1	51.8	42.9	47.8	39.4	42.1
OV.2	MFCC.0	36.6	43.1	35.5	35.6	41.0	43.8	40.9	43.4	36.3	39.8
manual	MFCC.0	37.2	40.0	34.7	32.2	39.7	35.6	41.7	39.3	33.9	37.4
baseline	CNC.1	32.3	40.5	28.6	28.4	44.6	47.4	41.3	41.3	34.0	37.8
OV.2	CNC.1	32.7	38.5	36.8	31.8	38.6	40.1	37.0	38.0	34.1	36.4
manual	CNC.1	38.2	35.8	33.1	27.8	34.5	32.1	40.2	33.9	30.7	34.3
baseline	CNC.2	28.9	36.0	22.3	26.5	38.9	35.2	34.1	36.5	29.6	32.3
OV.2	CNC.2	29.2	34.4	23.7	27.5	34.4	33.0	33.7	34.6	29.1	31.3
manual	CNC.2	28.7	32.4	21.7	25.2	29.5	23.2	32.7	29.3	26.0	27.9
baseline	CNC.3	28.9	35.6	20.7	26.1	37.9	31.1	33.5	35.2	28.8	31.1
OV.2	CNC.3	28.3	33.9	22.0	26.5	33.5	28.9	32.9	34.1	28.0	30.0
manual	CNC.3	27.7	31.9	20.2	24.6	29.2	21.9	31.9	28.4	24.4	27.0

Table 3: WERs for the baseline, OV.2 and manual segmentations using a single-pass development ASR system (MFCC.0) and a 3-pass ASR system (cf. Figure 3), for the individual meetings in `rt06s_eval`.

curs the same rate of insertions as human segmentation suggests that it successfully addresses crosstalk, widely believed to be the main source of segmentation errors in close-talk microphone recordings of multi-party meetings.

6. ACKNOWLEDGMENTS

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 INTERSPEECH*, Pittsburgh PA, USA, pp1962–1965.

[2] S. Burger, V. MacLaren, and H. Yu. 2002. The ISL Meeting Corpus: the Impact of Meeting Type on Speech Style. *Proc. of ICSLP*, Denver CO, USA, pp301–304.

[3] J. Dines, J. Vepa, and T. Hain. 2006. The Segmentation of Multi-channel Meeting Recordings for Automatic Speech Recognition. *Proc. of INTERSPEECH*, Pittsburgh PA, USA, pp1213–1216.

[4] 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 Evaluation System. *Proc. of MLMI (Springer Lecture Notes in Computer Science 4299)*, Washington DC, USA, pp407–418.

[5] K. Laskowski, Q. Jin, and T. Schultz. 2004. Crosscorrelation-based Multispeaker Speech Activity Detection. *Proc. of INTERSPEECH*, Jeju Island, South Korea, pp973–976.

[6] K. Laskowski and T. Schultz. 2006. Unsupervised Learning of Overlapped Speech Model Parameters for Multichannel Speech Activity Detection in Meetings. *Proc. of ICASSP*, Toulouse, France, I:993–996.

[7] K. Laskowski and T. Schultz. 2007. A Geometric Interpretation of Non-Target-Normalized Maximum Cross-channel Correlation for Vocal Activity Detection in Meetings. *Proc. of HLT-NAACL, Short Papers*, Rochester NY, USA, pp89-92.

[8] K. Laskowski and T. Schultz. 2007. Modeling Vocal Interaction for Segmentation in Meeting Recognition. *Proc. of MLMI*, Brno, Czech Republic.

[9] T. Pfau and D. Ellis and A. Stolcke. 2001. Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. *Proc. of ASRU*, Madonna di Campiglio, Italy, pp107–110.

[10] S. Wrigley, G. Brown, V. Wan, and S. Renals. 2003. Feature Selection for the Classification of Crosstalk in Multi-Channel Audio. *Proc. of EUROSPEECH*, Geneva, Switzerland, pp469–472.