

Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site

Özgür Çetin¹ and Elizabeth Shriberg^{1,2}

¹ International Computer Science Institute, Berkeley, CA, USA

² SRI International, Menlo Park, CA, USA
{ocetin, ees}@icsi.berkeley.edu

Abstract. We analyze speaker overlap in multiparty meetings both in terms of automatic speech recognition (ASR) performance, and in terms of distribution of overlap with respect to various factors (collection site, speakers, dialog acts, and hot spots). Unlike most previous work on overlap or crosstalk, our ASR error analysis uses an approach that allows comparison of the same foreground speech with and without naturally occurring overlap, using a state-of-the-art meeting recognition system. We examine 101 meetings. For analysis of ASR, we use 26 meetings from the NIST meeting transcription evaluations, and discover a number of interesting phenomena. First, overlaps tend to occur at high-perplexity regions in the foreground talker’s speech. Second, overlap regions tend to have higher perplexity than those in nonoverlaps, if trigrams or 4-grams are used, but unigram perplexity within overlaps is considerably lower than that of nonoverlaps. Third, word error rate (WER) after overlaps is consistently lower than that before the overlap, apparently because the foreground speaker reduces perplexity shortly after being overlapped. These appear to be robust findings, because they hold in general across meetings from different collection sites, even though meeting style and absolute rates of overlap vary by site. Further analyses of overlap with respect to speakers and meeting content were conducted on a set of 75 additional meetings collected and annotated at ICSI. These analyses reveal interesting relationships between overlap and dialog acts, as well as between overlap and “hot spots” (points of increased participant involvement). Finally, results from this larger data set show that individual speakers have widely varying rates of being overlapped.

1 Introduction

Speaker overlap is frequent in natural conversation. For example, in the 26 different meetings we look at in this work (from the last four year NIST meeting speech recognition evaluations) an average of roughly 12% of all foreground speaking time is overlapped by speech from one or more other talkers. The ratio is even higher (30 to 50%) if one considers pause-delimited regions as units, rather than speaking time [17].

While general effects of overlap are well reported in the literature (e.g., [3], [10], [12], [16], [17], and [21]), there is relatively little work quantifying such

effects under the different conditions we consider. In particular, to the best of our knowledge, the issue of the effect of overlaps on ASR errors *adjacent* to overlap regions has received little attention in earlier work. We explore this question by using a method that allows us to compare ASR results for the same foreground speech with and without naturally occurring overlap. For data from the six different meeting collection sites, we examine both ASR and language model perplexity as a function of the presence or absence of crosstalk, its severity, and time distance before and after the overlap.

To begin to better understand patterns of overlap with respect to meeting content, we use an additional set of 75 ICSI meetings that are independently hand-annotated for dialog acts and hot spots. We ask whether overlap is associated with specific dialog acts, and in turn whether such information can shed light on perplexity patterns and ASR results. We also ask to what degree hot spots are correlated with overlap, since increased involvement would be assumed to predict increased overlap. Finally, since the ICSI data set contains significant amounts of data per speaker, we ask how individual speakers vary in terms of how frequently they are overlapped by another talker.

2 Method

2.1 Data

For analyses of ASR performance, we use 19.8 hours of recordings from 26 different meetings from the 2002, 2004, and 2005 NIST meeting speech recognition evaluations [11]. These meetings were provided by the sites AMI (2), CMU (6), ICSI (6), LDC (4), NIST (6), and VT (2), with the number of meetings given in parentheses. The number of participants varies from three to nine, and the total amount of speech in the individual headset microphones (IHMs) after segmentation is about 3.5 hours. For further analyses, requiring human annotations, we use a set of 75 meetings from the ICSI meeting corpus [9]. In separate efforts, this set was extensively hand-marked for dialog acts [6] as well as for hot spots [20]. Since this data was included in training the ASR system, we did not use it in analyses of recognition or perplexity.

2.2 Recognition System

Recognition experiments are conducted using the 2005 ICSI-SRI meeting system [19]. This system is adapted from SRI’s conversational telephone speech system to the meeting domain using a variety of meeting data, excluding the test data. The standard n -gram language models (LMs) with order as high as four were trained on standard text and meeting transcriptions as well as on Web texts [4]. We use manual reference segmentations in our experiments and analyses to avoid confounds with the automatic speech segmentation errors.

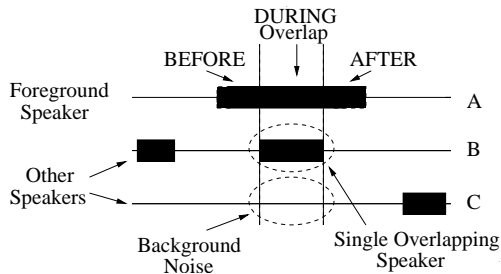


Fig. 1. Illustration of experiment conditions. When *A* is taken as the foreground speaker, *B* and *C* are background speakers. For the crosstalk condition, the original audio from *B* and *C* is added to *A*. For the background-noise condition, *B* and *C* are added only in the cases in which they do not contain any speech (e.g., inside DURING, *B* is not added to *A* and only *C* is added). BEFORE and AFTER in *A* are nonoverlaps. Solid rectangles denote speech segments.

2.3 Experiment Conditions

We use synchronously recorded speech from IHMs and speech/nonspeech alignments to create a rendition of crosstalk that is accurate in terms of speech that has overlapped and crosstalk severity. First, each channel is normalized to have unit energy using the average energy of speech samples in that channel. Next, to each channel the remaining channels are added in a time-synchronous fashion, after an appropriate linear weighting to adjust crosstalk severity (referred to as the “crosstalk condition”). To provide a contrast condition for isolating effects of background noises, we perform a second set of experiments, where a channel from the remaining channels is added only if no speech activity is marked for that channel (referred to as the “background-noise condition”). The performance differences between the crosstalk and background-noise conditions should indicate the crosstalk effects mainly due to the actual speech as opposed to background noise. See Figure 1 for an illustration of the design.

It is important to note that the crosstalk condition contains only speech that actually occurred at the same time. We do not create crosstalk using speech from different corpora or time spans. Nevertheless, the waveform addition is admittedly simplistic and does not capture some aspects of crosstalk such as nonlinear frequency weighting and reverberation. However, the effects from these factors would act only to exacerbate effects we report on here. Our study uses the performance difference between results with and without crosstalk in the same region of speech, and at this level of relative comparison such effects would be roughly normalized out. Also, to assess generalizability we repeat our crosstalk experiments with mixing powers $1/4$, $1/2$, and 1 .

3 Results

3.1 Rate of Overlap by Site

Table 1 provides rates of overlap in the evaluation test data from the six different sites, along with the rate overall. Rates are computed as the ratio of the

<i>Rate</i>	<i>All</i>	<i>AMI</i>	<i>CMU</i>	<i>ICSI</i>	<i>LDC</i>	<i>NIST</i>	<i>VT</i>
total overlap time/total speaking time	11.6	7.1	13.3	13.0	12.3	11.0	6.1
single-speaker overlap time/total overlap time	92.2	93.0	91.0	91.0	94.3	92.7	85.2

Table 1. Rates (%) of overlap by site. Line 1 provides the percentage of speech duration that is overlapped by any number of speakers. Line 2 considers all overlap events, and provides the percentage of overlaps that involve only two speakers.

time during which a foreground talker is speaking while overlapped, to the total amount of foreground speaking time over all foreground talkers. As shown, four sites have rates ranging from 11% to 13.3%, which is quite close considering that the meetings are of different natures. Two sites, AMI and VT, have significantly lower rates; this suggests that these two meeting types may be more artificial in terms of interaction patterns. For all sites but VT, over 90% of the overlaps involve only one background speaker, even though the meetings involved more than two speakers. VT shows a somewhat different pattern, with a higher rate of multiple-speaker overlaps, and yet a lower rate of overlap overall. This suggests that in VT meetings, overlap may be associated with a different function than it is in the five other meeting types.

3.2 ASR and Perplexity by Overlap Condition

WER results for various recognition conditions are provided in Table 2. WERs in this table are cumulative for all segments of the test data; analyses for overlaps and nonoverlaps are provided later. As one would expect, both the crosstalk and background noise significantly degrade recognition performance (up to 60% relative) and the degradation is more severe in the crosstalk condition. For future reference, we note that the 4-gram perplexity of all the evaluation data is 131.2, 111.1 for AMI, 148.4 for CMU, 102.3 for ICSI, 132.9 for LDC, 143.4 for NIST, and 188.3 for VT meetings.

<i>Condition</i>	<i>Mixing Power</i>	<i>All</i>	<i>AMI</i>	<i>CMU</i>	<i>ICSI</i>	<i>LDC</i>	<i>NIST</i>	<i>VT</i>
Clean	N/A	25.6	19.1	28.0	16.2	29.0	22.8	20.6
Background	1/4	29.1	22.5	31.2	20.7	32.1	26.2	23.3
Crosstalk	1/4	36.4	26.1	41.0	29.5	42.3	35.1	26.5
Background	1/2	30.6	23.4	32.8	22.6	33.1	27.8	24.4
Crosstalk	1/2	38.8	28.2	43.9	33.4	44.5	38.1	28.0
Background	1	32.6	24.6	34.9	25.6	34.6	30.2	24.8
Crosstalk	1	41.7	30.7	47.2	37.7	47.7	41.9	29.1

Table 2. WERs (%) under different recognition conditions. Clean refers to the case when the original IHM audio is used, and crosstalk and background are the crosstalk and background-noise conditions, respectively (cf. Figure 1). Mixing power is the square of the linear mixing coefficient for the interfering channels, assuming a coefficient of 1 for the channel with interference.

Using the time marks of the reference transcriptions obtained from a forced alignment and time marks in the recognition output, we found errors in the nonoverlap regions, and in the single- and two-speaker overlap regions [5]. WERs

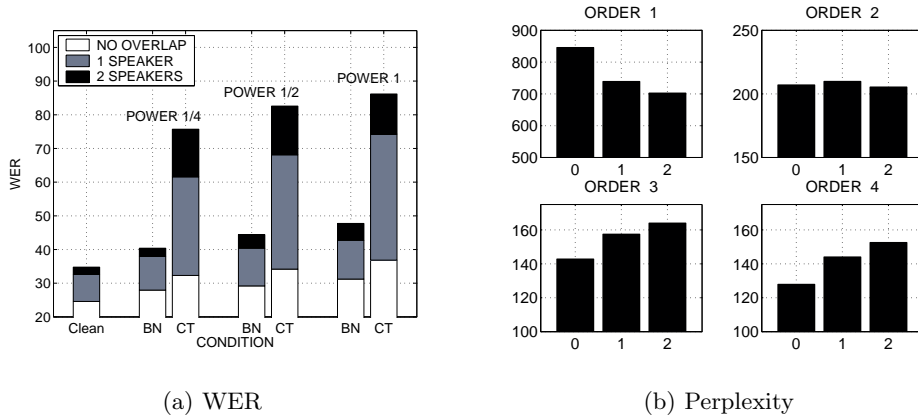


Fig. 2. (a) WERs for the clean, background noise (BN), and crosstalk (CT) conditions with the mixing powers 1/4, 1/2, and 1. For each condition, we display the WER in a stacked fashion for nonoverlaps, and single-speaker and two-speaker overlaps. (b) Perplexities of the foreground reference words during nonoverlaps (0), single-speaker overlaps (1), and two-speaker overlaps (2), for various n -gram LMs.

for each region type were calculated from the number of substitutions, insertions, deletions, and reference words assigned to the regions of that type. WERs for each recognition condition across overlap/nonoverlap types are displayed in Figure 2(a). We discover that crosstalk significantly increases WER, much more so than does background noise, and that two-speaker overlaps cause more errors than single-speaker overlaps. The pattern of results was similar across different sites (not shown here due to space restrictions).

Perplexities for the nonoverlap and single- and two-speaker overlap regions are displayed in Figure 2(b). The perplexities here are those of the reference words corresponding to these regions in the foreground speaker’s speech, since we would like to find out whether the speech from overlaps or nonoverlaps could be inherently more difficult to predict lexically. As shown in Figure 2(b), there is a reversal of the relationship between perplexity and the number of simultaneous speakers. Overlap regions tend to have higher perplexity than those in nonoverlaps if trigrams or 4-grams are used, but the unigram perplexity within overlaps is considerably lower than that of nonoverlaps. (While the perplexities were aggregated over the different sites, individual sites show a similar overall pattern, suggesting robustness of the results.)

An analysis of the frequent n -grams in the test data provided some insight. We found that overlaps contained far more backchannels and discourse markers than nonoverlaps, and the degree of increase for both types of events was larger when the number of simultaneous speakers was higher. Because backchannels are frequent unigrams in LMs trained on spontaneous speech, unigram perplexity is lower when the number of overlapping talkers is higher. The longer n -grams in nonoverlap regions tend to be within-sentence sequences, such as *might be able to*

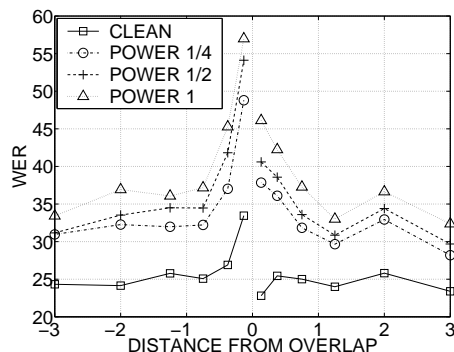


Fig. 3. WERs (%) for clean and crosstalk conditions with various gains, as a function of distance from the overlap (in seconds). Negative distances correspond to before overlaps, and positive distances to after overlaps. Note that the WER at each point represents data coming from that distance only (i.e., results are not cumulative).

and *just a matter of*, which are relatively common in ASR LMs. But, in overlap regions, we see far more cases like *right right right so* and *right i i am*, which are frequent at turn exchanges but not in ASR LMs, since most n -gram tokens come from regions inside single-speaker turns in which the speaker has already obtained the floor. In Section 3.3, we will provide a more detailed quantification of the relationship between the speaker overlaps and dialog acts.

Using the reference and recognition time marks, we looked at recognition errors associated with the nonoverlap regions directly before and after an overlap. We restricted the analysis to the errors that were completely included within such nonoverlap regions, in order to avoid any potential acoustic bias in error rates from the overlap region itself. In Figure 3, we plot WER over such before- and after-overlap regions for different recognition conditions, as a function of the distance from the overlap. As shown, WER decreases as a function of distance from the overlap. In addition, there is an asymmetry in the errors before and after overlaps: WERs are higher before the overlap than after it. This finding is consistent across different recognition conditions and across meetings from different sources (cf. Figure 4) and unlikely to be due to the recognizer itself because the decoding is not strictly forward in time, and not due to reverberation because its effects are smaller, for example, less than 250 ms in the ICSI meetings.

To investigate whether the lower error rate just after the overlap can be attributed to lexical effects, we calculated perplexities of the reference words in these regions (cf. Figure 5). For all n -gram orders, perplexity is lower after overlaps than before them, and the same general pattern holds for each of the different meeting collection sites except VT (cf. Section 2.1 for more observations about VT meetings) so it appears to be a robust finding. Although further investigation is needed, we hypothesize from inspection of the frequent n -grams that the lower perplexity after overlaps stems at least in part from a tendency to begin new sentences at this location.

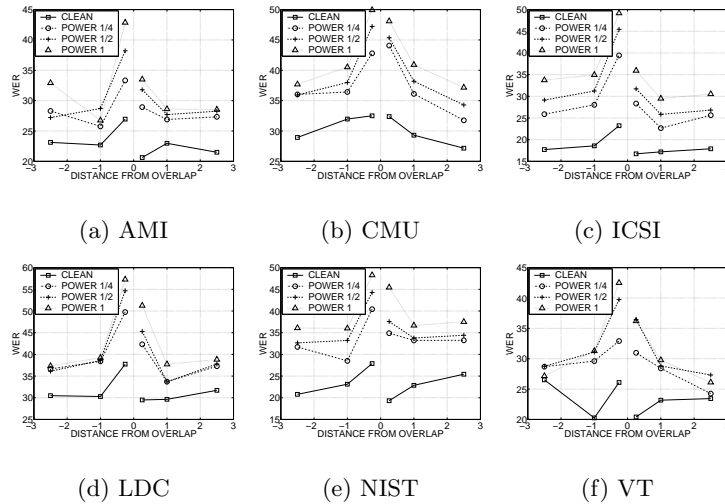


Fig. 4. WERs (%) as a function of time surrounding overlaps for various sites.

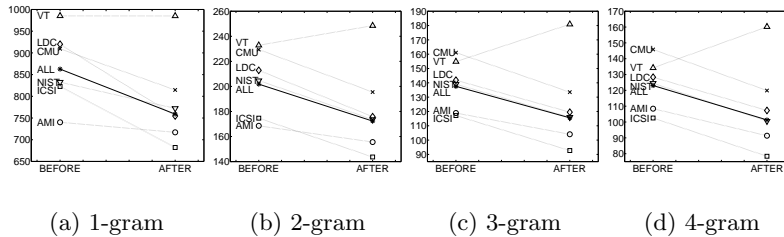


Fig. 5. Perplexities of the foreground reference words before and after overlaps with respect to various n -gram LMs.

3.3 Dialog Acts and Speaker Overlap

To further understand the pattern of results, which so far have treated all speech as one class, we look at various breakdowns of the speech. A first interesting breakdown is by basic dialog act, for example, whether an utterance is a statement, question, backchannel, or other type. As mentioned in Section 3.2, overlaps contain far more backchannels and discourse markers than nonoverlaps, and the degree of increase for both types of events was larger when the number of simultaneous speakers was higher. Here we produce the complete statistics on which speech regions in an overlap analysis are associated with which acts.

Fortunately, we can investigate the relationships between speaker overlaps and dialog acts by using the ICSI MRDA corpus [18], which contains hand-annotations for dialog acts [6] and hot spots [20] for the 75 meetings in the ICSI meeting corpus [9]. These meetings were recorded as part of the same data collection effort, and are similar in style and content to the meetings in the corpus. Roughly 16% of all speech in the annotated ICSI meeting corpus is

<i>Dialog Act</i>	<i>Example</i>
Backchannel	<i>right, uhuh</i>
Disruption	<i>so that's</i> - (self- or other-interruption)
Floor grabber	<i>well i</i> - (context suggests trying to gain floor)
Question	<i>and the new machines are faster?</i>
Statement	<i>it's easier just to buy new disks.</i>
Unlabelable	Unintelligible or some other issue

Table 3. Dialog acts in the ICSI meeting corpus and their examples.

overlapped, which is close to the value of 13% found for the unannotated ICSI evaluation data. The higher rate in the annotated corpus is most likely due to the subtype of ICSI meetings in each set, with the annotated set containing many meetings involving familiar participants who met regularly.

Because we want to know what dialog acts speakers actually produced, we look at human annotations based on reference transcripts. Dialog acts were labeled in detail [6], but collapsed into five broad classes for purposes of these analyses, which are listed in alphabetical order in Table 3. These basic classes have been used in much of the work on automatic detection of dialog acts for this data [2], [1], [8], [22]. Important for these analyses is that the annotation of dialog acts themselves does not depend explicitly on acoustic overlap [6]. For example, a backchannel, such as *uh-huh* can occur either during or after another speaker’s contribution. Similarly, a disruption (uncompleted utterance) can be disrupted by the same or a different speaker. A floor grabber (attempt to gain the floor) can occur during or outside of other speaker’s speech, and is labeled regardless of whether or not the floor is actually obtained.

We use time measures from a forced alignment of the reference transcriptions in the analyses to follow, because the average length of words in a dialog depends on the dialog act (e.g., words in backchannels or floor grabbers tend to be shorter than words in statements or questions). Overall, an average of just over 16% of 53.5 hours of speaking time on a foreground channel is overlapped by one or more other talkers. If we break down this 16% to see what it is made up of in terms of dialog acts, we find that there is a clear association between certain acts and overlap. Table 4 shows expected versus observed results for in-dialog-act times during overlap, and the rate of overlap from the perspective of dialog acts.

We observe in Table 4 that backchannels and fillers are much more likely to occur within overlap regions than would be expected from their distribution overall in the corpus. Disruptions and unlabelable utterances also occur more than expected. The longer, propositional-content-based utterances, questions and statements, are relatively less likely during overlap. Note that the large relative increase for backchannels and fillers is balanced out by a smaller relative increase in statements and questions, because the latter types have more words (and longer words) than the other utterance types. As explained earlier, the hand-coding of dialog acts was not based on whether or not an utterance occurred during overlap. Thus, the biases shown in Table 4 are not predetermined by the hand labels for the dialog acts. Rather, they reflect an association between

<i>Dialog Act</i>	<i>In-Dialog-Act Time</i>			<i>Overlap Time</i>		
	<i>Expected</i>	<i>Observed</i>	<i>Rel. Diff.</i>	<i>Expected</i>	<i>Observed</i>	<i>Rel. Diff.</i>
Backchannel	4.9	13.7	+179.6	16.0	69.5	+333.3
Disruption	12.8	15.7	+22.7	16.0	19.5	+48.0
Floor grabber	1.5	3.8	+153.3	16.0	19.5	+21.7
Question	7.3	5.9	- 19.2	16.0	15.2	- 5.0
Statement	71.5	58.6	- 18.0	16.0	12.4	- 22.5
Unlabelable	1.9	2.3	+21.1	16.0	28.8	+79.4

Table 4. Columns 2-4 display expected versus observed percentages of in-dialog-act times within the 16% of total speaker time that is overlapped. Expected values are based on the distribution of in-dialog-act times for the overall corpus. Columns 5-7 display expected versus observed percentages of overlap time, given a dialog-act class. Expected values are the rate of overlap in the overall corpus. Relative difference percentages are those of the observed values from the expected values.

certain dialog act types in foreground speech when the talker is overlapped, and the functions of these utterances in the meeting.

We can see in Table 4 that the most dramatic act for predicting overlap is the backchannel: If a foreground talker is producing a backchannel, the probability that he is being overlapped by one or more talkers is nearly 70%. Disruptions and unlabelable utterances are the next highest conditional predictors of overlap. One very interesting observation is that floor grabbers are only about 20% more likely to be uttered during overlap than expected. This suggests that when speakers try to grab the floor, they may be trying to do so during silent regions in the other talkers’ speech. The probability of overlap is lowest during statements and questions, suggesting that much of the overlap is not blatant interruption of propositional content, but rather occurs at potential turn-exchange regions in the discourse. This is consistent with long-standing work in conversation analysis [13], [14], [7], [15] but to our knowledge has not previously been analyzed using close study of acoustic overlaps in a large corpus of meeting data.

3.4 Hot Spots and Speaker Overlap

We were also interested in the relationship between overlap and hot spots, or locations in the meetings in which speakers become more affectively involved. The ICSI meeting corpus is hand-labeled for such hot spots, using a procedure described in [20]. Each hot spot consists of one or more utterances across different speakers, and has a number of internal structural and categorical markings (such as start, end, local peaks in hotness, level of hotness, and type of hotness). For purposes of this work, such codings were collapsed, and we asked simply whether an utterance was part of versus not part of a hot spot. Labeling of hot spots tried to capture speaker-normalized animation within utterances, rather than the rate of utterance exchanges. Starts and ends of hot spots were determined by semantic content, but their status as a hot spot relied on individual emotionally salient utterances within a talker. Hot spots were allowed to occur within only one talker’s speech, but in general we assumed that the animation of one speaker tended to produce more interaction with other talkers.

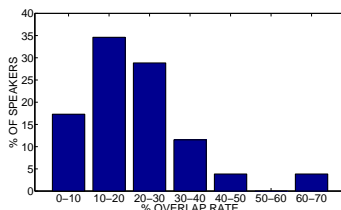


Fig. 6. Rates (%) at which an individual foreground talker is overlapped.

Table 5 shows that there is indeed an association between hot spots and overlap. As shown (see the expected column for hot spots in the corpus overall, under line 2 of the table) hot spots themselves are fairly rare overall in the data, occurring during less than 5% of speaking time. If we look only at overlap regions, hot spots are about 50% more probable. This means that there are many remaining hot spots whose overlap patterns match those of the overall corpus; the "hottness" in these cases must come from aspects of the individual speakers' utterances. Conversely, many overlap regions contain utterances that are not hot, since the 16% rate of overlap for the corpus increases to only 25% when conditioned on utterances in hot spots. Thus, while there is a significant association between hot spots and overlap, they appear to reflect distinct phenomena.

<i>Rate of</i>	<i>Given</i>	<i>Expected</i>	<i>Observed</i>	<i>Rel. Diff.</i>
Overlap	Hot spot	16.0	25.2	+57.5
Hot spot	Overlap	4.8	7.5	+36.0

Table 5. Expected versus observed values for association between overlap and hot spots. Expected values are the overall rate (%) of overlap (line 1) and the overall rate (%) of hot spots (line 2) in the corpus.

3.5 Overlap Rates by Speaker

As a final analysis, we looked at rates of overlap associated with individual speakers. These rates reflect the proportion of time that one or more other talkers overlap with the foreground talker, given that the foreground talker is speaking. We analyzed 52 speakers; the average amount of data per speaker was about an hour, 10 hours for a speaker with the most data. Results are shown in Figure 6. We discover that there is a very large range of behaviors from different talkers. While many speakers cluster near the 16% overlap value for the corpus overall, 20% of the talkers are overlapped by others more than 30% of the time—with two speakers overlapped between 60 and 70% of the time. Such speakers may be producing only backchannels most of the time, or they may be trying to grab the floor while others are talking and not succeeding.

4 Summary and Conclusion

We analyzed overlaps with respect to ASR performance and language model perplexity in regions before, during, and after the overlap. Using an approach

that allowed us to compare the same actually overlapped foreground speech with “clean” and “background-noise” versions, we assessed the relative detriment to ASR of overlapping speech under different crosstalk gain conditions. We found that overlap tends to start at times during which the foreground talker is producing relatively high perplexity word sequences, and that the relationship between perplexity and number of simultaneous talkers is positive for longer n -grams, but negative for unigrams. We discovered a robust asymmetry in ASR error rates before versus after overlaps, apparent across data from the different collection sites. The asymmetry suggests that after being overlapped, the foreground talker temporarily drops to lower-perplexity word sequences, often recycling such events before continuing to talk.

Analyses of a large amount of hand-labeled ICSI meeting data explored the relationship between overlap and content in meetings. Independent dialog act annotations, which did not use overlap as a labeling criterion, showed strong associations with overlap regions. Consistent with classical literature in conversation analysis, but to our knowledge not shown in an automatic analysis of large amounts of meeting data, dialog acts that manage interaction (backchannels, floor grabbers, and disruptions) were positively correlated with overlap, while dialog acts pertaining to propositional content (questions and statements) were negatively correlated. Overlap was also positively correlated with hot spots, or regions of high involvement. Many hot spots however showed default rates of overlap, indicating that speaker involvement ratings are based not only on turn-taking patterns but also on aspects of individual utterances. Finally, individual speakers varied widely in rates of being overlapped; a significant number of speakers showed rates over 30%, with some showing rates over 60%.

Overall, we hope these results illustrate that overlap is an inherent property of natural conversation, and that it shows systematic relationships with word sequences both during and surrounding the overlap. The correlations with word sequences reflect associations at the level of dialog acts, which serve different functions in interaction, as well as at the higher level of hot spots, or greater participant effect. From the engineering perspective, these associations show up as differences in perplexity and WER. Such differences suggest that we may benefit from more intelligent models of overlap in automatic meeting understanding.

Acknowledgments This work is supported in part by AMI (FP6-506811) and CALO (NBCHD-030010) funding at ICSI and SRI, respectively. The opinions and conclusions are those of the authors and not necessarily endorsed by the sponsors.

References

1. J. Ang, Y. Liu, and E. Shriberg, “Automatic Dialog Act Segmentation and Classification in Multi-party Meetings,” In *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing*, pages 1061–1064, 2005.
2. A. Clark and A. Popescu-Belis, “Multi-level Dialogue Act Tags,” In *SIGdial Workshop on Discourse and Dialogue*, pages 163–170, 2004.
3. M. Cooke and D.P.W. Ellis, “The Auditory Organization of Speech and Other Sources in Listeners and Computational Models,” *Speech Communication*, vol. 35, pages 141–177, 2001.

4. Ö. Çetin and A. Stolcke, *Language Modeling in the ICSI-SRI Spring 2005 Meeting Speech Recognition Evaluation System*, Technical Report TR-05-006, ICSI, 2005.
5. Ö. Çetin and E.E. Shriberg, "Speaker Overlaps and ASR Errors in Meetings: Effects Before, During, and After the Overlap," In *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing*, 2006.
6. R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, *Meeting Recorder Project: Dialog Act Labeling Guide*, Technical Report TR-04-002, ICSI, 2004.
7. G. Jefferson, "A Sketch of Some Orderly Aspects of Overlap in Natural Conversation," In G.H. Lerner (ed.) *Conversation Analysis*, pages 43–59, John Benjamins, 2004.
8. G. Ji and J. Bilmes, "Dialog Act Tagging Using Graphical Models," In *Proc. Intl. Conf. on Acoustics, Speech and Signal Process.*, pages 33–36, 2005.
9. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," In *Proc. Intl. Conf. on Acoustics, Speech and Signal Process.*, pages 364–367, 2003.
10. N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke "The Meeting Project at ICSI," In *Proc. Human Language Technologies Conf.*, pages 1–7, 2001.
11. NIST Speech Evaluations, <http://www.nist.gov/speech/tests/index.htm>.
12. T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," *Proc. Automatic Speech Recognition and Understanding Workshop*, pages 107–110, 2001.
13. H. Sacks, E. Schegloff, and G. Jefferson, "A Simplest Semantics for the Organization of the Turn-taking in Conversation," *Language*, vol. 50, pages 696–735, 1974.
14. E. Schegloff, "Recycled Turn Beginnings: A precise repair mechanism in conversation's turn-taking organisation," In G. Button and J.R.E. Lee (ed.'s) *Talk and Social Organisation*, pages 70–85, Clevedon, 1987.
15. E. Schegloff, "Overlapping Talk and the Organization of Turn-Taking for Conversation," *Language in Society*, vol. 29, pages 696–735, 2000.
16. R.T. Schultz, A. Waibel, M. Bett, F. Metze, Y. Pan, K. Ries, T. Schaaf, H. Soltau, M. Westphal, H. Yu, and K. Zechner, "The ISL Meeting Room System," In *Proc. Workshop on Hands-Free Speech Communication*, 2001.
17. E. Shriberg, A. Stolcke, and D. Baron, "Observations on Overlap: Findings and implications for automatic processing of multi-party conversation," In *Proc. European Conf. on Speech Communication and Technology*, pages 1359–1362, 2001.
18. E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," In *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, 2004.
19. A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-Text Evaluation System," In *Proc. NIST RT-05 Meeting Recognition Workshop*, 2005.
20. B. Wrede, S. Bhagat, R. Dhillon, and E. Shriberg. *Meeting Recorder Project: Hot Spot Labeling Guide*, Technical Report TR-05-004, ICSI, 2005.
21. S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and Crosstalk Detection in Multi-channel Audio," *IEEE Trans. on Speech and Audio Processing*, vol. 13, pages 84–91, 2005.
22. M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "A* based Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings," In *Proc. Automatic Speech Recognition and Understanding Workshop*, pages 215–219, 2001.