Analysis of Overlaps in Meetings by Dialog Factors, Hot Spots, Speakers, and Collection Site: Insights for Automatic Speech Recognition

Özgür Çetin† and *Elizabeth Shriberg*†‡

†International Computer Science Institute, Berkeley, CA, USA
‡SRI International, Menlo Park, CA, USA
{ocetin, ees}@icsi.berkeley.edu

Abstract

In previous work we found that automatic speech recognition (ASR) results on meetings show interesting patterns with respect to speaker overlaps, including a robust asymmetry in word error rates (WERs) before and after overlaps. The paradigm used allowed us to infer that these correlations are not due to crosstalk itself but to changes in how a person speaks around overlap regions. To better understand these ASR and perplexity results, we analyze speaker overlaps with respect to various factors, including collection site, speakers, dialog acts, and hot spots.

We examine a total of 101 meetings from the ICSI meeting corpus and the NIST meeting transcription evaluations of the last four years. We find that overlaps tend to occur at high-perplexity regions in the foreground talker's speech. We also find that 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. 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 reveal interesting relationships between overlap and dialog acts, as well as between overlap and "hot spots" (points of increased participant involvement). Finally, results from the ICSI meeting corpus show that individual speakers have widely varying rates of being overlapped.
Index Terms: automatic speech recognition, meeting recognition, crosstalk, speaker overlap, and dialog acts

1. Introduction

Speaker overlap is frequent in natural conversation For example, in the 26 different meetings from the NIST meeting speech recognition evaluations, the 12% of all foreground speaking time is overlapped by speech from one or more talkers [10]. The ratio is even higher (30 to 50%) if one considers pause-delimited regions as units, rather than the actual speaking time [1].

While the detrimental effects of overlap on the ASR performance are well known (e.g., [2], [3], [4], [1], and [5]), there is relatively little work analyzing overlaps with respect to speaker, meeting content, dialog factors, and other conversational phenomena prevalent in meetings. In previous work [10], we discovered that the detrimental effects of overlaps on WER extend multiple seconds before and after an overlap. The WER after the overlap was consistently lower than that before the overlap. This WER asymmetry cannot be attributed to acoustic effects because of the experimental methodology (only simultaneously recorded speech is used to introduce crosstalk), and the forward-backward nature of the decoding architecture. The method used natural overlap from individual speakers' recordings, in three conditions: actual (adding of simultaneous recordings, at various gains), foreground only (no other speaker's recordings), and foreground plus background noise (nonspeech regions from the other speakers' recordings). Thus, the correlations between overlap and ASR must be related to what the foreground speaker is doing in overlap regions.

In this paper, to begin to better understand patterns of speech recognition results in relationship to speaker overlaps reported in [10], we analyze overlaps with respect to the meeting type, content, dialog factors, hot spots, and speakers. We use 75 ICSI meetings that are independently hand-annotated for dialog acts and hot spots, for analysis with respect to meeting content. 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 look at the perplexities of different dialog acts in and around speaker overlaps for this purpose. 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 other speakers.

2. Data

We use about 20 hours of recordings from 26 different meetings from the 2002, 2004, and 2005 NIST meeting speech recognition evaluations. 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. For further analyses requiring human annotations, we use a set of 75 meetings from the ICSI meeting corpus [6]. In separate efforts, this set was extensively hand-marked for dialog acts [7] as well as for hot spots [8].

3. 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 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 nature. 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

Rate	*All*	*AMI*	*CMU*	*ICSI*	*LDC*	*NIST*	*VT*
Overlap	11.6	7.1	13.3	13.0	12.3	11.0	6.1
1 speaker	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 (total overlap time divided by total speaking time). Line 2 provides the percentage of overlaps that involve only two speakers (single-speaker overlap time divided by total overlap time).

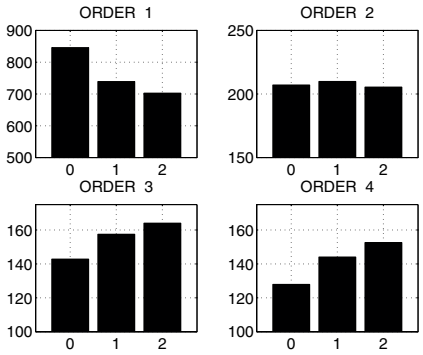

Figure 1: Perplexities of the foreground reference words during nonoverlaps (0), single-speaker overlaps (1), and two-speaker overlaps (2), for various n-gram LMs.

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 other meetings.

4. Perplexity by Overlap Condition

Perplexities for the nonoverlap and single- and two-speaker overlap regions are displayed in Figure 1, using a language model (LM) trained on a variety of meeting data (excluding the analysis data), conversational speech, broadcast news and Web data [9]. The perplexities 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 1, 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* 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.

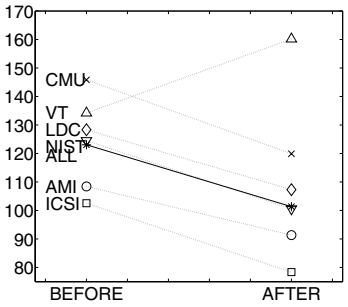

Figure 2: Perplexities of the foreground reference words before and after overlaps with respect to a 4-gram LM.

In Section 5, we will provide a more detailed quantification of the relationship between the speaker overlaps and dialog acts.

In [10], we found that WER decreases as a function of distance from the overlap, and there is an asymmetry in the errors before and after overlaps: WERs for the data analyzed here were higher before the overlap than after it. 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 shown in Figure 2. For all n-gram orders, perplexity is lower after overlaps than before them. The same pattern holds for nearly all sites, so it appears to be a robust finding. The one exception is the VT meetings, which show a markedly different pattern, suggesting that there is something unusual about interaction behavior in that particular data set.

5. 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 some other type. In general, overlaps contain far more backchannels and discourse markers than nonoverlaps. Here we produce the complete statistics on the association between overlap/nonoverlap regions and acts.

Fortunately, we can investigate the relationships between speaker overlaps and dialog acts by using the ICSI MRDA corpus [11], which contains hand-annotations for dialog acts [7] and hot spots [8] for the 75 meetings in the ICSI meeting corpus [6]. 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 overlapped, which is close to the 13% overlap rate 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 [7], but collapsed into five classes for purposes of these analyses (backchannel, disruption, floor grabber, question, statement, and unlabelable for unintelligible or some other issue). Important for these analyses is that the annotation of dialog acts themselves does not depend explicitly on acoustic overlap [7]. For example, a backchannel 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 the other speaker's speech, and is

Dialog Act	*In-Dialog-Act Time*		
	Expected	*Observed*	*Rel. Diff.*
Backchannel	4.9	13.7	+179.6
Disruption	12.8	15.7	+22.7
Floor grabber	1.5	3.8	+153.3
Question	7.3	5.9	– 19.2
Statement	71.5	58.6	– 18.0
Unlabelable	1.9	2.3	+21.1

Table 2: 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.

Dialog Act	*Overlap Time*		
	Expected	*Observed*	*Rel. Diff.*
Backchannel	16.0	69.5	+333.3
Disruption	16.0	19.5	+48.0
Floor grabber	16.0	19.5	+21.7
Question	16.0	15.2	– 5.0
Statement	16.0	12.4	– 22.5
Unlabelable	16.0	28.8	+79.4

Table 3: Expected vs. observed percentages of overlap time (%), given a dialog act class. Expected values are the rate of overlap in the overall corpus.

labeled as such, regardless of whether or not the floor is 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). If we break down the overlapped speech 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 2 shows expected versus observed results for in-dialog-act times during overlap, and Table 3 shows the rate of overlap from the perspective of dialog acts.

We observe in Table 2 that backchannels and fillers are much more likely to occur within overlap 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. 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 and longer words than the other utterance types. Note that the biases shown in Table 2 are not predetermined by the hand labels for the dialog acts, because the hand-coding of dialog acts was not based on whether or not an utterance occurred during overlap.

We can see in Table 3 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 (e.g., [12], [13], [14]) but to our knowledge has not previously been analyzed using close study of acoustic overlaps in a large corpus of meeting data.

Dialog Act	*# of overlaps*			*Around Overlaps*	
	0	*1*	*2*	*Before*	*After*
Backchannel	8.48	8.54	8.59	8.41	8.72
Disruption	1.60	2.05	2.21	1.66	1.55
Floor grabber	8.47	8.76	8.64	8.68	8.28
Question	1.61	2.21	2.51	1.66	1.54
Statement	1.18	1.28	1.29	1.17	1.18
Unlabelable	2.65	3.83	4.25	2.72	2.27

Table 4: Columns 2–4 display the bigram dialog-act perplexities in nonoverlap regions (0), and single-speaker overlaps (1), and two-speaker overlaps (2). Columns 5–7 display perplexities for the nonoverlap regions just before and after an overlap.

To better understand the relationship between dialog acts and overlaps, we calculated the dialog act perplexities in overlap and nonoverlap regions, shown in Table 4. Here the perplexity is calculated by replacing words by their dialog act tags, and placing sentence begin and end tokens at the dialog act boundaries. The bigram dialog act language models are estimated by leave-one-meeting-out cross-validation (the results for higher-order n-grams were very similar). We observe that consistent with the word perplexity patterns in Figure 1, the dialog act perplexities increase with the number of overlapping speakers for all dialog acts. Also consistent with the word perplexity patterns in Figure 2, the dialog act perplexity for all dialog acts but backchannels and statements is significantly higher before than after the overlap; backchannels show an opposite pattern, and the perplexity of statements does not change significantly before and after the overlap.

6. 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 [8]. 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.

Table 5 shows that there is indeed an association between hot spots and overlap. As shown (see the expected column 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 "hotness" in these cases must come from aspects of the individual speakers' utterances. Conversely, many overlap regions contain utterances that

Rate of	*Given*	*Expected*	*Observed*	*Rel. Diff.*
Overlap	Hot spot	16.0	25.2	+57.5
Hot spot	Overlap	4.8	7.5	+36.0

Table 5: Expected vs. observed rates (%) 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.

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 an association between hot spots and overlap, they appear to reflect distinct phenomena.

7. Overlap Rates by Speaker

As a final analysis, we looked at rates of overlap for individual speakers. These rates reflect the proportion of time that the other talkers overlap with the foreground talker, given that the foreground talker is speaking. We analyzed 52 speakers in the ICSI corpus; the average amount of data per speaker was about an hour, 10 hours for the speaker with the most data. Results are shown in Figure 3. We see 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 may be trying to grab the floor while others are talking, and not succeeding.

8. Summary and Conclusion

We analyzed overlaps with respect to meeting type, content, dialog factors, hot spots, and speakers. 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 language model perplexity 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

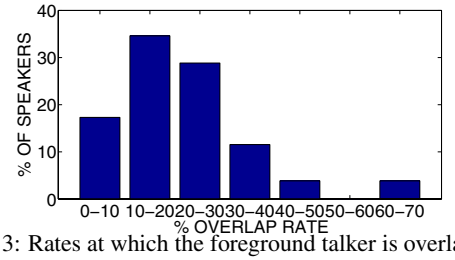

Figure 3: Rates at which the foreground talker is overlapped.

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, explained by well-known discourse principles. Such differences suggest that we may benefit from more intelligent models of overlap in automatic meeting understanding. For example, for better ASR, one can apply different LMs or adapt an existing LM depending on overlap. Another interesting application is to know how to best use training data from one corpus for testing on another, since meetings can have very different overlap patterns. The fact that we saw a significant difference in the VT meeting patterns even though they had a reasonable rate of overlap, provides a cue to mismatch with the other types, and to unnaturalness. For an automatic meeting participant, we may want to mimic these patterns around overlap so that the language generation sounds natural.

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

9. References

[1] E. Shriberg et al., "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. EUROSPEECH*, 2001, pp. 1359–1362.

[2] M. Cooke and D.P.W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, pp. 141–177, 2001.

[3] T. Pfau et al., "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. ASRU*, 2001, pp. 107–110.

[4] R.T. Schultz et al., "The ISL meeting room system," in *Proc. Workshop on Hands-Free Speech Communication*, 2001.

[5] S. Wrigley et al., "Speech and crosstalk detection in multi-channel audio," *IEEE Trans. on Speech and Audio*, vol. 13, pp. 84–91, 2005.

[6] A. Janin et al., "The ICSI meeting corpus," in *Proc. ICASSP*, 2003, pp. 364–367.

[7] R. Dhillon et al., *Meeting recorder project: Dialog act labeling guide*, Tech. Rep. TR-04-002, Intl. Computer Science Inst., 2004.

[8] B. Wrede et al., *Meeting recorder project: Hot spot labeling guide*, Tech. Rep. TR-05-004, Intl. Computer Science Inst., 2005.

[9] A. Stolcke et al., "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Proc. NIST RT-05 Meeting Recognition Workshop*, 2005.

[10] Ö. Çetin and E. Shriberg, "Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap," in *Proc. ICASSP*, 2006.

[11] E. Shriberg et al., "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. 5th SIGdial Workshop*, 2004, pp. 97–100.

[12] H. Sacks et al., "A simplest semantics for the organization of the turn-taking in conversation," *Language*, vol. 50, pp. 696–735, 1974.

[13] G. Jefferson, "A sketch of some orderly aspects of overlap in natural conversation," in *Conversation Analysis*, G.H. Lerner, Ed., pp. 43–59. John Benjamins, 2004.

[14] E. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, pp. 696–735, 2000.