Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation

Elizabeth Shriberg[1,3] *Andreas Stolcke*[1,3] *Don Baron*[2,3]

[1]SRI International, Menlo Park, CA [2]University of California at Berkeley, Berkeley, CA
[3]International Computer Science Institute, Berkeley, CA
{ees,stolcke,dbaron}@icsi.berkeley.edu

Abstract

We examine the distribution of overlapping speech in different corpora of natural multi-party conversations, including two types of meetings, and two corpora of telephone conversations. Analyses are based on forced alignment and speech recognition using an identical recognizer across tasks. Three results are discussed. First, all corpora show high overall rates of overlap, with similar rates for meetings and telephone conversations. Second, speech recognition performance in non-overlapped regions of meetings is no worse than that in single-channel telephone conversations, while recognition in overlap regions degrades considerably. Finally, interrupt locations are associated with endpoints of word-level events in a speaker's turn, including backchannels, discourse markers, and disfluencies. Results suggest that overlap is an important inherent characteristic of conversational speech that should not be ignored; on the contrary, it should be jointly modeled with acoustic and language model information in machine processing of conversation.

1. Introduction

Recent interest in automatic recognition and understanding of multi-party conversation such as meetings [10, 7] raises new problems related to highly frequent speaker overlap. Such overlap has serious consequences for processing models at all levels—from speech recognition to understanding to dialog modeling. Overlap has long been noted as a characteristic of natural conversation, particularly by researchers in conversation analysis and related fields (e.g., [4]). Little attention, however, has been paid to the phenomenon in work on automatic speech processing. For example, speech researchers have processed Switchboard [3] data by considering only one channel at a time. Dialog work on the same corpus has considered both channels, but imposed a strict linear ordering on speaker turns based on start times [5].

This paper provides an investigation into overlap characteristics of large databases of multi-party conversational speech, based on automatic time alignments. Data are drawn from a corpus of real meeting data, collected at the International Computer Science Institute (ICSI), as well as from two large corpora of telephone conversations. The study is admittedly crude in nature, compared with studies of hand-labeled data in conversation analysis. Nevertheless, because of the large scale of the data analyzed, and the potential for automatic processing using our simple categories, we hope that the general results will help to raise "overlap awareness" for the processing of conversational speech by machine.

2. Method

2.1. Speech data and transcriptions

We processed and analyzed data from three sources. Multi-party meetings were collected as part of the ICSI Meeting Project [7]. We drew data from two series of real-life group meetings, identified as "Meeting Recorder" (MR) and "Robustness" (ROB), with 4 to 8 participants each. Participants had been recorded on a variety of microphones, including close-talking and far-field types. For the present study we used data only from head-mounted and lapel microphones. For comparison we also used data from two corpora of two-person telephone conversations. The Switchboard (SWB) corpus [3] contains strangers talking about assigned topics, whereas the CallHome English (CH) corpus involves conversations between family members and friends. We used the retranscribed and resegmented version of Switchboard from [1]. The amount of data from each source is summarized in Table 1. (The Meeting Project has already collected an order-of-magnitude more data than we were able to use, the limiting factor here being the availability of transcriptions.)

Table 1: Types and amounts of data used in the study: Meeting Recorder meetings (MR), Robustness meetings (ROB), CallHome English (CH), and Switchboard (SWB). Speech duration includes overlapped speech multiple times. The notion of speech "spurt" is defined in Section 2.4.

	Meetings		Phone convs.	
	MR	ROB	CH	SWB
Meetings/Convs.	5	3	100	2437
Speech duration	7.7h	4.8h	16.6h	315h
Transcribed words	60,403	32,384	202,766	3,051,068
Speech spurts	5,688	4,100	23,693	298,825

2.2. Speech segmentation

Individual channel recordings were partitioned into "segments" of speech, based on a "mixed" signal (addition of the individual channel data, after overall energy equalization by channel). Segment boundary times were determined either by an automatic segmentation of the mixed signal followed by hand-correction, or by hand-segmentation alone. For the automatic case, the data were segmented with a speech/nonspeech detector consisting of an extension of an approach using an ergodic hidden Markov model (HMM). Currently, for simplicity and to debug the various processing steps, these segments are synchronous across channels.

2.3. Automatic speech recognition and alignment

The recognizer used in experiments was a subset of the March 2000 SRI large-vocabulary conversational speech recognition (LVCSR) system [9]. The system performs vocal-tract length normalization, feature normalization, and phone-loop-based speaker adaptation using all the speech collected on each channel (i.e., from one speaker, modulo crosstalk), and a bigram language model of about 30,000 words, trained from Switchboard,

CallHome, and Broadcast News data. As an expedient, we omitted more elaborate acoustic and language modeling which yield about a 20% relative error rate reduction on Hub-5 data. Notably, both the acoustic models and the language model of the recognizer were identical to those used in the Hub-5 domain (the acoustic front end downsamples the wide-band signal to telephone bandwidth). This allows us to compare results directly across corpora.

The same system without speaker adaptation was used for forced-aligning the reference transcripts to the acoustic meeting data from each speaker's channel, thereby giving us the approximate locations of the foreground speech on each channel. We hand-checked a sample of the segments resulting from forced alignment, and estimated the combined word error rate due to faulty hand-transcriptions and/or automatic alignment to be about 7% for meeting data. Since each speech region analyzed for overlap must contain at least one word, this figure also gives us a (generous) upper bound on the error in the overlap measures discussed in the next section. Switchboard transcriptions and segmentations should have a very low error rate, as they underwent extensive quality control procedures [1], but the accuracy of CallHome transcripts is not known at this point (although we expect it to be better than meeting data).

Initially, we expected that crosstalk between microphones could be reduced or eliminated by a simple adaptive filter tracking the cross-coupling between speakers as recorded by their own head-mounted microphones and picked up on the other channels. A colleague [2] investigated using a Block Least Squares algorithm [11] to estimate the coupling and cancel the crosstalk; however, the problem was more complex than we expected. The head movements of either the speaker or the listener (i.e., the crosstalk pickup microphone) result in very significant changes in the required cancellation filter. It appears that in a meeting environment significant movements are both fast and frequent. Informal observation suggests that they often closely follow the start of a new speaker (as listeners turn to face that speaker), thereby making it very hard to cancel at least the first portion of these intrusions. Our group is continuing to study this problem in the hope of generating "purer" recordings of individual speakers.

2.4. Overlap measures

We considered speech from each foreground talker, in turn. For that talker, we defined "overlapped" words as words spoken while one or more other talkers were also speaking during some portion of the word. Overall overlap rates were weighted averages of these foreground-speaker measures. We also computed a rate based on "spurt" units, where spurts were defined as *speech regions uninterrupted by pauses longer than 500 ms*. Rates of overlapping spurts were computed in a manner similar to that for overlapping words; spurts were considered overlapped if background speech was present at any time during the spurt. These measures are illustrated in Figure 1. We also created versions of these measures in which backchannels (such as "uh-huh") were excluded, i.e., effectively treated as non-speech regions.

The word transcripts were also annotated (by a combination of hand-labeling and automatic methods) to indicate "hidden" locations of interest, including sentence boundaries and disfluency boundaries, and special word types such as filled pauses, coordinating conjunctions, and discourse markers that are used to manage interaction in conversation [6]. Such locations are hidden from the point of view of the speech recognizer, which outputs only a word stream, but are potentially automatically detectable using a combination of lexical and prosodic information [8].

Table 2: Relative frequencies of overlapped speech in different corpora. Frequencies are given as percentages of overlapped words (in plainface) and "spurts" (in italics).

	Meetings		Phone convs.	
Backchannels	MR	ROB	CH	SWB
Included				
words	17.0	8.8	11.7	12.0
spurts	*54.4*	*31.4*	*53.0*	*54.4*
Excluded				
words	14.1	5.6	7.9	7.8
spurts	*46.4*	*21.0*	*38.8*	*38.9*

3. Results and Discussion

3.1. Overall rates of overlap across corpora

Table 2 summarizes rates of word and spurt overlaps across corpora. Comparing meetings to two-party telephone conversations, we see that meetings are not special in terms of overlap. Telephone conversations fall somewhere in between the two meeting types, with MR meetings containing more overlaps than phone conversations, and ROB meetings containing fewer. Furthermore, overlap is an issue for modeling conversations even in the corpora the speech community has been using; their pervasiveness has so far been hidden by the fact that current systems typically process Switchboard and CallHome as isolated individual channels. All the rates we found are also significantly higher than what has been reported by researchers in conversation analysis (less than 5% according to [6, p. 296], although methodological differences may account for some of the discrepancy).

Second, Switchboard and CallHome are very similar to each other on all measures, even though it has been conjectured that there is more overlap in CallHome. This conjecture has been based on the assumption that there are more overlaps among friends and people one is familiar with, than among strangers. CallHome contains family and friends on free real-world phone calls with time limits, so one could expect overlaps to maximize use of the time. In Switchboard, on the other hand, strangers with no set task were asked to talk for a certain length of time. Furthermore, it is not the case that Switchboard overlaps are not just due to polite backchanneling, since the rates with backchannels removed are still nearly identical for the two corpora.

Third, we see a sizable difference in overlap rates depending on meeting type. The MR and ROB meetings were chosen from the set of meeting types collected at ICSI because they represent two very different types of social interaction. In the first, there is a fairly open exchange between many of the participants; in the second, one speaker directs the flow of the meeting. In MR meetings, high rates of overall words come from 3 to 5 main participants per meeting, whereas in the ROB meetings, 56% of the total words are attributable to the main speaker (thus there is less opportunity for overlap).

3.2. Overlap and ASR

In the absence of effective signal cross-cancellation—which, as mentioned earlier, is a nontrivial problem in meeting settings—background speech can degrade ASR performance significantly. Background speech in regions where the foreground speaker is silent will be recognized as foreground speech, generating a large number of insertion errors.

Foreground speaker: (pause) <spurt$_A$> So that's the scoop. </spurt$_A$> (pause) <spurt$_B$> Let's move on. </spurt$_B$>
Background speaker 1: <ba>Uh-huh. </ba> Great!
Background speaker 2: <dm> Well, </dm> I'm not sure

Figure 1: Illustration of the word and spurt overlap classification scheme. Overlap rates for words and spurts are computed relative to speech from the foreground speaker. In the figure, word beginnings are spatially aligned to indicate synchrony in time. In this example, the overlapped foreground words are: "that's the scoop". Spurt A is overlapped, while spurt B is not. Backchannels are indicated by a <ba> tag, discourse markers by <dm>. With backchannels excluded from the overlap computation, "scoop" remains as the only overlapped foreground word.

Table 3: Recognition performance in various conditions of overlapping speech, percent word error rate. Word counts reflect the foreground reference transcripts only, excluding background speech.

Overlap condition	Headset	Lapel	Total	#Words
Non-overlap segs.	41.7	38.7	41.3	36,532
Overlap segs.	50.4	70.4	53.4	38,524
Non-overlap+overlap	46.1	57.4	47.5	75,056
Foreground speech	42.5	43.3	42.6	75,056
No-background speech	45.0	47.6	45.3	64,508

To analyze the effect of overlapped speech we split the test set into regions containing non-overlapped and overlapped speech, and into speakers using head-mounted close-talking microphones versus those wearing lapel microphones. The latter are much more prone to pick up other speakers and were thus expected to exacerbate the effects of overlaps. Non-native speakers were excluded from the study to ensure comparability with ASR performance on well-known corpora, in particular so we could compare results with Switchboard recognition.

There were four overlap test conditions. "Non-overlap segments" includes segments of speech with only the foreground speaker talking (as per the hand-segmentations described earlier). "Overlap segments" comprise all segments containing speech from more than one speaker (not necessarily overlapping over the entire duration of the segment). "Foreground speech" includes all speech by the foreground speaker, excluding regions with only background speech. "No-background speech" includes all regions with only foreground speech and no background speech, i.e., all the "non-overlap segments" plus portions of the "overlap segments" free of background speech. The extent of foreground speech in overlap segments was determined automatically using forced alignments of the reference transcripts, and is thus somewhat prone to errors. We added 50 ms around automatically determined foreground speech regions, but note that this could also allow extra background speech into the test regions.

Table 3 summarizes the results in all test conditions. As shown, there is indeed a significant increase in error rate in segments containing overlapping speech (+12% absolute). The increase is especially marked (+32%) on lapel microphones, indicating that crosstalk is to blame for the degradation. Furthermore, if the scoring only considers speech regions tightly bounded around the foreground speech, the error rate is almost that of the non-overlapping segments. We can infer from this result that the high error rate in the overlap segments stems from recognized background speech, not from the foreground speech being harder to recognize. This is confirmed by the "non-overlapping" result, where we scored only over regions without background speech, without a further improvement in accuracy.

A breakdown of the word recognition errors by type (into substitutions, deletions, and insertions) further confirms that segments containing background speech, especially with lapel microphones, suffer from excessive insertions, as shown in Figure 2. Inspection of recognition output shows that these insertions do indeed correspond to the background speakers' utterances. Figure 2 also provides a comparison with Switchboard and CallHome recognition accuracy, using the same recognition system, and measured on subsets of the 2001 and 2000 Hub-5 LVCSR development test sets, respectively. Remarkably, recognition performance using Hub-5 acoustic and language models (and telephone-band-limited meeting recordings) is on par with accuracy on matched Hub-5 data, if we exclude background speech regions. It is quite possible that background speech has subtle effects on the speech production of the foreground speaker (cf. the Lombard effect in noisy conditions), but at least at current error rates this does not seem to affect ASR performance.

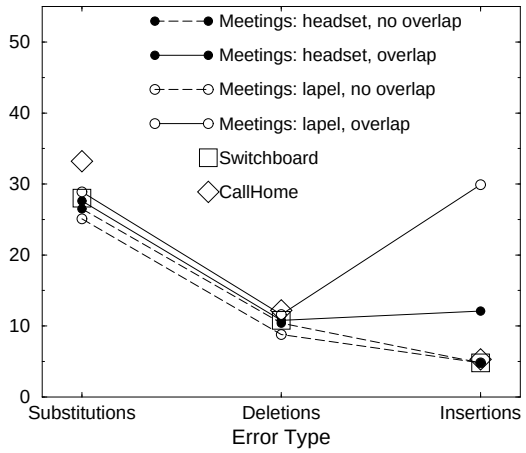

Figure 2: Comparison of recognition error types across overlap conditions, microphone types, and corpora.

Another important conclusion is that Switchboard seems to be representative of the acoustic-phonetic and stylistic properties of conversational speech even in other settings, including meetings, making it a good target for continued research in large-vocabulary recognition. This leaves us hopeful that with increasing amounts of matched meeting data for training, considerably better meeting recognition will be obtained, provided effective foreground speech detection can be achieved.

3.3. Locations of overlap

We also examined the location of "interrupts" in the meeting data. We define interrupts as *within-sentence locations in the speech of the current speaker at the time points at which another speaker begins a sentence*. Interrupts are thus defined to occur only at within-sentence locations in the foreground speech. The majority of interrupts (72.9%) occur at spurt boundaries, as might be expected, since spurts are defined to be followed by pauses. However, the remaining 27% of interrupts—a fairly high number—occur within spurts, when speakers are talking continuously. This suggests that it is not sufficient to allow speaker change only at pause boundaries: over one quarter of interrupting sentence onsets would be lost that way.

To further examine where interrupts occur, we looked at their distribution with respect to the location of certain common events, including backchannels (e.g., "uh-huh"), coordinat-

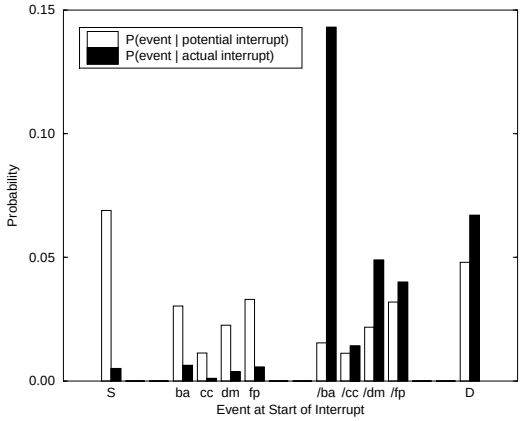

Figure 3: Distribution of events overall (white bars) and at interruption points (black bars). 'S" = sentence onset, "ba" = backchannel, "cc" = coordinating conjunction, "dm" = discourse marker, "fp" = filled pause, "D" = repetition, repair or false start boundary. For the lower-case events, no "/" = start of event, "/" = end of event.

ing conjunctions (e.g., "and"), discourse markers (e.g., "well"), filled pauses ("uh" and "um"), and interruption points in disfluencies (e.g., repetitions, repairs, and false starts). In the case of sentence and disfluency boundaries, there is only one event location; in the case of the other events, we recorded both the location preceding the event, and that following the event. Note that events are not mutually exclusive, for example:

<S> <dm> well </dm> <fp> uh </fp> i
<D> i think that's great </S>

The start-event tags are associated with utterance onsets, since many of these elements (discourse markers, filled pauses, coordinating conjunctions) are used to start or hold a turn. Points right after these elements (i.e., the end-markers for the same cases) are typically locations in which the speaker has obtained the floor, but may pause before continuing. Disfluency boundaries (<D>) are also in this category. We might expect that at such locations, it would be somewhat rude to interrupt. However, this is not what we found.

Figure 3 shows the distribution of events at interrupt locations, and compares the event distribution to the overall distribution, i.e., over all locations at which an interrupt could occur. A clear pattern in the figure is that while interrupts are dispreferred at the onsets of the events, there is a strong tendency to interrupt right after the same events, even though the speaker may still be "holding the floor" from a strictly lexical analysis. Because overlaps are highly associated with hidden events, an overall model of speaker segmentation and overlap detection could benefit from an integrated approach. Since the main preference for overlap locations coincides with the *ends* of such events, it is reasonable to propose that online event detection could benefit overlap detection by continuously updating speech boundary locations for the probability of speech onset from a new speaker.

4. Conclusions

We have studied overlap in four different styles of conversational speech data. Results show that both meetings and telephone conversations have high rates of overlap, suggesting that overlap is an inherent characteristic that should not be ignored in computational models of conversation. Results on word error rates using the same speech recognition system across tasks, reveal that in regions of no overlap, recognition performance for real meetings is similar to that for telephone conversations. Speakers must therefore be using fairly consistent pronunciation patterns in meetings and telephone conversations—implying that progress on recognition of telephone speech should benefit recognition of meeting speech, and vice versa. Recognition of meetings does suffer from overlaps, even on close-talking microphone data. The errors are in the form of insertions, which should be partially addressable by cross-cancellation techniques, but which present an important challenge for further research. Finally, interrupts do not occur in random locations, but rather are associated with hidden events (such as disfluencies and discourse markers) in the foreground speech. The interrupts tend to start after such events, suggesting an integrated acoustic/language model for speaker segmentation in natural conversation.

5. Acknowledgments

We thank our ICSI collaborators: Nelson Morgan, Chuck Wooters, Jane Edwards, Dan Ellis, Dave Gelbart, Adam Janin, and Thilo Pfau. Thilo Pfau conducted work on segmentation; Dan Ellis (also Columbia U.) conducted work on cross-cancellation. This work was funded under a DARPA Communicator project (via U. Washington), supplemented by an award from IBM, by SRI's LVCSR project, and by SRI's NSF-STIMULATE IRI-9619921.

6. References

[1] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone. Resegmentation of Switchboard. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, pp. 1543–1546, Sydney, 1998. Australian Speech Science and Technology Association.

[2] D. Ellis. ICSI Meeting Recorder Project: Notes on direct crosstalk cancellation. http://www.icsi.berkeley.edu/~dpwe/-research/mtgrcdr/blscancel.html.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP*, vol. 1, pp. 517–520, San Francisco, 1992.

[4] G. Jefferson. A case of precision timing in ordinary conversation: overlapped tag-positioned address terms in closing sequences. *Semiotica*, 9:47–96, 1973.

[5] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 88–95, Santa Barbara, CA, 1997.

[6] S. C. Levinson. *Pragmatics*. Cambridge University Press, Cambridge, 1983.

[7] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The ICSI Meeting Project. In *Proceedings of the Human Language Technology Conference*, San Diego, 2001.

[8] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, 2000. Special Issue on Accessing Information in Spoken Audio.

[9] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng. The SRI March 2000 Hub-5 conversational speech transcription system. In *Proceedings NIST Speech Transcription Workshop*, College Park, MD, 2000.

[10] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting Browser: Tracking and summarizing meetings. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 281–286, Lansdowne, VA, 1998. Morgan Kaufmann.

[11] E. Woudenberg, F. Soong, and B. Juang. A block least squares approach to acoustic echo cancellation. In *Proc. ICASSP*, vol. 2, pp. 869–872, Phoenix, AZ, 1999.