SPEAKER OVERLAPS AND ASR ERRORS IN MEETINGS: EFFECTS BEFORE, DURING, AND AFTER THE OVERLAP

Özgür Çetin†

†International Computer Science Institute
Berkeley, CA
ocetin@icsi.berkeley.edu

Elizabeth Shriberg†‡

‡SRI International
Menlo Park, CA
ees@speech.sri.com

ABSTRACT

We analyze automatic speech recognition (ASR) errors made by a state-of-the-art meeting recognizer, with respect to locations of overlapping speech. Our analysis focuses on recognition errors made both *during* an overlap and in the regions immediately *preceding* and *following* the location of overlapped speech. We devise an experimental paradigm to allow examination of the same foreground speech both with and without naturally occurring cross-talk. We then analyze ASR errors with respect to a number of factors, including the severity of the cross-talk and distance from the overlap region. In addition to reporting effects on ASR errors, we discover a number of interesting phenomena. First, we find that overlaps tend to occur at high-perplexity regions in the foreground talker's speech. Second, word sequences within overlaps have higher perplexity than those in nonoverlaps, if using trigrams or 4-grams, but the unigram perplexity within overlaps is considerably lower than that of nonoverlaps. An explanation for this behavior is proposed, based on the preponderance of multiple short dialog acts found in overlap regions. Third, we discover that the word error rate (WER) after overlaps is consistently lower than that before the overlap. This finding cannot be explained by the recognition process itself; rather, the foreground speaker appears to reduce perplexity shortly after being overlapped. Taken together, these observations suggest that the automatic modeling of meetings could benefit from a broader view of the relationship between speaker overlap and ASR in natural conversation.

1. INTRODUCTION

Speaker overlap is frequent in natural conversation, especially if one considers units such as dialog acts, stretches of pause-delimited speech, or speaker turns. For example, in a study of overlap in both telephone conversations and multiparty meetings, it was found that 30% to 50% of all speech *spurts* (regions of speech in which a particular talker does not pause for more than half a second) include one or more frames of simultaneous speech by another talker [7]. As described in classic work on conversation analysis [5], speakers do not alternate sequentially in a conversation, but rather they predict the end of a current speaker's turn using syntax, semantics, and prosody, and often start speaking before the current speaker finishes.

In this work we examine overlap and its effects on ASR for speech from recorded meetings. We hypothesize that overlaps could affect recognition performance not only because of the well-known effects of acoustic cross-talk, but also because speech near overlaps could be inherently different in style or content from speech elsewhere. We also hypothesize that the effect of overlaps may not be confined to the regions in which they actually occur, but rather that the effects extend in time before and/or after the overlap itself. Note that we will use the term *overlap* to indicate situations in which multiple talkers are speaking simultaneously. We will use the term *cross-talk* to indicate cases in which a microphone associated with one talker picks up the voice of another speaker during an overlap.

While general effects of overlap are well reported in the literature (e.g., [1, 3, 4, 7]), there is relatively little work quantifying such effects under the different conditions that we consider. 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 analyze both the errors made during overlaps, and errors made in nonoverlap regions directly before and after an overlap. We examine various factors, including the number of speakers involved in the overlap, the presence or absence of cross-talk and its severity, and the time distance from the overlap. We also look at language model perplexity, as well as at words associated with particular discourse roles (filled pauses, backchannels, and discourse markers), in an attempt to better understand the pattern of results.

2. METHOD

2.1. Data

We use 19.8 hours of recordings from 26 different meetings from the 2002, 2004, and 2005 NIST meeting speech recognition evaluations. These meetings were from AMI (2), CMU (6), ICSI (6), LDC (4), NIST (6), and VT (2), with the number of meetings from each source 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. About 88%, 11%, and 1% of the speech frames are from nonoverlaps, single-speaker overlaps (i.e., one additional speaker), and two-speaker overlaps, respectively, as determined from a forced alignment of the reference transcripts.

2.2. Recognition System

Recognition experiments are conducted using the 2005 ICSI-SRI meeting system [8]. This system is adapted from SRI's conversational telephone speech system to the meeting domain using a variety of meeting data (including about 72 hours from the ICSI meeting corpus, excluding our test data). 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. See [8] for full details. To avoid confounds with automatic speech segmentation, we use manual reference segmentations in our experiments, as our goal is to study the effects of overlaps on automatic speech recognition.

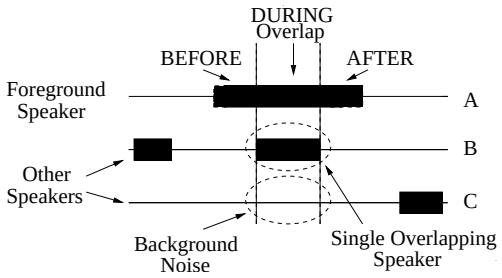

Fig. 1. Illustration of experiment conditions. When A is taken as the foreground speaker, B and C are background speakers. For the cross-talk condition, full original audio from B and C are added to A. For the background-noise condition, B and C are added only in the cases in which they do not contain any speech (for example, during the overlap marked DURING, B is not added to A, and only C is added). The regions marked BEFORE and AFTER in A are nonoverlaps. Solid rectangles denote speech segments, obtained from a forced alignment of reference segmentations.

2.3. Experiment Conditions

Since our goal is to study effects of overlaps on the ASR performance, we need a principled method for controlling cross-talk and its severity. To the best of our knowledge, there are no large publicly available data sets with careful recording of the same speech with and without cross-talk. In addition, the signal processing methods for introducing or removing cross-talk are imperfect, especially for the purposes of this study. Instead, we use synchronously recorded speech from IHMs and speech/nonspeech alignments to create a rendition of cross-talk that is accurate in terms of speech that has overlapped and cross-talk severity. Speech activity regions are defined to be consecutive segments of spoken words uninterrupted by pauses longer than 0.5 secs (the same definition as used for *spurts* in [7]).

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 weighting by a factor to adjust cross-talk severity. We refer to the recognition with such modified audio as the 'cross-talk condition'. When a particular channel is added to another one, in addition to the speaker's voice in that channel, any background noise that is also captured by that channel is added as well. 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. We will refer to this recognition condition as the 'background-noise condition'. The performance differences between the cross-talk and background-noise conditions should indicate the cross-talk 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 cross-talk condition contains *only speech that actually occurred at the same time*. (We do not create cross-talk using speech from different corpora or time spans!) Nevertheless, the waveform addition is admittedly simplistic and does not capture some aspects of cross-talk such as nonlinear frequency weighting, room geometry, and reverberation. It also does not take into account the cross-talk that might already present in the IHMs. However, the effects from these factors should be smaller than those due to overlapping speech, and would act only to exacerbate effects we report on here. Our study uses the performance *difference* between results with and without cross-talk in the same region of speech, and at this level of relative comparison, small effects such as reverberation would be roughly normalized out. To assess generalizability of our results, we repeat cross-talk experiments with mixing powers $1/4$, $1/2$, and 1, corresponding to mild to severe cross-talk.

Condition	*Mixing Power*	*WER*	*Sub*	*Del*	*Ins*
Clean	N/A	25.6	12.8	10.8	2.0
Background	1/4	29.1	13.6	13.7	1.9
Cross-talk	1/4	36.4	16.4	16.2	3.8
Background	1/2	30.6	13.9	14.8	1.8
Cross-talk	1/2	38.8	17.5	17.1	4.2
Background	1	32.6	14.2	16.6	1.7
Cross-talk	1	41.7	18.6	18.7	4.4

Table 1. WERs, and substitution (Sub), deletion (Del), and insertion (Ins) rates for different recognition conditions. The condition Clean refers to the case when the original IHM audio is used, and Cross-talk and Background are the cross-talk 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 being interfered.

3. RESULTS

3.1. Results During Overlap

3.1.1. WER

WERs and their the breakdown into substitution, deletion, and insertion rates for different recognition conditions are given in Table 1. WERs reported in this table are cumulative for all segments of the test data; analyses for overlaps and nonoverlaps are provided later. We observe that both the cross-talk and background noise significantly degrade recognition performance, the degradation being more severe in the cross-talk condition. As expected, the background noise does not introduce any additional insertions over the clean condition, and most additional errors in the cross-talk condition are insertions and deletions (which tend to associate with each other). Such a dramatic increase in insertions and deletions for the cross-talk condition is in agreement with the results in [8] for real-world cross-talk, and provides a sanity check for the design.

To perform an analysis of errors with respect to overlaps, we need a way to associate the recognition output with overlap/nonoverlap regions. Our recognition system outputs start and end times for each recognized word. Thus, correctly recognized words as well as insertions and substitutions are easily associated with the overlap/nonoverlap regions using the time information (when a word falls within more than one region, we assign it to the one with which it intersects most). Deletions are absent from the recognition output; we use the time marks in a forced alignment of reference transcripts to associate them with overlap/nonoverlap regions. Overlap/nonoverlap time boundaries are determined entirely from a forced alignment of reference transcripts, and are thus largely independent of the recognition condition (cf. Figure 1).

Using this method, we found the errors in the nonoverlap regions, and in the single- and two-speaker overlap regions. The WERs for each region type are calculated from the number of substitutions, insertions, deletions, and reference words assigned to the regions of that type. To facilitate the analysis, we display the WERs first with respect to the each recognition condition across overlap/nonoverlap types in Figure 2, and then with respect to the

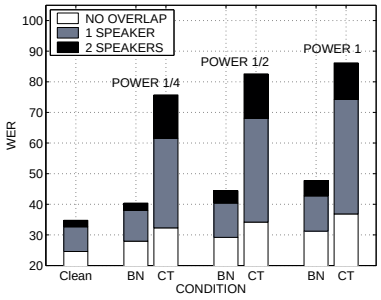

Fig. 2. WERs for the Clean, Background Noise (BN), and Cross-Talk (CT) conditions with the mixing powers $1/4$, $1/2$, and 1. For each condition, we display the WER in a stacked fashion for nonoverlaps, and single-speaker and two-speaker overlaps.

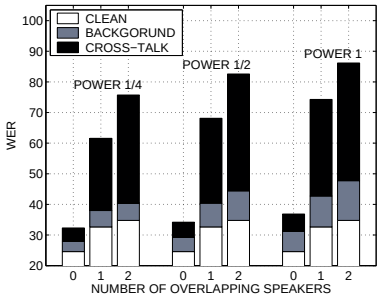

Fig. 3. WERs for nonoverlaps (0) and single-speaker (1) and two-speaker (2) overlaps. For each nonoverlap/overlap type and mixing power, the WERs are displayed in a stacked fashion for the clean, background-noise, and cross-talk conditions, in order.

overlap/nonoverlap types across different recognition conditions in Figure 3. We find that cross-talk significantly increases WER (much more so than background noise), and that two-speaker overlaps cause more errors than single-speaker overlaps. The observations from these plots are expected, but provide a quantification of errors due to the different conditions.

3.1.2. Perplexity

Perplexities for the nonoverlap and single- and two-speaker overlap regions are displayed in Figure 4. 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 4, there is a curious reversal of the relationship between perplexity and the number of simultaneous speakers. Typically, perplexity of higher-order n-grams should follow the same pattern as that for lower orders. In overlap regions, however, something different occurs. We note that while perplexities here are aggregated over the different sites at which meetings were collected, individual sites show a similar overall pattern, suggesting robustness of the results.

From inspection of individual n-grams, we believe the behavior can be explained as follows. We looked first at unigrams, and hand-coded each case as either a backchannel (e.g., "uhhuh","yeah"), discourse marker (e.g., "well"), filled pause ("um", "uh"), or none of the above. We found that overlaps contained far more backchannels and discourse markers than nonoverlaps, and the degree of increase

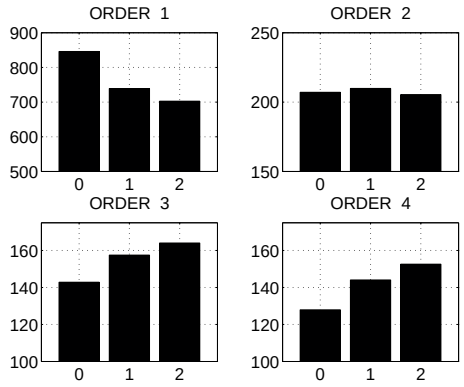

Fig. 4. Perplexities of the foreground reference words during nonoverlaps (0), and single-speaker overlaps (1), and two-speaker overlaps (2), for various n-gram LMs.

for both types of events was larger when the number of simultaneous speakers was higher. Relative rates of filled pauses, on the other hand, were stable or slightly decreasing with the number of simultaneous talkers. These findings make sense; most of these overlaps are associated with backchanneling rather than holding the floor. Because backchannels are frequent unigrams in LMs trained on spontaneous speech, unigram perplexity is *lower* when the number of overlapping talkers is *higher*.

What is very interesting is what happens for longer n-grams. We illustrate using 4-grams. In nonoverlap regions, 4-grams tend to be within-sentence sequences, such as "might be able to" and "just a matter of". If we look at overlap regions, however, we see far more cases like those below:

(a) "right right right so" (has repeated backchannel)
(b) "with yeah yeah an" (has backchannel inside phrase)
(c) "right i i am" (has sentence-initial disfluency)

In (a) and (c), the speaker produces multiple dialog acts; in (a) and (b) he produces multiple backchannels; in (b) he inserts a backchannel *within* a syntactic unit; in (c) he makes sentence-initial disfluencies. All of these behaviors are frequent at turn exchanges, where speakers reinforce each other, negotiate for the floor, and produce turn-initial discourse elements until the floor is determined [5]. In ASR language models however, such 4-grams are relatively rare, since most n-gram tokens come from regions inside single-speaker turns in which the speaker has already obtained the floor. Inside these single-talker stretches, there is little reason to backchannel. And while disfluencies can occur anywhere, their floor-grabbing function is used more at turn exchanges than within turns [6].

3.2. Results Surrounding Overlaps

Using the method described in the previous section, we associated errors in the recognition output with nonoverlap regions directly before and after an overlap (cf. Figure 1). We restricted the analysis to errors completely included within such nonoverlap regions, in order to avoid any bias from overlaps.

3.2.1. WER

In Figure 5, we plot WER over before- and after-overlap regions for different recognition conditions, as a function of the distance from

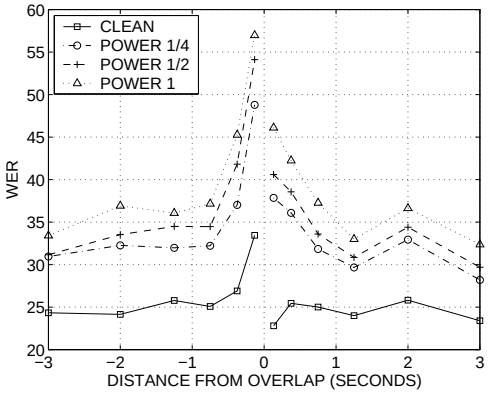

Fig. 5. WERs for Clean and Cross-Talk conditions with various gains, as a function of distance from the overlap (in seconds). Negative distances correspond to before overlaps, and positive distances to after overlaps. Note that the WER at each point represents data coming from that distance only (i.e., results are not cumulative).

the overlap. As expected, the WER decreases as a function of distance from the overlap, but there is an asymmetry in the errors before and after overlaps: WERs are significantly higher just before the overlap than it is after the overlap. This finding is consistent across different recognition conditions, and across meetings from different sources. This asymmetry is unlikely to be due to the recognizer itself, because the decoding is not strictly forward in time (i.e., it proceeds in a "forward-backward" fashion as opposed to Viterbi decoding, using word-confusion networks, n-best list scoring, speaker adaptation, and so on [8]). It is also unlikely to be due to reverberation because the meetings take place in relatively small rooms where reverberation effects are less than 250 msecs.

3.2.2. Perplexity

To further 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 6). A first observation from the figure is that rates during overlaps (middle bars, at 0) follow the same patterns as shown earlier, in Figure 4. They show low unigram perplexities, but high trigram and 4-gram perplexities. The new information in Figure 6 is the relative perplexities before and after overlap. For all n-gram orders, perplexity is lower after overlaps than before them. The same general pattern holds for each of the different meeting collection sites, so it appears to be a robust finding. Although further investigation is needed, we hypothesize from inspection that that the lower perplexity after overlaps stems at least in part from a tendency to begin new sentences at this location.

4. DISCUSSION

We have analyzed ASR errors in multiparty meetings with respect to regions before, during, and after speaker overlaps. 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 cross-talk gain conditions. Further analyses addressed questions that to our knowledge have not been studied on automatic recognition of meeting speech. We found that overlap tends to start at times during which the foreground talker is producing relatively

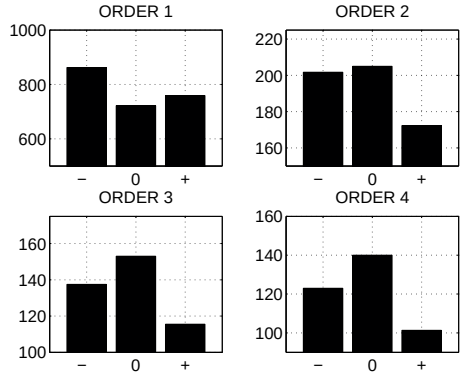

Fig. 6. Perplexities of the foreground reference words before (-), during (0), after overlaps (+) with respect to various n-gram LMs.

high perplexity word sequences. We also found that the relationship between perplexity and the number of simultaneous talkers is positive for longer n-grams, but negative for unigrams. This appears to be due to the preponderance of multiple short dialog acts within overlaps, particularly of backchannels. The short dialog acts are frequent unigrams, but their sequencing is not well represented in statistical LMs, since such events typically occur only at turn beginnings. We discovered an asymmetry in ASR errors made before versus after overlaps, which cannot be explained by properties of ASR nor of the experimental setup. The asymmetry occurred for each of the different sites represented in the test data, and appears to reflect differences in the speech itself. After being overlapped, the foreground talker temporarily drops to lower-perplexity word sequences. These results suggest that automatic modeling of meetings can benefit from a broader view of the relationship between overlap and ASR. For instance, one may want to use separate models adapted by proximity to overlaps. Future work should further investigate the relationship between overlap, ASR, and discourse phenomena.

Acknowledgments We thank Andreas Stolcke for useful discussions. This work is supported 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.

5. REFERENCES

[1] 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.

[2] A. Janin, *et al.*, "The ICSI Meeting Corpus," In *Proc. of ICASSP*, pages 364–367, 2003.

[3] N. Morgan *et al.*, "The Meeting Project at ICSI," In *Proc. of HLT*, pp. 1–7, 2001.

[4] T. Pfau *et al.*, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," *Proc. of ASRU Workshop*, pp. 107–110, 2001.

[5] H. Sacks *et al.*, "A Simplest Semantics for the Organization of the Turn-taking in Conversation," *Language*, vol. 50, pp. 696–735, 1974.

[6] E.A. Schegloff, "Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organisation," In *Talk and Social Organisation*, Clevedon, 1987.

[7] E. Shriberg *et al.*, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-party Conversation," In *Proc. of Eurospeech*, pp. 1359–1362, 2001.

[8] A. Stolcke *et al.*, "Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-Text Evaluation System," *Proc. of NIST RT-05 Meeting Recognition Workshop*, 2005.