A Corpus-Based Study of Interruptions in Spoken Dialogue

Agustín Gravano[1], *Julia Hirschberg*[2]

[1]Departamento de Computación & Laboratorio de Investigaciones Sensoriales,
Universidad de Buenos Aires, Buenos Aires, Argentina
[2]Department of Computer Science, Columbia University, New York, NY, USA
gravano@dc.uba.ar, julia@cs.columbia.edu

Abstract

We examine interruptions in a corpus of spontaneous task-oriented dialogue. We present evidence that interruptions occur at particular places in conversation. They are likely to occur during or after speech with certain acoustic/prosodic properties. We also examine the speech of interruptions themselves and find a number of significant differences between interrupting and non-interrupting turns.
Index Terms: interruption, turn-taking, prosody, dialogue.

1. Introduction

The speech science and technology communities have become increasingly interested in recent years in the study of turn-taking phenomena. Two applications benefit from this knowledge: interactive voice response (IVR) systems and automatic meeting processing systems. As IVR systems are increasingly able to support natural-seeming dialogues with users, a new challenge consists in handling more sophisticated system and user behaviors, including different kinds of interruptions [1, 2]. Likewise, as speaker segmentation and automatic speech recognition (ASR) technologies continue to improve, meeting processing systems must be able to deal a wider variety of turn exchanges that occur naturally in human-human dialogue [3].

Human interruptions and responses to them have been extensively studied by linguists and psychologists. Some such studies depict interruptions as violations of a principal rule of conversation, that only one party speaks at a time [4, 5], and are seen as indicative of power, control or dominance [6, 7] or as rude displays of indifference, aggressiveness or hostility towards the speaker [6]. Goldberg [8] and others, on the other hand, claim that interruptions may indeed be competitive, but they may also be neutral (e.g., requests for clarification) or used even to convey rapport with the interlocutor; these are often termed collaborative interruptions, in which a speaker helps their interlocutor, e.g. by completing their utterance. Collaborative interruptions are described as indicators of coordination and alignment in dialogue [9], and their production presents prosodic and gestural differences from competitive interruptions [10]. Cross-cultural studies show differences both in the frequency of interruptions and in the sociocultural value attached to them [11, 12].

A number of studies examine the acoustic/prosodic characteristics of interruptions. According to Yang [13], competitive interruptions have high pitch and intensity levels, while collaborative interruptions have a relatively lower pitch level. From a series of machine learning experiments, Lee and Narayanan [14] report that intensity-based features from the current speaker, as well as gestural features from the interlocutor such as eyebrow movement and mouth opening, are good predictors of the occurrence of interruptions in face-to-face dialogue. Related to interruptions are *speech overlaps*, during which both speakers speak at the same time, and thus may briefly compete for the conversation floor. Schegloff [15] argues that speech overlaps are usually resolved within two syllables or less, by means of devices such as a higher intensity or pitch level and a faster or slower speaking rate. A subclass of overlaps are *initiative conflicts*, in which both speakers begin speaking at roughly the same time after a silence. Initiative conflicts have been studied by Yang and Heeman [16], who report that these normally take less than two syllables to resolve and tend to be resolved in favor of the speaker displaying the higher intensity level.

In this study we address two questions that have received less attention in the literature. *Q1*: Where are interruptions likely to occur? Are there points in the current speaker's speech at which the interlocutor is more likely to interrupt? *Q2*: What characterizes the speech of interruptions? Is the onset of interruptions different from that of other conversational turns? We investigate these questions by analyzing the acoustic, prosodic, lexical and syntactic characteristics of the speech immediately preceding and following interruption points and comparing it to non-interruptions. Section 2 describes our corpus and the different types of of interruptions used in this study. Sections 3 and 4 present the methodology and results of our analyses of speech immediately preceding and following interruption points, respectively. Finally, Section 5 discusses the results, and outlines future research directions.

2. Corpus

The data for our experiments is the Columbia Games Corpus, a collection of 12 spontaneous task-oriented dyadic conversations elicited from 13 native speakers of Standard American English (SAE).[1] In each session, two subjects were paid to play a series of computer games requiring verbal communication to achieve joint goals of identifying and moving images on the screen. Subjects were recorded in a soundproof booth divided by a curtain to ensure that all communication was verbal. The subjects' speech was not restricted in any way, and the games were not timed. The corpus contains 9 hours of dialogue, which were orthographically transcribed, with the transcription time-aligned to the source by hand. Roughly 6 hours were intonationally transcribed using the ToBI framework [18].

We automatically extracted a number of acoustic features from the corpus using the Praat toolkit [19], including pitch, intensity, jitter, shimmer, and noise-to-harmonics ratio (NHR).

[1]A detailed description of the corpus, annotation methodologies and inter-labeler agreement measures may be found in [17].

Pitch slopes were computed by fitting least-squares linear regression models to the F_0 track extracted from given portions of the signal. Features were normalized by speaker using z-scores: $z = (x - \mu)/\sigma$, where x is a raw measurement, and μ and σ are the mean and standard deviation for a speaker.

For our turn-taking studies, we define an INTER-PAUSAL UNIT (IPU) as a maximal sequence of words surrounded by silence longer than 50 ms.[2] A TURN then is defined as a maximal sequence of IPUs from one speaker, such that the gap between any two adjacent IPUs contains no speech from the interlocutor. Two trained annotators classified all turn transitions in the corpus using the labeling scheme described in [17], identifying, inter alia (see Figure 1):

SMOOTH SWITCH (**S**): Transition from speaker A to speaker B such that A manages to complete her utterance, and no overlapping speech occurs between the two turns.

OVERLAP (**O**): Same as **S** but with some overlapping speech.

PAUSE INTERRUPTION (**PI**): Transition from speaker A to speaker B such that A does **not** manage to complete her utterance and no overlapping speech occurs.

SIMPLE INTERRUPTION (**SI**): Same as **PI** but with some overlapping speech.

BUTTING-IN (**BI**): Failed attempt from speaker B to interrupt speaker A, who thus continues speaking.

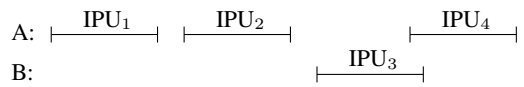

Figure 1: Hold transition (**H**) from IPU_1 to IPU_2; smooth switch (**S**) or pause interruption (**PI**) from IPU_2 to IPU_3; overlap (**O**) or simple interruption (**SI**) from IPU_3 to IPU_4.

Additionally, each **PI** and **SI** transition was annotated as a COLLABORATIVE COMPLETION (**CC**) when the speaker completes, or attempts to complete, an utterance from their interlocutor, as if trying to help them. Finally, all continuations from one IPU to the next IPU within the same turn were automatically labeled as HOLD (**H**) transitions. The Columbia Games corpus has 3250 **S** labels, 1067 **O**, 275 **PI** (38 of which are **CC**), 158 **SI** (6 **CC**), 104 **BI**, and 8123 **H**. In this study we consider only successful interruptions (either **PI** or **SI**), and thus exclude butting-ins (**BI**) from our analysis. In future research, we plan to contrast successful and unsuccessful interruption attempts and compare those results with [16].

3. Speech preceding interruptions

3.1. Pause interruptions

In previous research [17], we presented evidence of the existence of seven measurable events that take place with a significantly higher frequency in IPUs preceding smooth switches (**S**) than in IPUs preceding holds (**H**), summarized as follows: longer IPU duration; reduced lengthening of IPU-final words; lower intensity level; lower pitch level; higher values of three voice quality features: jitter, shimmer, and NHR; falling or high-rising intonation at the end of the IPU; and ending at a point of lexico-syntactic completion. These seven events represent potential TURN-YIELDING CUES, such that when several cues occur simultaneously, the likelihood of a subsequent turn-taking attempt by the interlocutor increases linearly; in the Games Corpus, the percentage of IPUs followed by a turn-taking attempt ranges from 5% when none of these turn-yielding cues are present to 65% when all seven cues are present, thus providing empirical support for Duncan's [4] general hypothesis. Recently, Hjalmarsson [20] has found additional empirical evidence that most of these cues also affect the listener's expectations of a speaker change.

To investigate our first question, *Q1*, regarding the existence of acoustic/prosodic cues preceding the occurrence of an interruption, we examine how IPUs immediately preceding pause interruptions (**PI**) differ from IPUs immediately preceding holds (**H**) or smooth switches (**S**). We hypothesize that, if a **PI** occurs during a random pause in the current speaker's turn, then its preceding IPU should be more similar to a **H** (and thus contain fewer turn-yielding cues) than to a **S** (containing more turn-yielding cues). Another possibility would be that speech preceding **PI** contains *some* turn-yielding cues but not all of them, thus lying between the **H** and **S** categories.

Figure 2 shows the speaker-normalized mean of the six acoustic/prosodic cues distinguishing **H** from **S**, now including **PI** as well. When we compare **PI** with the other two, one-way ANOVA tests reveal significant differences ($p < .05$) in a number of features.

First, we find that **IPU duration**, measured both in seconds and in number of words, is significantly *longer* in IPUs preceding **PI** than IPUs preceding **H**, and significantly *shorter* than IPUs preceding **S**. Likewise, the **speaking rate** of IPUs preceding **PI**, measured in phones per second, lies between the values of the two other groups, with both differences being significant. Turning our attention to acoustic features, we find that IPUs followed by **PI** have a **mean intensity** significantly *higher* than IPUs followed by **S**, but *not* significantly different from IPUs followed by **H**. Additionally, speech before **PI** appears to have a *lower* **mean pitch** than speech before **H** (approaching significance, $p \approx .05$), but *not* different from **S**. For our three **voice-quality** features (jitter, shimmer and NHR) we observe the same results as for intensity: IPUs preceding **PI** have a significantly *lower* mean value than IPUs preceding **S**, but do *not* differ significantly from IPUs preceding **H**.

For **final intonation**, we tabulate the phrase accent and boundary tone labels assigned to the end of each IPU, and compare their distribution for the **H**, **PI** and **S** turn exchange types, as shown in Table 1. A chi-square test indicates that there is a significant departure from a random distribution ($\chi^2 = 1196.2$, $d.f. = 10$, $p \approx 0$). An analysis of the residuals reveals that IPUs preceding **PI** are significantly more likely than the two other groups to a) end an intermediate phrase or b) end an intonational phrase with a plateau contour ([!]H-L%) or c) not end a prosodic phrase at all. All of these have been found in the literature to function as turn-holding cues [17, 20]. Furthermore, the proportion of IPUs followed by **PI** that end in either a falling contour (L-L%) or a high rise (H-H%) – two turn-yielding cues – is significantly lower than IPUs preceding the two other transition types. The absolute value of the speaker-normalized F_0 slope computed over the final 300 ms of each IPU works as an objective acoustic approximation of this perceptual feature: a plateau corresponds to a value of F_0 slope close to zero; rising or falling pitch corresponds to a high absolute value of F_0 slope. As shown in the rightmost chart of Figure 2, we find that the final slope before **PI** is significantly lower than before **S**, but not different than the slope before **H**, which supports the findings from our categorical prosodic labels.

Finally, to analyze the **lexico-syntactic completion** of IPUs preceding **PI**, we use the best-performing machine learning classifier developed in [17] to automatically classify all IPUs

[2] 50 ms was identified empirically to avoid stopgaps.

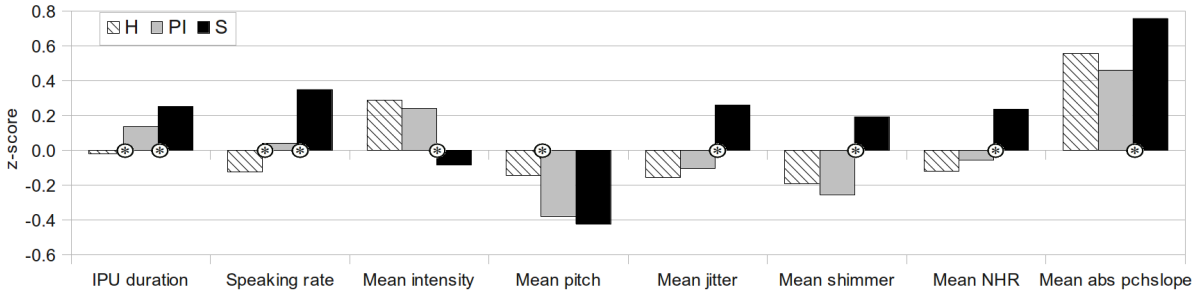

Figure 2: Speaker-normalized values of several variables for IPUs preceding holds (**H**), pause interruptions (**PI**) or smooth switches (**S**). Significant differences between the **PI** group and the two other groups are marked with '*' at the base of the corresponding bars. Intensity, pitch and voice quality means were computed over the IPU-final 1000 ms.

	H		PI		S	
H-H%	513	(9.1%)	8	**(3.5%)**	484	(22.1%)
[!]H-L%	1680	(29.9%)	90	**(39.1%)**	289	(13.2%)
L-H%	646	(11.5%)	16	(7.0%)	309	(14.1%)
L-L%	1387	(24.7%)	31	**(13.5%)**	1032	(47.2%)
No b.tone	1261	(22.4%)	78	**(33.9%)**	16	(0.7%)
Other	136	(2.4%)	7	(3.0%)	56	(2.6%)
Total	5623	(100%)	230	(100%)	2186	(100%)

Table 1: ToBI phrase accent and boundary tone for IPUs preceding **H**, **PI** and **S**.

in the corpus as either *complete* or *incomplete*. This classifier is trained on data annotated by human experts, who were asked to "determine whether what speaker B has said up to this point could constitute a complete response to what speaker A has said in the previous turn/segment." Our classifier uses a number of lexical and syntactic features extracted from the beginning of the turn up to the target IPU, without access to later material, and reaches an accuracy of 80.0% (majority class baseline: 55.2%; human labelers mean agreement: 90.8%). Of the 274 IPUs preceding a pause interruption, 207 (75.6%) are labeled *incomplete*. Also, 81.6% of all IPUs preceding a smooth switch (2649/3246) and 52.6% of all IPUs preceding a hold (4272/8123) are labeled *complete*. These differences are significant ($\chi^2 = 961.68, df = 2, p \approx 0$). The high proportion of incomplete IPUs preceding **PI** is to be expected, given the fact that the labelers of turn exchanges judged (with access to both transcripts and speech audio) that the current speaker had not managed to complete their utterance. In fact, these results suggest that most IPUs preceding **PI** are incomplete – independently of their acoustic/prosodic characteristics.

Summing up, these findings indicate that the IPUs that immediately precede a **PI** are more similar to turn-medial IPUs (i.e., those followed by a **H**), than to turn-final IPUs (i.e., those followed by a **S**). However, IPUs preceding **PI** and **H** are not identical: on average, the former have a higher duration, a lower mean pitch, a faster speaking rate and a much higher likelihood of being lexico-syntactically incomplete.

3.2. Simple interruptions

Next, we analyze the speech immediately preceding simple interruptions (**SI**), which differ from pause interruptions (**PI**) in that the former involve some overlap between the two conversational turns. In this section we compare the speech immediately preceding simple interruptions (**SI**), holds (**H**) and overlaps (**O**). Unlike **PI**s, it is not easy to determine where differences should be looked for before **SI**, since the current speaker is interrupted during the production of an IPU rather than during a pause separating two contiguous IPUs.

We first repeat the procedure described above for comparing IPUs preceding **H**, **SI** and **O**. We obtain results very similar to the ones shown in Figure 2 for **PI**. The order relations between the groups are preserved for all variables; however, due to the lower counts of **SI**, some of the statistical significance disappears, and thus we omit these results here.

In [17] we describe a procedure for contrasting speech before overlaps and holds over smaller portions of overlapped IPUs, which consists in comparing the penultimate ToBI intermediate phrases (*ip*s) preceding each transition type. Now we use this procedure for comparing the penultimate *ip*s preceding **SI**, **H** and **O**. Here, we do not observe significant differences between the *ip*s followed by **SI** and those followed by **H**; only a subset of the differences between the former and the *ip*s preceding **O** approach significance ($p < .1$): *ip* duration, speaking rate, shimmer and jitter.

These findings suggest that the speech immediately preceding simple interruptions (**SI**) is more similar to turn-medial speech (**H**) than to turn-final speech (**S**). However, speech preceding **SI** and **H** are not identical: we observe similar differences as in the **PI** vs. **H** comparison in the previous section.

4. Speech of interruptions

Considering our second research question (*Q2*), we now examine the speech of interruptions themselves. We constrast the initial IPUs of interrupting turns (**SI** and **PI**) with those of non-interrupting turns (**S** and **O**, respectively). Figure 3 summarizes the significant results (ANOVA, $p < .05$) for the duration of turn-initial IPUs and for speaking rate, mean intensity and mean pitch, each computed over the first 1000ms of the IPU. No significant differences were found for our voice quality features (jitter, shimmer and NHR).

Interrupting IPUs tend to start with a **higher intensity** than non-interrupting IPUs, a result consistent with previous work that singles out intensity as the most salient property of fights for initiative in dialogue [14, 16]. We also observe that interrupting IPUs usually have a **faster speaking rate**, thus providing empirical evidence supporting Schegloff's claim about the role of rate as a device for resolving speakership conflicts [15]. Also, for **SI**s, the initial IPU is on average longer than for **O**s; the difference is not significant for **PI**s.

PIs begin with a **lower mean pitch** than **S**s, but **SI**s start with a ***higher*** **mean pitch** than **O**s; this final result only approaches significance at $p = .13$. Following Yang [13], these differences for pitch may be explained by the collaborative vs.

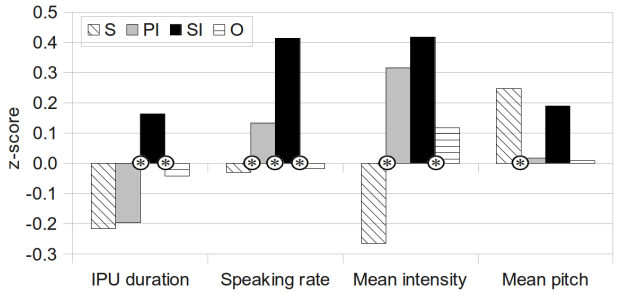

Figure 3: Speaker-normalized values of several variables for IPUs following smooth switches (**S**), pause interruptions (**PI**), simple interruptions (**SI**) or overlaps (**O**). Significant differences between each pair of adjacent groups are marked with '*' at the base of the corresponding bars.

competitive nature of interruptions: **PI**s may be collaborative more frequently than **SI**s; in our corpus, collaborative completions (**CC**) are more likely to occur as **PI**s than as **SI**s (recall the distribution of labels from Section 2).

Additionally, we compare the two interruption types, and find that **SI**s tend to start with longer IPUs and a faster speaking rate than **PI**s. In other words, speakers use fewer words, at a slower rate, when the interruption takes place during a pause from the current speaker.

4.1. Timing of interruptions

By definition, the onset of **PI**s occurs during a pause from the current speaker. In our corpus, the mean latency for **PI**s is 373 ms (sd=411). We examine how latency correlates with the variables described above, computed over the beginning of **PI** turns; we find that **latency** correlates slightly but significantly with the **speaking rate** of the first word ($\rho = .155$, $p = .011$) and of the entire IPU ($\rho = .136$, $p = .027$), in both cases measured as the number of phones per second. A plausible explanation for this finding is that, as the pause continues, a continuation from the current speaker becomes more likely. This leads the interlocutor to hurry their contribution to avoid an initiative conflict [16].

The onset of **SI**s always overlaps the current speaker's turn. In our corpus, the mean overlap duration of **SI**s is 303 ms (sd=253) and the mean syllable duration is 216 ms (sd=143). This supports Schegloff's claim that speech overlaps in conversation are usually resolved within two syllables or less [15]. We have run correlation tests between overlap duration and our variables, computed over the beginning of **SI** turns, and have found that **maximum intensity** correlates with **overlap duration** ($\rho = .147$, $p = .06$). As noted in [16], intensity is an important factor for resolving initiative conflicts. Likewise, our results suggest that during prolongated overlaps, interrupters tend to raise their voices to increase the chances of success of their interruptions. We also find that overlap duration correlates with the duration of the interrupting IPU ($\rho = .247$, $p = .001$).

5. Conclusions and future work

For our first research question, *Q1*, results from our corpus suggest that, in task-oriented dialogue, interruptions usually take place during or after IPUs that are more similar to turn-medial IPUs than turn-final ones, but that are still not identical to the former. In other words, we present evidence that interruptions apparently do *not* occur at random, but rather, they are more likely to occur during or after certain types of IPUs. For *Q2*, our comparison of the onset of interruptions and other turn transitions yields a number of significant differences in intensity and pitch level, speaking rate and IPU duration. These results might be useful to speech scientists and developers for identifying and processing interruptions in spontaneous conversation. In future work we plan to study the role of filled pauses (e.g., *um*, *uh*), affirmative cue words (*yeah*, *okay*) and other lexical classes both in and before interruptions. We will also use the annotations of butting-ins (**BI**) and collaborative completions (**CC**) in our corpus to characterize these special interruption types.

6. Acknowledgements

This work was funded in part by NSF IIS-0307905, NSF IIS-0803148, UBACYT 20020090300087, ANPCYT PICT-2009-0026, and CONICET.

7. References

[1] A. Raux, D. Bohus, B. Langner, A. W. Black, and M. Eskenazi, "Doing research on a deployed spoken dialogue system: One year of Let's Go! experience," in *Proc. of Interspeech*, 2006.

[2] N. Crook, C. Smith, M. Cavazza, S. Pulman, R. Moore, and J. Boye, "Handling user interruptions in an embodied conversational agent," in *Proc. of AAMAS*, 2010.

[3] Z. Yu and Y. Nakamura, "Smart meeting systems: A survey of state-of-the-art and open issues," *ACM Comp Surveys*,42(2),2010.

[4] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, 23(2), pp. 283–292, 1972.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, 50(4), pp. 696–735, 1974.

[6] C. West, "Against our will: Male interruptions of females in cross-sex conversation," *Annals of the New York Academy of Sciences*, 327(1), pp. 81–96, 1979.

[7] J. Orcutt and L. Harvey, "Deviance, rule-breaking and male dominance in conversation,"*Symbolic Interaction*,8(1), pp.15-32, 1985.

[8] J. Goldberg, "Interrupting the discourse on interruptions:: An analysis in terms of relationally neutral, power-and rapport-oriented acts," *Journal of Pragmatics*, 14(6), pp. 883–903, 1990.

[9] M. Poesio and H. Rieser, "Completions, coordination, and alignment in dialogue," *Dialogues & Discourse*, 1(1), pp. 1–89, 2010.

[10] C. Lee, S. Lee, and S. Narayanan, "An analysis of multimodal cues of interruption in dyadic spoken interactions," in *Proc. of Interspeech*, 2008.

[11] K. Murata, "Intrusive or co-operative? a cross-cultural study of interruption," *J. of Pragmatics*, 21(4), pp. 385–400, 1994.

[12] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speech overlaps in conversation," in *ICPhS*, 2007.

[13] L. Yang, "Visualizing spoken discourse: Prosodic form and discourse functions of interruptions," in *Proc. of SIGdial*, 2001.

[14] C. Lee and S. Narayanan, "Predicting interruptions in dyadic spoken interactions," in *Proc. of ICASSP*, 2010.

[15] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Lang. in Society*, 29(1), pp.1-63, 2000.

[16] F. Yang and P. A. Heeman, "Avoiding and resolving initiative conflicts in dialogue," in *Proc. of HLT/NAACL*, 2007.

[17] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Comp. Speech and Language*, 25(3), pp.601-634, 2011.

[18] M. E. Beckman and J. Hirschberg, "The ToBI annotation conventions," *Ohio State University*, 1994.

[19] P. Boersma and D. Weenink, "Praat," *http://www.praat.org*, 2001.

[20] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Commun.*, 53(1), pp. 23–25, 2011.