Entrainment and Turn-Taking in Human-Human Dialogue

Rivka Levitan [1,2], **Štefan Beňuš** [3], **Agustín Gravano** [4,5], **Julia Hirschberg** [2]
[1] Department of Computer and Information Science, Brooklyn College CUNY, USA
[2] Department of Computer Science, Columbia University, USA
[3] Constantine the Philosopher University in Nitra & Institute of Informatics, Slovak Academy of Sciences, Slovakia
[4] National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina
[5] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
levitan@sci.brooklyn.cuny.edu, sbenus@ukf.sk, gravano@dc.uba.ar, julia@cs.columbia.edu

Abstract

Interlocutors in spoken conversations have been shown to entrain, or become similar to each other, in multiple dimensions. We investigate the relationship between entrainment and turn-taking. We show that speakers entrain on turn-taking behaviors such as the distribution of turn types and degree of latency between turns, and that entrainment at turn exchanges is related to some extent to the type of turn exchange.

1 Introduction

In spoken and written conversation, conversational partners tend to align features of their speech with those of their interlocutor. This tendency, known as *entrainment*, *convergence*, or *alignment*, has been shown to exist for a speaker's choice of referring expressions (Brennan and Clark 1996), syntax (Branigan, Pickering, and Cleland 2000; Reitter, Moore, and Keller 2010), linguistic style (Niederhoffer and Pennebaker 2002; Danescu-Niculescu-Mizil, Gamon, and Dumais 2011), pronunciation (Pardo 2006), and acoustic-prosodic features such as intensity, pitch, and voice quality (Natale 1975; Gregory, Webster, and Huang 1993; Ward and Litman 2007; Levitan and Hirschberg 2011; Levitan 2014), in human-computer as well as human-human conversation (Bell et al. 2000; Bell, Gustafson, and Heldner 2003; Coulston, Oviatt, and Darves 2002; Brennan 1991; Brennan and Clark 1996; Thomason, Nguyen, and Litman 2013).

This study investigates two aspects of entrainment as it relates to turn-taking. The first part of this study continues from our work in (Levitan, Gravano, and Hirschberg 2011), which showed that interlocutors entrain on the use and expression of backchannel-inviting cues, by exploring entrainment on two other aspects of turn-taking behavior. We compare the similarity of a speaker's behavior to that of her conversational partner with its similarity to that of those speakers with whom she is never paired, to test the hypotheses that speakers entrain on the kinds of turns that they are likely to use and on the latency between their turns.

Next, we look at how turn-taking affects acoustic-prosodic entrainment. In (Levitan and Hirschberg 2011; Levitan 2014), we showed that speakers entrain *locally* on an array of acoustic-prosodic features — that is, the beginning of each turn matches, to some extent, the features of the ending of the interlocutor's previous turn. We compare the degree of local entrainment across multiple turn types and eight acoustic-prosodic features for a broad view of how local entrainment is mediated by turn type. We control for latency in order to isolate the effect of the specifically pragmatic characteristics of a given turn type on local entrainment.

One hypothesis examined here is that local acoustic-prosodic entrainment at turn exchanges systematically relates to the discourse structure of spoken task-oriented interactions. Previous research on prosodic signals to discourse boundaries showed that pause duration and resets in pitch and intensity correlate well with the presence and strength of a discourse boundary (e.g. (Swerts, Geluykens, and Terken 1992)). Most of these findings, however, come from data in monologues and narratives (e.g. (Hirschberg and Nakatani 1996; Oliveira Jr 2003; Swerts 1998)). Hence, informally, self-entraining across speech chunks with short latencies and pitch/intensity (declination) trends hints at discourse coherence while disentraining (long latencies and resets) hints at discourse boundaries. We hypothesize that, like self-entrainment in monologues, entrainment to an interlocutor at turn exchanges might be lower when a turn begins a new discourse segment than when it continues with the current one.

The results of this study provide direction for the design of an interactive voice response system that can entrain to a human user's prosody and turn-taking behavior in a human-like manner. Additionally, they suggest that a system can promote desirable turn-taking behavior in its human user by modeling such behavior itself, presenting additional motivation for reducing system response time.

2 Columbia Games Corpus

The experiments in this study were conducted on the Columbia Games Corpus (Gravano 2009), a corpus of twelve spontaneous, task-oriented dyadic conversations elicited from native speakers of Standard American English. During the collection of the corpus, each pair of participants played a set of computer games that required them to verbally cooperate to achieve a mutual goal. Neither sub-

ject could see the other's laptop screen. In the Cards games, one speaker (the information *giver* described the cards she saw on her screen, and her partner (the *follower*) attempted to match them to the cards on his own screen. In the Objects games, one speaker (the *giver* described the location of an object on her screen, and her partner (the *follower*) attempted to place the corresponding object in exactly the same location on his own screen. For both games, the participants received points based on how exact a match was; they later were paid for each point. Each of the twelve sessions consists of two Cards games and one Objects game.

Thirteen subjects participated in the collection of the corpus. Eleven returned on another day for another session with a different partner. Their ages ranged from 20 to 50 years (M = 30.0, SD = 10.9). They were recruited through flyers on the Columbia University campus, word of mouth, and the classified advertisements website www.craigslist.org. All reported being native speakers of Standard American English and having no hearing or speech impairments. Six subjects were female, and seven were male; of the twelve dialogues in the corpus, three are between female-female pairs, three are between male-male pairs, and six are between mixed-gender pairs. All interlocutors were strangers to each other.

Recording took place in a double-walled soundproof booth using close-talk microphones. The corpus consists of approximately nine hours of recorded dialogue; on average, each session is approximately 46 minutes long. It has been orthographically transcribed and annotated with prosodic and turn-taking labels.

2.1 Turn type annotation

All turns in the Games Corpus were labeled according to the turn-taking behavior exhibited at their initiation, following a scheme proposed in (Gravano 2009). An *inter-pausal unit*, or IPU, is defined as a pause-free unit of speech from a single speaker. A **turn** is defined as a maximal sequence of contiguous IPUs from a single speaker, as opposed to the usual discourse sense of the term; backchannels, in which the speaker does not attempt to take the floor, are considered turns according to this definition.

Utterances in which the speaker did not intend to take the floor are termed **backchannels** (**BC**, or **BC_O** if simultaneous speech is present). The remaining labels differentiate between the different kinds of turn exchanges in which the speaker *does* intend to take the floor. When simultaneous speech was present and the speaker was successful in taking the floor, the turn is called an **overlap** (**O**) if the previous utterance was complete (as judged by the annotator) and an **interruption** (**I**) if it was not. If simultaneous speech was present and the speaker's attempt to take the floor was *not* successful, the turn is termed a **butting-in** (**BI**). If simultaneous speech was *not* present, the turn is called a **smooth switch** (**S**) if the previous utterance was complete and a **pause interruption** (**PI**) if it was not. Turns beginning a new game task were labeled **X1**, and turns that continued a previous utterance that had been interrupted by a backchannel were labeled **X2**. Finally, all IPUs that were continuations of the same turn were labeled **holds** (**H**).

The Objects Games were labeled separately by two trained annotators, with a Cohen's κ of 0.99. Since inter-labeler agreement was so high, the Cards Games were labeled by only one of the original trained annotators. A complete description of the annotation of the Games Corpus can be found in (Gravano 2009).

Although we do not have independent labeling of discourse structure in this corpus, the turn-taking annotation provides partial and indirect access to this structure. It is plausible that turn-types can be ordered based on their likelihood to begin a new discourse segment. **Backchannels** inevitably continue with the current discourse segment. Similarly, continuations after backchannels (**X2**) signal strong discourse coherence to the previouse speech and an almost certain absence of a boundary. Next, the negative latency of overlaps and interruptions also suggests low probability of a discourse juncture. **Pause interruptions** present medial likelihood for a discourse boundary. Finally, **smooth switches** are most likely of the turn types to follow a discourse boundary. These turns follow acknowledgments and agreements from the interlocutor or pose questions, all of which might plausibly constitute a discourse boundary.

3 Entrainment on Turn-Taking Behaviors

In (Levitan, Gravano, and Hirschberg 2011), we showed that interlocutors entrain on their use and expression of *backchannel-inviting cues*, and suggested that they were likely to entrain on other kinds of turn-taking behavior as well. Here, we explore entrainment on the distribution of turn types used by a speaker, and on the latency of his or her turns.

3.1 Entrainment on Turn Types

An important aspect of a speaker's conversational behavior is what kinds of turns he or she tends to use. Some speakers, for example, may be interrupters; others may backchannel frequently, or produce lengthy pauses before they speak. We hypothesize that some of this interpersonal variation may be explained by entrainment: that is, we predict that speakers' turn-taking behavior will be more similar in this regard to that of their interlocutor than to that of other speakers in the corpus.

We measured the similarity of two speakers' turn type distributions using the Kullback-Leibler (KL) divergence, an asymmetric measure of the difference between two probability distributions (Kullback and Leibler 1951). For each speaker s, we computed a *partner* similarity (the negated KL divergence between s's turn type distribution and that of her conversational partner) and a *non-partner* similarity (the mean of the negated KL divergence between s's turn type distribution and that of each of her non-partners). Non-partners are defined as speakers in the corpus with whom neither s nor her current partner is ever paired, and who are of the same gender as s's partner. The gender restriction normalizes for any gender-specific turn-taking behaviors.

The *non-partner* similarity serves as a baseline for the degree of similarity we might expect if neither speaker s nor her partner entrain. We argue that significant evidence of similarity stronger than this baseline is evidence of entrainment, or adaptation towards one's partner, rather than

circumstancial similarity, which might arise from the fact that the two conversational partners are speaking about the same things in the same environment. However, the conditions of the collection of the Columbia Games Corpus make the comparison to non-partners a valid baseline. No participant in the corpus had previously met his or her partner. All conversations were strictly task-oriented and constrained by the progress of the game, which was the same for every speaker pair. In addition, as stated earlier, all conversations were recorded in a soundproof booth. Since circumstances for all conversations were the same, it is reasonable to employ the comparison to non-partners as a baseline to control for circumstancial similarity.

We compared partner and non-partner KL divergences over turn type distributions with a paired t-test, which showed that speakers' turn type distributions are more similar to those of their partners than to those of their non-partners ($t(23) = -2.04, p = 0.05$). That is, conversational partners use similar proportions of interruptions, backchannels, smooth switches, and other turn types.

3.2 Entrainment on Latency

Turn latency is the difference between the start time of a turn and the end time of the previous turn. Sometimes, as in the case of overlaps and interruptions, turn latency can be negative, when a turn starts while the interlocutor is still speaking.

Latency is an important part of the flow of a conversation. Conversations that have long latencies can be perceived as awkward or badly coordinated. Conversations with frequent negative latencies may be badly coordinated as well, with one or more of the participants either deliberately interrupting or misreading the other's turn-taking cues. In previous work, we showed that mean latency was negatively correlated with entrainment on intensity, pitch, voice quality and speaking rate (Levitan and Hirschberg 2011; Levitan 2014).

Here, we look at whether conversational partners entrain on turn latency. We compare a speaker's mean session latency to that of her partner's and to those of her non-partners — those speakers in the corpus with whom neither she nor her partner is ever paired. The non-partner comparison, as in Section 3.1, serves as a baseline for the degree of similarity we can expect if the partner's behavior has no effect on the speaker. We compare the partner similarities and the averaged non-partner similarities with a paired t-test, which shows that on average, speakers are more similar in mean latency to their partners than to their non-partners ($t(23) = 4.04, p = 0.00051$), leading us to conclude that interlocutors do entrain on this aspect of turn-taking behavior.

As Figure 1 shows, different turn types tend to have different latencies. Some of these differences are by definition — interruptions (I) and overlaps (O), for example, must always have negative latency, while the other categories must always have positive latency. Other differences, however, emerge from the pragmatic differences between turn types. Smooth switches (S), for example, have significantly larger latencies than backchannels (BC), pause interruptions (PI), and continuations after backchannels (X2), while interruptions overlap with their preceding turns significantly more than overlaps do.

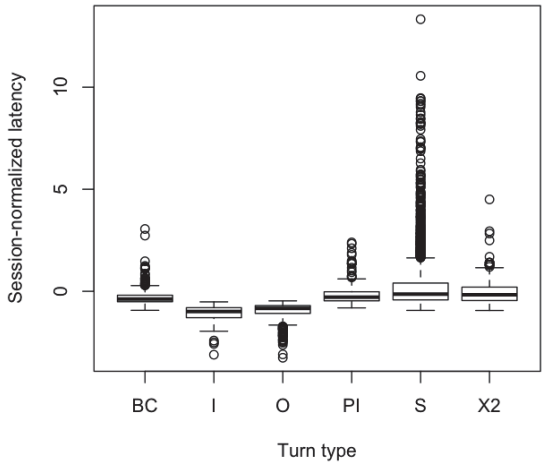

Figure 1: Session-normalized latency by turn type.

Since latency is affected by turn type, we must consider the possibility that our finding that speakers entrain on mean latency is redundant with our finding that speakers entrain on turn type distribution. We therefore calculate the z-score[1] of each raw latency value relative to its turn type's latency distribution and repeat the comparisons between partners and non-partners using the turn type-normalized scores. The result is essentially the same ($t(23) = 4.14, p = 0.00040$), showing that entrainment on mean latency holds independently of entrainment on turn types.

We calculate each speaker's average turn type-normalized latency for each turn type, and compare partner similarities in average turn type relative latency with the corresponding non-partner similarities in order to test the hypothesis that in addition to entraining on overall average latency, speakers entrain on the *relative* length of latencies used for each turn type. However, as Table 1 shows, we do not see any evidence that this is so: partner similarities in turn-specific relative latency are no greater than their corresponding non-partner similarities. It seems that while a speaker's average latency values are affected by those of her interlocutor, the relative latency of each turn type is not subject to entrainment.

Feature	*t*	*df*	*p*	*Sig.*[1]
S	-0.1	23	0.92	
I	0.47	23	0.64	
PI	1.24	23	0.23	
O	2.88	23	0.0085	.
BC	0.65	23	0.52	
X2	-0.26	21	0.8	

Table 1: Entrainment on mean turn type-normalized latency.

[1] $z = \frac{x-\mu}{\sigma}$

[1] All tests for significance correct for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The kth smallest p value is considered significant if it is less than $\frac{k \times \alpha}{n}$.

Locally, we find some evidence that the latency of a given turn is related to the latency of the previous turn from the opposite speaker. The session-speaker normalized latencies of alternating turns from opposite speakers are weakly correlated ($r = 0.037, df = 2404, p = 0.067$).

4 Entrainment Degree by Turn Type

In addition to entrainment on the kinds of turns a speaker uses, we are interested in how *local entrainment* — that is, how well the beginning of a speaker's turn matches the ending of her interlocutor's previous turn (Levitan and Hirschberg 2011) — is affected by the properties of the turn exchange. (Heldner, Edlund, and Hirschberg 2010) showed that **backchannels** (**BC**) are more similar in pitch to their preceding turns than other non-overlapping turns, (**smooth switches (S)** and **pause interruptions (PI)**) are in the Columbia Games Corpus. They suggested that this is due to backchannels' characteristic of being unobtrusive: the speaker does not intend to take the floor, and the backchannel is not necessarily even acknowledged by the interlocutor. Producing the backchannel at a pitch similar to that of the preceding turn helps "background" the backchannel by making it less conspicuous.

Similar reasoning may be applied to other kinds of turn exchanges. **Interruptions (I)**, which begin before the previous turn has ended, and when the interlocutor does not complete his or her utterance, may need to be more conspicuous than **overlaps (O)**, which also partly coincide with the previous turn, although that interlocutor's utterance is ultimately completed. Since interruptions can represent a more aggressive attempt to take the floor, they may require a greater degree of salience. On the other hand, interruptions may be viewed as collaborative completions (Local 2005), in which the speaker cooperatively completes her interlocutor's utterance. In that case, a speaker may begin where the interlocutor left off prosodically as well as syntactically and semantically, and locally entrain *more* than when the interlocutor's previous utterance is complete. An analogous argument may be made for pause interruptions, which are interruptions without simultaneous speech, as compared to smooth switches, for which the previous utterance is complete.

Alternatively, we hypothesize that local acoustic-prosodic entrainment is likely to be low at discourse boundaries, based on research on monologue data that found that pitch and intensity resets are associated with the presence and strength of a discourse boundary (Swerts, Geluykens, and Terken 1992). If this is true for inter-personal entrainment as well, we can expect turn types such as backchannels or X2, which cannot begin a discourse segment, or pause interruptions, which are unlikely to begin a discourse segment, to exhibit greater local similarity to their antecedent turns than smooth switches, which are the most likely type of turn to follow a discourse boundary. Interruptions and overlaps are both unlikely to follow a turn boundary, but overlaps, which partly coincide with an utterance that is ultimately completed, may be more likely to do so than interruptions, whose antecedent turn is left unfinished.

A difficulty that arises when attempting to isolate the effect of turn type on local entrainment is that different categories of turn exchanges tend to have different latencies (Figure 1). It is reasonable to hypothesize that a turn that begins more immediately after its preceding turn will be more similar to that turn, since the relevant representations of the interlocutor's prosody will be more strongly activated in the speaker's cognition. Thus it will be more likely for the speaker to prosodically continue where the interlocutor ended. An analysis of the relationship between local entrainment and turn latency in our data shows that the difference between adjacent IPUs is slightly correlated with turn latency (Table 2); the correlations are very small but lend weight to our hypothesis that this tendency exists. If a certain kind of turn tends to be more similar to its preceding turn, this may owe to the fact that turns in that category tend to have lower latencies, rather than any to pragmatic characteristics of the turn type. To correctly analyze the effect of turn type, it is therefore necessary to control for the effect of latency.

Feature	*r*	*df*	*p*	*Sig.*
Intensity mean	0.07	4892	3.1e-07	*
Intensity max	0.06	4892	4.4e-05	*
Pitch mean	0.05	4866	0.0016	*
Pitch max	0.04	4866	0.0047	*
Jitter	0.02	4855	0.094	
Shimmer	0.04	4800	0.0047	*
NHR	0.02	4888	0.15	
Speaking rate	0.02	4939	0.16	

Table 2: Normalized local differences correlated with normalized turn latency.

Normalizing for the difference between adjacent turns by the latency between them is an obvious way to control for latency. However, this may overly penalize turn categories that have short latencies, and inflate our estimate of the effect of entrainment in the case of turn categories with long latencies. Instead, we compare the local entrainment of each IPU with the local entrainment of a corresponding IPU with a different turn type but the same (or nearly the same) latency. Comparing IPUs with the same latency means that any differences we find can be attributed to turn type alone.

In our analyses, we define local entrainment as the negative absolute value of the difference between the features of the initial IPU of each reference turn and the features of the final IPU of the preceding turn (Levitan and Hirschberg 2011). Since our analyses include data from multiple sessions, we normalize each local entrainment measure and latency value by session means and standard deviations to produce z-scores.

We look at the relationship between turn type and entrainment on the following acoustic-prosodic features: mean and max intensity, which measures loudness; mean and max pitch; syllables per second, a measure of speaking rate; and jitter, shimmer, and noise-to-harmonics ratio (NHR), three measures of voice quality. Jitter describes the irregularity in the frequency of the vocal cord vibrations; shimmer describes the irregularity in the intensity of the vocal cord vibrations; and NHR is the ratio of the periodic portion of the

speech signal to the aperiodic or noise component of the signal. While the relationship between these measures and how speech is perceived is unclear, jitter and shimmer are generally associated with vocal harshness, while NHR is associated with hoarseness. All features were computed in Praat (Boersma and Weenink 2012); syllables were counted using an online dictionary.

4.1 Smooth Switches vs. Pause Interruptions

Smooth switches and pause interruptions both begin without simultaneous speech from the interlocutor. They differ in that smooth switches begin after the interlocutor has completed her utterance (as judged by the rater based on intonation, syntax, and meaning), while pause interruptions begin while the interlocutor's utterance is unfinished.

As stated above, we expect pause interruptions to be more similar to their preceding turns than smooth switches are, since they are less likely to begin a new discourse segment. Additionally, they can function as collaborative completions, beginning syntactically and semantically where the interlocutor left off, and may therefore be aligned prosodically as well. When we compare normalized local entrainment for pause interruptions and smooth switches, we do in fact find that entrainment is higher for pause interruptions for intensity max ($t(316.38) = 3.55, p = 4.4e-04$), pitch mean ($t(334.54) = 3.01, p = 2.8e-03$), and pitch max ($t(327.40) = 4.04, p = 6.7e-05$).

However, as can be seen in Figure 1, pause interruptions tend to have shorter latencies than smooth switches. This difference is significant according to a paired t-test ($t(739.16) = 10.72, p \approx 0$). On average, pause interruptions tend to start after 373 ms of silence, and smooth switches tend to start after 718 ms. The higher levels of entrainment for pause interruptions may be due to the fact that they begin more immediately after their preceding turns.

We therefore match every pause interruption with a smooth switch of the same latency (within $\epsilon = 0.0005$). 245 out of 274 pause interruptions have a corresponding smooth switch, with similar numbers of instances taken from all twelve sessions. We then compare their normalized local differences for each feature with a paired t-test. The differences in local entrainment on intensity max and pitch mean are no longer significant according to the corrected p-value (intensity max: $t(233) = 1.87, p = 0.063$; pitch mean: $t(235) = 2.43, p = 0.016$), although the differences do approach significance. The difference in local entrainment on pitch max is still significant ($t(235) = 3.11, p = 0.0021$). We can therefore conclude that speakers locally entrain *more* on pitch max when they begin their utterance after an incomplete utterance from their interlocutor, as opposed to when the interlocutor's utterance is complete. The same is true for intensity max and pitch mean, but while the effect on pitch max is present (though less significant) even when turn latency is controlled for, the effect of turn type on intensity max and pitch mean may be associated with the tendency of pause interruptions to have lower latencies than smooth switches.

4.2 Smooth switches vs. backchannels

Backchannels are short segments of speech uttered by a speaker to indicate that she is paying attention and to encourage the other speaker to continue, without attempting to take the floor. They typically fulfill interactional functions rather than conveying information, and usually go unacknowledged by the other speaker. In Figure 2, the word "okay" is a backchannel.

Speaker A: All right so I have a- a nail on top
Speaker B: ***okay***
Speaker A: with an owl in the lower left.

Figure 2: Example of a backchannel (BC).

(Heldner, Edlund, and Hirschberg 2010) showed that backchannels are more similar in pitch to their preceding turns than smooth switches are, and proposed that this entrainment is related to the "backgrounding" of backchannels. However, their analysis did not account for the fact that backchannels begin, on average, after approximately 444 ms less latency than smooth switches do ($t(3369.34) = 18.57, p \approx 0$) (Figure 1). We apply our method of comparing turns with the same latency to isolate the effect of turn type from the effect of latency, matching 511 of 553 backchannels with smooth switches of the same latency (within $\epsilon = 0.0005$); again, these instances are distributed fairly evenly among all twelve sessions.

Backchannels were more similar to their preceding turns than smooth switches with the same latency were for all features we examined except speaking rate (Table 3). This result supports the finding of (Heldner, Edlund, and Hirschberg 2010) with respect to pitch mean, using a different measure of pitch and controlling for the effect of latency, and extends the comparison of local entrainment to seven other features. We can conclude that for pitch, intensity, and voice quality — three major aspects of prosody — speakers entrain more for backchannels then they do for smooth switches, and that this behavior is apparent even when the effect of turn latency is controlled for.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	7.02	502	7.1e-12	*
Intensity max	3.7	502	0.00024	*
Pitch mean	3.34	502	0.00089	*
Pitch max	3.06	502	0.0023	*
Jitter	6.08	498	2.4e-09	*
Shimmer	6.42	490	3.2e-10	*
NHR	8.62	501	8.8e-17	*
Speaking rate	1.17	510	0.24	

Table 3: T-tests for differences in local entrainment between smooth switches and backchannels with the same latencies.

One way in which backchannels differ from smooth switches is that they tend to be shorter. All backchannels in our data consist of a single word in a single IPU. Smooth switches, on the other hand, can consist of multiple IPUs.

Backchannels may be more similar to their preceding turns than smooth switches are because in the case of backchannels, the speaker does not intend to continue speaking, and his or her pitch and intensity may therefore be lower. However, of the 511 backchannels matched with smooth switches of the same latency, 312 are matched with smooth switches that also have only one IPU, and a further 100 are matched with smooth switches that have two. Additionally, we obtain comparable results when restricting the comparison to smooth switches that have only one IPU, and when restricting the comparison to smooth switches that have more than two IPUs, suggesting that the differences in local entrainment between smooth switches and backchannels can be ascribed to other discourse functions such as the unobtrusiveness of backchannels, or the fact that they can never begin a new discourse segment, rather than the disparity in turn length.

4.3 Overlaps vs. Interruptions

Overlaps and interruptions are two kinds of turns that begin while the interlocutor is still speaking — that is, they have negative latency. When the previous turn is syntactically, semantically and prosodically complete (as judged by the rater), the overlapping turn is termed an overlap; when the previous turn is incomplete, the overlapping turn is called an interruption. An overlap is the counterpart of a smooth switch, and an interruption is the counterpart of a pause interruption. We might therefore expect speakers to locally entrain more when producing an interruption, as is the case with pause interruptions.

Our standard definition of local entrainment, the negated absolute difference between the first IPU of the reference turn and the last IPU of its preceding turn, cannot be applied to turns that overlap with their preceding turns, since the first IPU of the reference turn may begin before the last IPU of the previous turn has even been uttered, and therefore cannot be said to be aligned to it. Instead, we compare the first IPU of the reference turn to the last IPU *that does not overlap with it.* For example, in Figure 3, IPU 4 will be compared to IPU 1.

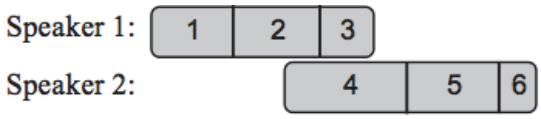

Figure 3: Overlapping turns with numbered IPUs.

This generalized definition can be applied to the calculation of local entrainment in non-overlapping turns as well. However, according to this definition, overlapping turns may have speech from the interlocutor between the two IPUs under comparison (as in Figure 3), which cannot occur in the case of non-overlapping turns. We therefore do not compare entrainment in overlapping turns with entrainment in non-overlapping turns.

We match each interruption with an overlap of the same latency. Latency here is the difference between the ending time of the first IPU under comparison and the start time of the second IPU under comparison, as opposed to the difference between the two turns, which is negative in the case of overlaps and interruptions. Due to sparsity, we use an ϵ of 0.01 (instead of 0.0005, as we do for other turn types), such that $|$(latency of the first IPU) $-$ (latency of the second IPU)$| < \epsilon$. 61 of 158 interruptions have matches. Examples here are fewer than they are for other turn types, but at least one data point is included from each session, with a reasonable spread over all sessions.

When latencies are *not* matched, there are no differences in local entrainment between interruptions and overlaps. The same is true when local entrainment is compared between interruptions and overlaps with the *same* latencies (Table 4), although the differences in local entrainment on pitch mean and jitter approach significance, with overlaps displaying greater local similarity. This accords well with our hypothesis that local entrainment is associated with discourse coherence, since while neither interruptions nor overlaps are likely to begin a new discourse segment, because of their negative latency, overlaps are more likely to do so, since the turn that they partly coincide with is ultimately completed. The difference between the results of the interruptions/overlaps comparison and the pause interruptions/smooth switches comparison may also possibly be attributed to the fact that according to our calculation of local entrainment, overlapping turns may have some speech from the interlocutor intervening between the two IPUs being compared, which cannot occur in the case of non-overlapping turns.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	0.72	58	0.47	
Intensity max	0.67	58	0.5	
Pitch mean	2.27	57	0.027	
Pitch max	1.67	57	0.1	
Jitter	2.32	55	0.024	
Shimmer	1.87	54	0.066	
NHR	1.93	58	0.059	
Speaking rate	-1.52	60	0.13	

Table 4: T-tests for differences in local entrainment between interruptions and overlaps with the same latencies.

4.4 Smooth Switches vs. Continuations After Backchannels

In the Columbia Games Corpus, continuations after backchannels are labeled X2. In Figure 2, Speaker A's second utterance (*"with an owl in the lower left"*) is an X2. On average, latency before X2 utterances is significantly smaller than the latency before smooth switches ($t(1565.02) = 9.15, p \approx 0$). When local entrainment of smooth switches is compared with local entrainment of X2 utterances, without accounting for this difference in latency, every feature except speaking rate is significantly more similar between the beginnings of X2 utterances and the endings of their previous turns than between the beginnings of smooth switches and the endings of their previous turns. Local entrainment

on speaking rate is in fact *greater* for smooth switches (Table 5).

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	7.65	697.75	6.8e-14	*
Intensity max	5.03	704.25	6.1e-07	*
Pitch mean	3.41	714.82	0.00068	*
Pitch max	3.63	650.74	0.00031	*
Jitter	5.09	694.83	4.6e-07	*
Shimmer	6.55	726.28	1.1e-10	*
NHR	8.29	720.14	5.6e-16	*
Speaking rate	-2.65	599.60	0.0083	*

Table 5: T-tests for differences in local entrainment between smooth switches and X2 when latencies are not matched.

Several of these differences disappear when latency is controlled for in the comparison between X2 and smooth switches. Table 6 shows the results of t-tests for differences in local entrainment between smooth switches and X2 utterances with the same latencies (381 of 449 X2 utterances have matches). Intensity mean, pitch max, jitter, shimmer, and NHR show significantly higher local entrainment for X2 as compared to smooth switches, but the differences in local entrainment on intensity max, pitch mean, and speaking rate are most likely related to the fact that X2 utterances tend to have lower latency than smooth switches. The greater entrainment of X2 utterances is consistent with our hypothesis that entrainment is associated with discourse coherence, since X2 utterances can almost never follow a discourse boundary.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	3.82	374	0.00015	*
Intensity max	1.68	374	0.093	
Pitch mean	1.72	373	0.085	
Pitch max	2.64	373	0.0087	*
Jitter	4.40	373	1.4e-05	*
Shimmer	4.46	370	1.1e-05	*
NHR	7.14	378	4.7e-12	*
Speaking rate	-0.07	380	0.95	

Table 6: T-tests for differences in local entrainment between smooth switches and X2 with the same latencies.

5 Discussion and Conclusions

We have explored how entrainment relates to turn-taking in two dimensions in human-human conversation. First, we have shown that interlocutors entrain to each other on two aspects of turn-taking behavior: partners' distributions of turn types are similar and their mean latency between turns is similar. These findings can be exploited in the design of spoken dialogue systems to facilitate turn-taking behaviors that are associated with good system performance. Long latencies, for example, are undesirable in human-computer interactions, since they cost unnecessary bandwidth and may even interfere with the performance of the ASR (automatic speech recognition). A system may be able to promote shorter latencies on the part of its human interlocutor by shortening its own response time. While system response time is highly constrained by the throughput of its components, the potential to shorten the human user's latency provides additional motivation for implementing strategies such as incremental processing to reduce system latency.

Similarly, a system may promote the use of certain turn types on the part of the user by producing those turn types itself. This is useful in the case of backchannels, a behavior that is desirable, if the system can process it, because it is a low-latency way of validating what the system is doing. Our results suggest that a system can encourage a user to backchannel by producing backchannels of its own.

In addition to how speakers entrain on turn-taking behaviors, we show that turn-taking behaviors affect local entrainment on acoustic-prosodic features. Specifically, we show that latency is negatively associated with local entrainment, possibly because the features of the interlocutor's previous turn are more strongly activated in the speaker's cognition when the reference turn closely follows the previous one. We further show that certain types of turns have higher local entrainment than others, even when the different latencies that are characteristic of different turn types are controlled for. Speakers entrain more closely on pitch max when the turn type is a pause interruption (speech does not overlap, but the previous utterance is incomplete) than when it is a smooth switch (the previous utterance is complete). They entrain more on all features except speaking rate when producing a backchannel — that is, when not intending to take the floor — than when producing a smooth switch, an effect that we show is independent of the effects of utterance length as well as turn latency. Overlaps and interruptions (turns overlapping with the previous utterance that differ in whether that utterance is complete) show no differences in local entrainment. Continuations after backchannels are more similar to their previous utterances than smooth switches are in intensity mean, pitch max, jitter, shimmer, and NHR.

(Heldner, Edlund, and Hirschberg 2010) suggested that backchannels are more similar to their preceding utterances because they are meant to be unobtrusive, and entraining to the preceding turn is a way of "backgrounding" an utterance; this reasoning is consistent with our results here, which show that backchannels match their preceding utterances in nearly every acoustic-prosodic feature examined here, independently of utterance length or latency. Backchannels' entrainment to their preceding utterances, in combination with the fact that backchannels do not disrupt the ongoing utterance may explain the entrainment of continuations after backchannels to their preceding backchannels: Speaker A produces an utterance. Speaker B produces a backchannel with the same prosody. Speaker A continues with the same prosody as before, which is the same as the prosody of B's backchannel.

A complementary explanation relates to the fact that smooth switches are most likely to begin a new discourse segment. Backchannels and continuations after backchannels can *never* begin a new discourse segment, and pause interruptions, overlaps and interruptions are unlikely to do so. Differences in local entrainment may be an indicator

of discourse structure, with turns at discourse boundaries entraining less. This would explain all of the differences in local entrainment by turn type reported here, and suggests the potential of using local entrainment — a low-level, automatically-derived measure — as a feature for automatically discovering discourse structure. Future work should explore this possibility and explicitly test the relationship between inter-personal entrainment and discourse structure with data whose discourse boundaries have been manually identified.

Acknowledgments

This material is based in part upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract # HR0011-12-C-0016 and by the Slovak Ministry of Education under VEGA 2/0197/15. Any opinions, findings and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of DARPA.

References

Bell, L.; Boye, J.; Gustafson, J.; and Wirn, M. 2000. Modality convergence in a multimodal dialogue system. In *Proceedings of Gtalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue*, 29–34.

Bell, L.; Gustafson, J.; and Heldner, M. 2003. Prosodic adaptation in human-computer interaction. In *Proceedings of ICPHS'03*, 833–836.

Boersma, P., and Weenink, D. 2012. Praat: doing phonetics by computer [computer program]. Version 5.3.23, retrieved 21 August 2012 from http://www.praat.org.

Branigan, H. P.; Pickering, M. J.; and Cleland, A. A. 2000. Syntactic co-ordination in dialogue. *Cognition* 75(2):B13–B25.

Brennan, S. E., and Clark, H. H. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 22(6):1482–1493.

Brennan, S. E. 1991. Conversation with and through computers. *User modeling and user-adapted interaction* 1(1):67–86.

Coulston, R.; Oviatt, S.; and Darves, C. 2002. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of ICSLP'02*.

Danescu-Niculescu-Mizil, C.; Gamon, M.; and Dumais, S. 2011. Mark my words! linguistic style accommodation in social media. In *Proceedings of WWW*.

Gravano, A. 2009. *Turn-taking and affirmative cue words in task-oriented dialogue*. Ph.D. Dissertation, Columbia University.

Gregory, S.; Webster, S.; and Huang, G. 1993. Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language & Communication* 13(3):195–217.

Heldner, M.; Edlund, J.; and Hirschberg, J. B. 2010. Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech*.

Hirschberg, J., and Nakatani, C. H. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 286–293. Association for Computational Linguistics.

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 79–86.

Levitan, R., and Hirschberg, J. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech*.

Levitan, R.; Gravano, A.; and Hirschberg, J. 2011. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Levitan, R. 2014. *Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue*. Ph.D. Dissertation, Columbia University.

Local, J. 2005. On the interactional and phonetic design of collaborative completions. *A figure of speech: A festschrift for John Laver* 263–282.

Natale, M. 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology* 32(5):790–804.

Niederhoffer, K. G., and Pennebaker, J. W. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21(4):337–360.

Oliveira Jr, M. 2003. Pitch reset as a cue for narrative segmentation. *Evaluation* 3(4):7–9.

Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustic Society of America* 19(4).

Reitter, D.; Moore, J. D.; and Keller, F. 2010. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation.

Swerts, M.; Geluykens, R.; and Terken, J. M. 1992. Prosodic correlates of discourse units in spontaneous speech. In *ICSLP*.

Swerts, M. 1998. Filled pauses as markers of discourse structure. *Journal of pragmatics* 30(4):485–496.

Thomason, J.; Nguyen, H. V.; and Litman, D. 2013. Prosodic entrainment and tutoring dialogue success. In *Artificial Intelligence in Education*, 750–753. Springer.

Ward, A., and Litman, D. 2007. Measuring convergence and priming in tutorial dialog. Technical report, University of Pittsburgh.