Entrainment in Spoken Dialogue Systems: Adopting, Predicting and Influencing User Behavior

Rivka Levitan
Department of Computer Science
Columbia University
New York, NY 10027, USA
rlevitan@cs.columbia.edu

Abstract

Entrainment is the phenomenon of the speech of conversational partners becoming more similar to each other. This thesis proposal presents a comprehensive look at entrainment in human conversations and how entrainment may be incorporated into the design of spoken dialogue systems in order to improve system performance and user satisfaction. We compare different kinds of entrainment in both classic and novel dimensions, provide experimental results on the utility of entrainment, and show that entrainment can be used to improve a system's ASR performance and turn-taking decisions.

1 Introduction

Entrainment is the phenomenon of interlocutors becoming more similar to each other in their speech in the course of a conversation. Entrainment has been observed in numerous domains and for multiple levels of communication. In addition, many studies have shown associations between entrainment and desirable dialogue characteristics. The proposed work aims to improve spoken dialogue system performance both qualitatively and quantitatively by exploiting this prevalent and significant phenomenon. Spoken dialogue systems can significantly improve the quality of their user interactions by incorporating entrainment into their design:

- A spoken dialogue system can entrain to its users, adjusting its own output to align with theirs. This should improve the dialogue quality as perceived by the user.
- It can exploit the concept of entrainment by changing the parameters of its own output when it wants the user to speak differently. For example, when the user is speaking too quickly, the system can slow down its own output, causing the user to do the same.
- It can use an entrainment model along with information about its own behavior to more accurately predict how the user will behave.

Our proposed work explores the role of entrainment in human conversations and looks at how it can improve interactions with spoken dialogue systems. In addition to presenting an in-depth study of the characteristics of human entrainment, we will demonstrate that spoken dialogue systems can use this information to predict characteristics of the user's speech, improve the user's impression of the dialogue quality and system persona by adopting the user's speech characteristics, and improve recognition accuracy by influencing the user to abandon prosodic characteristics associated with ASR error.

This thesis proposal is organized as follows: Section 2 discusses the literature related to the proposed work. Section 3 describes the corpus used in these studies. Section 4 addresses the question of how humans entrain and how this information can be used to more accurately predict a user's behavior. Section 5 discusses how entrainment affects the perceived quality of human and human-computer conversations, and Section 6 explores how entrainment can be used to influence user behavior. Section 7 describes the main contributions of this work.

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 84–90,
Atlanta, Georgia, 13 June 2013.

2 Related work

Entrainment has been shown to occur at almost every level of human communication: lexical (Brennan and Clark, 1992), syntactic (Reitter and Moore, 2007; Ward and Litman, 2007), stylistic (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011), acoustic-prosodic (Natale, 1975; Coulston et al., 2002; Ward and Litman, 2007) and phonetic (Pardo, 2006).

Entrainment in many of these dimensions has also been associated with measures of dialogue success. Chartrand and Bargh (1999), for example, demonstrated that subjects who interacted with confederates who mimicked their posture and behavior reported greater liking for the confederate and a smoother interaction. Lee et al. (2010) found that entrainment measures derived from pitch features were significantly higher in positive interactions between married couples in therapy than in negative interactions. Looking at more objective measures, Nenkova et al. (2008) found that the degree of entrainment on high-frequency words was correlated with task score and turn-taking features.

These studies have been motivated by theoretical models such as Giles' Communication Accommodation Theory (Giles et al., 1987), which proposes that speakers promote social approval or efficient communication by adapting to their interlocutors' communicative behavior. Another theory informing the association of entrainment and dialogue success is the coordination-rapport hypothesis (Tickle-Degnen and Rosenthal, 1990), which posits that the degree of liking between conversational partners should be correlated with the degree of nonverbal coordination between them. In contrast, Chartrand and Bargh (1999) posit that entrainment is a purely automatic process, a product of the perception-behavior link, which predicts that the act of observing a behavior makes the observer more likely to engage in that behavior as well.

3 Columbia Games Corpus

Many of the studies in this work were conducted on the Columbia Games Corpus (Gravano, 2009), a collection of twelve 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. In the Cards games, one speaker described the cards she saw on her screen, and her partner attempted to match them to the cards on his own screen. In the Objects games, one speaker described the location of an object on her screen, and her partner 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.

The corpus consists of approximately nine hours of recorded dialogue. It has been orthographically transcribed and annotated with prosodic and turn-taking labels. Thirteen subjects participated in the collection of the corpus, and nine returned on another day for a second session with a different partner. This is useful for our study of entrainment, since we can compare a single speaker's behavior with two different interlocutors. In addition, the corpus is representative of the kind of speech we are interested in: task-oriented dialogue between strangers.

4 Entrainment in human conversations

We begin our study of entrainment by looking at entrainment in human conversations. Aside from the interest inherent in advancing our understanding of this human behavior, research in this area can inform the design of spoken dialogue systems. A system that entrains the way a human does will seem more natural, and a system that knows how humans entrain can use this information to better predict how a user will behave, improving its own performance.

4.1 Acoustic-prosodic entrainment

This study, previously presented in (Levitan and Hirschberg, 2011), creates a cohesive view of entrainment by directly comparing entrainment on a set of acoustic-prosodic features, measured in five different ways. By comparing these different measures of entrainment, we bring clarity to three aspects of entrainment:

- *Is it global or local?* Two speakers may fluctuate around similar means, while diverging widely at any specific point. Conversely, they may be globally dissimilar, but locally they may be relatively similar.
- *Is it by value or by direction?* If a speaker en-

trains to her partner's actual value, if he lowers his voice, she may raise her own in order to match his new intensity. If she matches the direction of the change rather than the new value, she will lower her voice as well, even if this results in a value less similar to his.

- *Is the degree of entrainment static, or does it improve?* Do speakers converge—become more similar—as the conversation progresses?

The features we examine are intensity mean and max, pitch mean and max, jitter, shimmer, noise-to-harmonics ratio (NHR), and syllables per second[1]. We look for evidence of *global* entrainment by comparing the similarities in feature means between partners with the similarities between speakers who are not conversational partners.

We see an effect of entrainment for almost all the features. In addition, the difference between partners for several of the features is smaller in the second half of the conversation, constituting evidence of *convergence*. We also find a strong effect of *local* entrainment: for every feature, adjacent turns are significantly ($p < 0.001$) more similar to each other than non-adjacent turns. We conclude that entrainment is by value rather than by direction; that global entrainment exists in addition to local matching for several features, most notably intensity; and that entrainment is dynamic for some features, improving as the conversation progresses.

4.2 Entrainment on outliers

Since entrainment is generally considered an unconscious phenomenon, it is interesting to consider entrainment when a feature is particularly salient. The theory that the perception-behavior link is the mechanism behind entrainment (Chartrand and Bargh, 1999) would predict that the effect of entrainment would be stronger in this case, since such features are more likely to be observed and therefore imitated. We test this hypothesis by looking at cases in which one speaker in a pair has a feature value in the 90th or 10th percentile. This study was previously described in (Levitan et al., 2012).

[1]Intensity mean is an acoustic measure perceived as loudness, and intensity max represents the range of loudness. Jitter, shimmer and NHR are three measures of voice quality; jitter and shimmer are perceived as harshness, and NHR as hoarseness. Syllables per second measure speaking rate.

As in our tests for global entrainment (Section 4.1), we compute a partner and non-partner similarity for each speaker. The partner similarity should be lower for outlier pairs (pairs in which one speaker has an outlier feature value), and the non-partner similarity should be lower as well, since the outlier speaker diverges from the norm. We therefore can expect the difference between these two values to be the same for outlier and typical pairs. If this difference is lower for outlier pairs, we can conclude that the effect of entrainment is weaker in outlier cases. We find, in fact, that this difference is *greater* for outlier pairs for several features, indicating that speakers entrain *more* to outlier values of these features. This finding supports the perception-behavior link. In addition, it has implications for cases in which it is an objective to induce one's interlocutor to entrain, as we will discuss in Section 6.

4.3 Entrainment and backchannel-inviting cues

Backchannels are short, nondisruptive segments of speech that a speaker utters to let his interlocutor know that he is keeping up. They are extremely prevalent in task-oriented conversation. Gravano and Hirschberg (2009) identified six acoustic and prosodic features that tend to be different before backchannels, hypothesizing that these features serve as cues to one's interlocutor that a backchannel would be welcome. Individual speakers use different sets of cues, and can differ in their realization of a cue. We look for evidence of entrainment on backchannel-inviting cues. This work, previously discussed in (Levitan et al., 2011), represents a first look at entrainment in a pragmatic dimension.

We measure backchannel-inviting cues in three ways. Firstly, we measure the similarity of the speaker pairs' cue sets by counting the number of cues they have in common, and find that partners have more cues in common than non-partners. Secondly, we measure the similarity of cue realization, and show that feature values before backchannels for pitch, intensity and voice quality are more similar between partners. In addition, this measure shows evidence of convergence for pitch and intensity, which are more similar before backchannels in the second half of a conversation. Finally, we measure the local effect of this entrainment by correlating feature values before consecutive backchannels

and find that pitch and intensity before backchannels are moderately correlated.

4.4 Future work

We have shown that a speaker's conversational behavior is influenced by that of her interlocutor. We therefore propose to develop a framework for using entrainment information to label or predict a speaker's behavior. An example of such a task is predicting backchannels. Based on the work of Gravano and Hirschberg (2009), a system deciding whether to produce a backchannel or take the floor should compare the user's most recent utterance to a backchannel-preceding model and a turn-yielding model. Since each speaker uses a different count of backchannel-preceding cues, a model trained on other speakers may not be useful. However, data from the user may not be available and is likely to be sparse at best.

Since interlocutors use similar backchannel-inviting cues, we can use information from the interlouctor – the system – to build the model. The influence of this interlocutor information can be weighted according to the probable strength of the entrainment effect, which can depend, as we have shown, on the feature being predicted, the respective genders of the participants, whether a feature value is an outlier, and where in the conversation the speech segment occurs.

5 Entrainment and dialogue quality

This section addresses two main research questions:

1. What kinds of entrainment are most important to conversational quality?
2. Will the passive benefits of entrainment apply when it is a computer that is entraining?

To answer the first question, we look at the entrainment correlates of social and objective variables in the Games Corpus (previously reported in Levitan et al., 2012). We address the second question with a Wizard of Oz study that looks at subjects' reactions to an entraining spoken dialogue system.

5.1 Entrainment correlates of dialogue characteristics

Lexical entrainment has been associated with measures of smooth turn-taking and task success (Nenkova et al., 2008). Here, we correlate entrainment on intensity mean and max, pitch mean and max, jitter, shimmer, noise-to-harmonics ratio (NHR), and syllables per second with four objective measures of dialogue coordination: number of turns, mean turn latency, percentage of overlaps, and percentage of interruptions. We interpret a high number of turns and percentage of overlaps (cases in which one person begins speaking as her interlocutor finishes his turn) as signs of a smoothly flowing, well-coordinated conversation. We therefore expect them to be positively associated with entrainment, in line with previous work and the theory that entrainment facilitates communication. In contrast, high turn latency (the lag time between turns) and percentage of interruptions (cases in which one person begins speaking before her interlocutor has finished his turn) are signs of poor turn-taking behavior and an awkward conversation. We therefore expect them to be negatively correlated with entrainment measures.

To look at more perceptual measures of dialogue quality, we used Amazon Mechanical Turk[2] to annotate each task (the sub-units of each game) in the Games Corpus for what we term *social variables*, the perceived social characteristics of an interaction and ints participants. Details on the annotation process can be found in (Gravano et al., 2011). In this study, we focus on four social variables: *trying to be liked*, *giving encouragement*, *trying to dominate*, and *conversation awkward*. Based on Communication Accommodation Theory (Giles et al., 1987), we expect the first two social variables, which represent the desire to minimize social distance, to be positively correlated with entrainment. Someone who is *trying to dominate*, on the other hand, will try to increase social distance, and we therefore expect this variable to correlate negatively with entrainment, as should *conversation awkward*.

We report separate results for female, male and mixed-gender pairs. In general, we see correlations in the expected directions: the number of turns, percentage of overlaps, and *giving encouragement* are positively correlated with entrainment for all gender groups, latency is negatively correlated with entrainment for male and female pairs, and *trying to be liked* is positively correlated with entrainment for

[2] http://www.mturk.com

male and mixed-gender pairs. We see no correlations for *trying to dominate*, possibly because annotators were confused between the socially weak position of *trying* to dominate, and the socially powerful position of actually dominating.

For objective variables, we see the strongest and most numerous correlations for male pairs, while for objective variables, this is true for mixed-gender pairs, leading us to conclude that entrainment is most important to the coordination of a conversation for male pairs and to the perceived quality of a conversation for mixed-gender pairs. We identify intensity as an important entrainment feature, as well as shimmer for dialogue coordination for female or mixed-gender pairs. In future work, we plan to correlate these social and objective variables with measures of local entrainment and convergence (Section 4.1).

5.2 Entrainment and dialogue quality in spoken dialogue systems

In this study (currently ongoing), we look at whether subjects will attribute more positive qualities to an interaction with a system whose voice is more similar to their own. To answer this question, we create a Wizard of Oz setup in which a subject interacts with an *entrained* voice and a *disentrained* voice. We chose to employ a wizard instead of a fully functional dialogue system in order to neutralize possible intrusions from other components of a dialogue system and isolate the entrainment effect.

The subjects are given three tasks modeled on reasons for which someone might call 311, New York City's phone number for government information. In the *taxi* scenario, for example, the subject is given a description of an incident in which a taxi drove unsafely, and is told to report the incident to the system, using the given date, time and location. Using this paradigm, we can collect spontaneous speech while still being able to use prerecorded prompts: the content is predetermined, but the sentence form and word choice is up to the subject.

For the first task, *alternate side parking*, the experimenter prints prompts to the subject's screen using a chat program, and the subject responds by speaking into a headset that plays into the experimenter's computer. The purpose of this first task is to get a sample of the subject's speech. The subject then fills out some demographic forms and the NEO-FFI personality test, while the experimenter calculates the vocal intensity and speaking rate of the subject's speech. A set of prerecorded prompts is then scaled to match the subject's vocal parameters, forming an *entrained* set, and then scaled away from the subject's parameters, forming the *disentrained* set. The parameters for the disentrained set were chosen empirically to result in a voice perceptibly different from the entrained set while remaining undistorted and natural-sounding.

The subject then completes two more tasks, one with the *entrained* voice and one with the *disentrained* voice. We vary the order and combination of tasks and voices so we can test for effects of order and task. After each task, the subject fills out a survey containing questions like "I liked the system's personality" or "I found talking with the system annoying." We hypothesize that they will agree more with positive statements about the entraining version of the system.

We also crudely measure each subject's perceptual sensitivity to vocal characteristics by asking them to describe each voice by choosing from a list of adjectives like "high-pitched," "fast," or "loud." We will look at how this sensitivity, as well as gender and personality, interact with the subjects' reactions to the system's entrainment.

6 Influencing user behavior

In human conversations, it is common for a speaker to attempt to affect his interlocutor's behavior by modeling a desired change. For example, a speaker may raise his own voice if he is having trouble hearing and wishes his interlocutor to speak more loudly. Since humans have been shown to entrain to computers (Coulston et al., 2002; Stoyanchev and Stent, 2009; Bell et al., 2003), it is reasonable for a spoken dialogue system to use this strategy to influence its user to speak in a way that will optimize the performance of its automatic speech recognition (ASR). A previous study (Lopes et al., 2011) successfully induced users to abandon words prone to ASR error simply by removing those words from the system's prompts. In this work, we attempt to influence users to abandon prosodic characteristics associated with ASR failure by modeling the desired change in the system's prompts.

Hirschberg et al. (2004) found that utterances that followed longer pauses or were louder, longer, or pitched higher were less likely to be recognized correctly. Our method looks for these undesirable prosodic features in utterances with low ASR confidence and attempts to induce the user to abandon them. We hypothesize that abandoning prosody associated with ASR failure will result in improved ASR performance.

Our approach is as follows. When the system's ASR returns a hypothesis with low confidence for an utterance, it finds the utterance's intensity, pitch and duration. If any of these features fall within the range of utterances that tend to be misrecognized, the system employs one of four strategies. The **explicit** strategy is to ask the user to make the desired change, e.g. "Please speak more quietly." The **entrainment** strategy is to model the desired change, e.g. lowering the intensity of the system's output. The **explicit+entrainment** strategy combines the two, e.g. by saying "Please speak more quietly" in a quieter system voice. We hypothesize that one strategy may increase the efficacy of the other. We will also try a **no strategy** condition as a baseline for how often the user independently abandons the undesirable prosody.

Each strategy will be embodied in a simple request for repetition. For each strategy, we will look at how often the subsequent turn displays the desired change in prosody. In addition, we will see how often the ASR performance improves on the subsequent turn. A third measure of a strategy's success will be the durability of its effect—that is, how likely the undesirable prosody is to recur later in the conversation.

Within the entrainment condition, we will test how pronounced a change must be in order to induce a corresponding change on the part of the user. Our research on outlier entrainment suggests that a more extreme change is more likely to be entrained to. However, the most attractive feature of the entrainment condition is its nondisruptiveness, and this quality will be lost if the change in the system's voice is too extreme. We will therefore begin with a slight change, and test how much the degree of change must be increased before the user will imitate it.

Fandrianto and Eskenazi (2012) implemented a similar approach, lowering the system's vocal intensity or increasing its speaking rate when its classifiers detected the speaking styles of shouting or hyperarticulation. By responding to individual prosodic features instead of higher-level speaking styles, we avoid the layer of error introduced by classifiers. Furthermore, our approach can account for cases in which ASR error is caused by prosodic features that do not comprise an identifiable speaking style. Finally, our detailed analysis will give more information about the advantages and limitations of each strategy.

7 Contributions

The studies of human-human conversations in this thesis will advance current understanding of how people entrain. We provide a cohesive picture of entrainment by directly comparing different measures on a single corpus, establishing that entrainment is both a global and a local phenomenon, that people entrain by value rather than by direction, and that it is a dynamic process, improving with the course of a dialogue. We show that speaker pairs entrain in a novel dimension, backchannel-inviting cues, and that this entrainment is associated with task success and dialogue coordination. We also show that the effect of entrainment is stronger in outlier cases, lending experimental support to the perception-behavior link.

This work provides experimental results on the utility of entrainment in conversations with both humans and spoken dialogue systems. In human conversations, we show that entrainment is correlated with positive social characteristics and turn-taking features. In our Wizard of Oz experiments, we will show how entrainment affects a user's perception of the quality of a spoken dialogue system.

Finally, this work shows how the principles of entrainment can be used to actively improve spoken dialogue systems. We will build a framework for implementing the results of our studies of entrainment in human conversations into prediction models, which we hypothesize will improve their accuracy and can be used to improve a system's performance. In our influencing experiments, we will attempt to influence a user to speak in a way that will optimize ASR performance simply by changing the system's own voice.

References

Linda Bell, Joakim Gustafson, and Mattias Heldner. Prosodic adaptation in human-computer interaction. In *Proceedings of ICPHS'03*, pages 833–836, 2003.

Susan E. Brennan and Herbert H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22(6): 1482–1493, 1992.

T. L. Chartrand and J. A. Bargh. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.

Rachel Coulston, Sharon Oviatt, and Courtney Darves. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of ICSLP'02*, 2002.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words! linguistic style accommodation in social media. In *Proceedings of WWW*, 2011.

Andrew Fandrianto and Maxine Eskenazi. Prosodic entrainment in an information-driven dialog system. In *Proceedings of Interspeech*, 2012.

H. Giles, A. Mulac, J.J. Bradac, and P. Johnson. *Speech accommodation theory: the first decade and beyond*. Sage, Beverly Hills, CA, 1987.

Agustín Gravano. *Turn-taking and affirmative cue words in task-oriented dialogue*. PhD thesis, Columbia University, 2009.

Agustín Gravano and Julia Hirschberg. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of Interspeech*, 2009.

Agustín Gravano, Rivka Levitan, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. Acoustic and prosodic correlates of social behavior. In *Proceedings of Interspeech*, 2011.

Julia Hirschberg, Diane Litman, and Marc Swerts. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43:155–175, 2004.

Chi-Chun Lee, Matthew Black, Athanasios Katsamanis, Adam Lammert, Brian Baucom, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth Narayanan. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Proceedings of Interspeech*, 2010.

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

Rivka Levitan, Agustín Gravano, and Julia Hirschberg. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N12-1002`.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. Towards choosing better primes for spoken dialog systems. In *ASRU'11*, pages 306–311, 2011.

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

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL/HLT*, 2008.

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

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

David Reitter and Johanna D. Moore. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, 2007.

Svetlana Stoyanchev and Amanda Stent. Lexical and syntactic priming and their impact in deployed spoken dialog systems. In *Proceedings of NAACL HLT*, 2009.

Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1 (4):285–293, 1990.

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