Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue

Rivka Levitan

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue

Rivka Levitan

Entrainment (sometimes called adaptation or alignment) is the tendency of human speakers to adapt to or imitate characteristics of their interlocutors' behavior. This work focuses on entrainment on acoustic-prosodic features. Acoustic-prosodic entrainment has been extensively studied but is not well understood. In particular, it is difficult to compare the results of different studies, since entrainment is usually measured in different ways, reflecting disparate conceptualizations of the phenomenon. In the first part of this thesis, we look for evidence of entrainment on a variety of acoustic-prosodic features according to various conceptualizations, and show that human speakers of both Standard American English and Mandarin Chinese entrain to each other globally and locally, in synchrony, and that this entrainment can be constant or convergent. We explore the relationship between entrainment and gender and show that entrainment on some acoustic-prosodic features is related to social behavior and dialogue coordination. In addition, we show that humans entrain in a novel domain, backchannel-inviting cues, and propose and test a novel hypothesis: that entrainment will be stronger in the case of an outlier feature value. In the second part of the thesis, we describe a method for flexibly and dynamically entraining a TTS voice to multiple acoustic-prosodic features of a user's input utterances, and show in an exploratory study that users prefer an entraining avatar to one that does not entrain, are more likely to ask its advice, and choose more positive adjectives to describe its voice.

This work introduces a coherent view of entrainment in both familiar and novel domains. Our results add to the body of knowledge of entrainment in human-human conversations and propose new directions for making use of that knowledge to enhance human-computer interactions.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my warmest gratitude to my advisor, Julia Hirschberg, who defines everything a mentor should be. During my time at Columbia, I have benefited profoundly from her support, encouragement, knowledge, advice, and unstinting efforts on my behalf. She has always been available, always encouraging, and her guidance in every aspect of professional life has always been outstanding. I am endlessly grateful.

I would like to thank my dissertation committee — Kathleen McKeown, Mike Collins, Mari Ostendorf, and Susan Brennan — for the time and effort they generously devoted to the review of this thesis, and for their many valuable comments.

I would also like to thank Agustín Gravano, Stefan Benus, Shirley Xia, and Ani Nenkova, who collaborated on several portions of this thesis, and with whom I had many fruitful conversations. I am especially grateful to Agus, whose data tables and scripts were extremely useful, and who was always available for questions.

It has been such a pleasure working alongside the members of the Speech Lab, past and present, with whom I had many enjoyable conversations about entrainment, and even more enjoyable conversations about everything else: Bob Coyne, Fadi Biadsy, Erica Cooper, Morgan Ulinski, Daniel Bauer, Laura Willson, Anna Prokofieva, Victor Soto, Sarah Ita Levitan, Svetlana Stoyanchev, and all the interns and visitors.

My deepest thanks go to my mother, Debbie Sturm, my personal and professional role model. I never would have started graduate school if not for you, and I definitely could never have finished it without you. I hope to someday be half the teacher that you are.

I want to thank the rest of my family for their love and support, both emotional and tangible, and especially my husband and best friend, Chaim Simcha Lieberman, who took care of everything so I could work on this thesis, and who always reminded me to stop working, and my girls, Racheli and Tehila, who did nothing to help with this thesis and in fact materially impeded its progress, but who light up every day. You are my greatest joy.

Chapter 1

Introduction

In recent years, spoken dialogue systems (SDS) have begun to replace or supplement human operators, and devices such as cars, phones and televisions have moved towards eyes-free interfaces powered by speech technologies. These shifts have been made possible by advances in text to speech (TTS) and automatic speech recognition (ASR), whose performance under optimal conditions can approach human levels. However, although TTS can produce speech that is intelligible and clear, it often is perceived as stiff and unnatural. With the problem of intelligibility solved, the next step for TTS is to implement characteristics of human speech that contribute to what we perceive as natural speech behavior.

Entrainment, also known as coordination, adaptation, or alignment, is one of the most fascinating of such behaviors. It is the tendency of human speakers to adapt to, or to some extent imitate, the conversational behavior of their interlocutors. Humans have been shown to entrain to each other in multiple aspects of speech, including gestural [Chartrand and Bargh, 1999], acoustic [Natale, 1975], phonetic [Pardo, 2006], lexical [Brennan and Clark, 1996], and syntactic [Reitter and Moore, 2007]. The evidence of entrainment at nearly every level of communication attests to its importance in human dialogue.

Entrainment is of interest to a variety of disciplines, both because of its prevalence in human dialogue and its association in the literature with numerous positive conversational qualities. It has been studied in the context of hostage negotiation [Taylor and Thomas, 2008], tutorial dialogues [Ward and Litman, 2007], criminal trials [Gnisci, 2005], oral arguments in the Supreme Court [Danescu-Niculescu-Mizil *et al.*, 2011], Wikipedia editor discussions [Danescu-Niculescu-Mizil *et al.*, 2011], and therapy interactions between married couples [Lee *et al.*, 2010], among others. In many domains, it has been linked to rapport, increased liking for the interlocutor, improved negotiation outcomes, power dynamics, task success, and positive therapy outcomes, in both experimental and real-world conditions. This relationship is often explained by Communication Accommodation Theory [Giles *et al.*, 1991], which posits that speakers dynamically adjust their communication behaviors, converging to or diverging from their interlocutor in order to diminish or increase social distance.

This body of literature suggests that a conversational agent that entrains to its interlocutor – a human user – can thereby implicitly improve the user's perception of its performance. Even if the relationship between entrainment and an improved conversation is not causal – although some of the literature indicates that a human's perception of an interlocutor's quality can in fact be improved by causing the interlocutor to entrain – the fact of the agent's adopting this humanlike behavior is likely to make the experience of interacting with it feel more natural to the user. Implementing entrainment in a conversational agent can provide an increase in perceived output quality that is orthogonal to the separate efforts of improving pronunciation, intelligibility, and prosody.

While entrainment has been extensively explored, most studies have focused on single aspects of entrainment. This thesis presents a broad, multidimensional study of entrainment, in which multiple aspects of the phenomenon are explored using a single paradigm and a single corpus. This approach yields a comprehensive view of entrainment in various dimensions and aspects, which is one of the two main contributions of this thesis. In addition to the scientific value inherent in the exploration of a phenomenon so prevalent in human behavior, such a study can motivate the design of conversational agents to implement entrainment as we see it in human conversations.

The second main contribution of this thesis is a method for implementing dynamic entrainment in a conversational agent. This method applies our findings from studies of human-human conversations to implement the capability to align acoustic-prosodic features of the agent's output to the user's input. Instead of maintaining constant prosody throughout, an entraining agent can dynamically respond and adapt to the user's speech, just as a human interlocutor would. This part of the thesis includes a study of how an agent's entrainment behavior affects how it is perceived and related to by a human user.

This thesis focuses on entrainment on acoustic-prosodic features. Specifically, we look at intensity (loudness), pitch, speaking rate, and three measures of voice quality: jitter, shimmer, and NHR (noise-to-harmonics ratio). The features represent some of the major characteristics of a person's voice. They are especially suitable to our research goals because they can be automatically measured, and so systems can extract these features and entrain on them in real time.

This thesis is organized as follows. Part I describes a series of empirical studies of entrainment in numerous aspects of human-human conversation. Part II presents a method for implementing entrainment in a conversational agent, and a study of the effect an entraining agent has on user interactions with a system. Part III discusses the conclusions and significance of this work.

Part I

Entrainment in Human Conversations

Chapter 2

Motivation and Research Goals

Entrainment, a term commonly used to refer to the tendency of a speaker to converge her communicative behavior to that of her interlocutor, has long been of interest to researchers. The phenomenon has been shown to occur at all levels of communication, from acoustics to syntax, and has been studied in numerous and varied contexts, including marital therapy, tutorial dialogues, and hostage negotiations.

Entrainment is often explained by Giles's Communication Accommodation Theory [Giles *et al.*, 1991], which states that interlocutors dynamically adjust their communication behavior in order to create, maintain, or diminish the social distance between them. For example, a salesman may mimic a customer's posture; an trial lawyer may reject a witness's terminology; a talk show host may adopt her guest's accent.

An alternative explanation was proposed by Chartrand and Bargh [Chartrand and Bargh, 1999], who posit that the perception-behavior link is the mechanism behind entrainment. In human cognition, the processes of perception and production are closely linked, such that the fact of perceiving a behavior makes a person more likely to adopt it. According to this view, entrainment is an automatic rather than social process, although it may be mediated by social factors related to the degree to which the interlocutor's behavior is observed. Although we do not attempt to explicitly test either of these theories in the scope of this thesis, they inform our analyses and interpretation.

Most research on entrainment follows one of two paths. One line of research attempts to demonstrate entrainment *explicitly* using statistical corpus analysis. The other demonstrates entrainment *implicitly* by showing that measurements of similarity correlate with some external metric. The first path makes a strong claim that humans do entrain to their interlocutors on a given feature ("entrainment" may denote different aspects of similarity, as discussed in Chapter 5), while the second claims only that the extent to which humans entrain to their interlocutors is associated with an orthogonal feature. Most of the studies in this section follow the first path: applying statisical methods to determine whether humans entrain to their interlocutors in various ways.

Our main research questions are the following:

On what features do people entrain? In Chapter 5, we explore different definitions of entrainment and whether the data supports each one for a set of acoustic-prosodic features. This analysis is replicated in Chapter 8 on Mandarin Chinese data for the first cross-linguistic comparison of acoustic-prosodic entrainment. Chapter 7 introduces entrainment in a novel dimension, backchannel-inviting cues.

What factors affect how people entrain? We then consider each interlocutor-interlocutor gender pair separately to see how entrainment interacts with gender (Chapter 9). In Chapter 6, we look at how entrainment may be affected in the case of feature values that diverge from the norm.

What happens when people entrain? Finally, in Chapter 10 we look at the associations between entrainment and various descriptors of social behavior, both objective and subjective.

Studies of entrainment have largely focused on individual aspects of the phenomenon. In addition to the individual contributions of each study in this part of the thesis, a broader contribution of this section is a comprehensive view of entrainment, which emerges from the exploration of multiple aspects of entrainment using a single paradigm and a single corpus.

Most studies in this section were conducted on the Columbia Games Corpus, a collection of spontaneous Standard American English task-oriented conversations between strangers, which is described, along with units of analysis and the acoustic-prosodic features under discussion, in Chapter 4. Our conclusions, therefore, may not be applicable to other types of dialogue; in fact, we might expect entrainment behavior to be different in situations with a different social dynamic, such as when the power relationship is imbalanced or intimacy exists between the interlocutors, or when a task is competitive rather than cooperative. However, the dialogue in the Columbia Games Corpus is closely related to the genre of greatest interest to us: task-oriented conversation between a user and a conversational agent.

Chapter 3

Related Work

The literature uses numerous terms to refer to the phenomenon of entrainment, including "accommodation", "alignment", "adaptation", and "convergence." In the psychological literature, "entrainment" usually refers specifically to the synchronization of temporal cycles (as in [Manson *et al.*, 2013]), while "convergence" is the general term for the phenomenon when it results in increased similarity (as in [Giles *et al.*, 1991; Street, 1984; Natale, 1975]). We follow the usage of the computer science literature, which uses "entrainment" as a general term (as in [Brennan and Clark, 1996; Nenkova *et al.*, 2008; Lee *et al.*, 2010; Fandrianto and Eskenazi, 2012; Friedberg *et al.*, 2012], etc.). Following [Edlund *et al.*, 2009], we use the term "convergence" to refer specifically to an *increase* in similarity, and "synchrony" to refer to the degree to which interlocutors vary their speech in synchrony.

"Accommodation" and "adaptation" usually refer to a process of audience design in which the speaker takes the specific communication needs of her interlocutor into account when formulating an utterance (such as speaking more slowly to a child) but not necessarily striving for similarity (the child may not have spoken slowly). [Brennan and Hanna, 2009] is one example of this usage; we use both terms in this way. "Alignment" is usually used in the context of the cognitive processes associated with entrainment ([Pickering and Garrod, 2004]) but is also commonly used to refer to dynamic lexical entrainment (as in [Brockmann *et al.*, 2005; Buschmeier *et al.*, 2009; de Jong *et al.*, 2008]). We use "alignment" to refer to all forms of dynamic matching.

Several theoretical models of entrainment have been proposed in the literature. The most influential, Communication Accommodation Theory (CAT) [Giles *et al.*, 1991], proposes that speakers converge to or diverge from their interlocutors in order to attenuate or accentuate social differences. For example, [Bourhis and Giles, 1977] described Welsh speakers speaking with a much more strongly marked accent when their interlocutor was a British speaker who referred to Welsh as “a dying language which had a dismal future.” An alternative pragmatic view of entrainment is the “communication model” of convergence proposed by [Natale, 1975] in an earlier work, which posits that “matching or convergence of non-content speech behavior is an automatic process used by the speaker to attain the *optimal* ‘format’ of his speech behavior so that his message will be intelligible” (emphasis added). Both theories view entrainment as functional, but while CAT holds that speakers attempt to become *similar* to their interlocutor, the communication model considers entrainment to be a form of audience design in which a speaker assumes that her interlocutor is utilizing the speech behavior that he would prefer her to use. Convergence of speech rates, duration of utterance, speech latency and vocal intensity have been shown to occur and can be explained according to this model, which, although Natale conceptualizes it specifically in terms of automatic speech behavior, can also be applied to lexical entrainment. Natale showed that the degree to which a speaker converges to her interlocutor’s vocal intensity can be predicted by the speaker’s *social desirability*, or “propensity to act in a social manner,” which is evidence in support of both the communication model and CAT.

[Chartrand and Bargh, 1999] challenge the notion that mimicry of one’s interlocutor is intentional and goal-directed. They suggest that humans behave like chameleons rather than apes, passively and unconsciously reflecting the social behavior of their interlocutor. Many functions of social perceptual activity are automated — that is, unconscious — and numerous studies have shown that social perception has a passive, unintended effect on social behavior. They posit that the psychological mechanism behind entrainment is the perception-behavior link, the finding that the act of observing another’s behavior increases the likelihood of the observer’s engaging in that behavior. In a series of experiments, they showed that subjects mimicked the mannerisms and facial expressions of a confederate without being aware of the confederate’s mannerisms. The effect of this mimicry was not affected by whether the confederate was smiling or unsmiling throughout the interaction, which they consider a proxy for likability; they conclude that entrainment is *not* mediated by social goals. Instead, they theorize that because of the link between perspective taking and social skills, high-perspective takers are likely to be better at promoting smooth interactions, and therefore more likely to mimic their interlocutors; this hypothesis was supported by their data.

In the same study, Chartrand and Bargh found that subjects who interacted with a confederate who mirrored their behavioral mannerisms liked the confederate more and thought the interaction went more smoothly. When asked in a debriefing whether they had noticed anything in particular about the confederate's behavior or mannerisms, no subject reported noticing the mimicry. Although Chartrand and Bargh, like the proponents of Communication Accommodation Theory, hypothesize an association between entrainment and positive social goals, they view the process of entrainment as automatic rather than intentional. [Pickering and Garrod, 2004] propose an analogous theory relating specifically to language, which holds that production is automatically linked to perception.

In a study of perceptual learning and production changes in an ambiguous pronunciation, [Kraljic *et al.*, 2008] decouple the processes of perception and production and show that at least in this domain, changes in perception did not trigger corresponding changes in production, suggesting that the link between processes of perception and production is *not* automatic, but can be mediated by pragmatic goals. They further suggest that the strength of this link may relate to the levels of representation activated in perception and production, since at the motor level it may be difficult to override the effect of practice. Characterizing speech features in terms of their levels of representation may help reconcile disparate findings regarding perception, production, and pragmatic context.

Empirical evidence of acoustic-prosodic entrainment has been well-documented in the literature. In one of the earliest studies of entrainment, [Matarazzo and Wiens, 1967] showed that an interviewer could manipulate an interviewee's response time latency by increasing or decreasing his own duration of silence before responding. Similarly, using time series regression procedures, [Street, 1984] showed that interviewees converged towards their interviewers on response latency and speech rate. [Natale, 1975] manipulated a confederate's within-conversation intensity levels to show that subjects who were engaged in open-ended conversation entrained to each new intensity level; this entrainment increased (converged) over the course of a conversation and again when the subject returned for subsequent sessions. Using Fourier series analysis as implicit measures of pitch and intensity, [Gregory *et al.*, 1993] found that similarity was greater in true conversations than in conversations simulated by splicing together utterances from speakers who did not actually interact. [Ward and Litman, 2007] measured local entrainment on loudness and pitch by fitting regression lines to the feature values found at each distance d from a prime, which was defined as a feature value more than one standard deviation away from the speaker mean; this measure was able to successfully distinguish randomized from naturally-ordered data. According to this measure, students in tutorial dialogues converged to their tutor on max and mean amplitude and diverged on min pitch. At the phonetic level, [Pardo, 2006] found that listeners judged an item spoken in the course of a task dialogue to be more similar to a corresponding production spoken by the talker's conversational partner in the course of that dialogue than to productions spoken by the same partner before or after the task.

Entrainment has been shown to occur at lexical and syntactic levels of communication as well. [Brennan and Clark, 1996; Brennan, 1996; Bortfeld and Brennan, 1997] showed that conversational partners entrain on referring expressions: having negotiated a given conceptualization of an object, they continued to refer to the object in that way even when a new context made the expression overly informative. [Metzing and Brennan, 2003] showed that when an interlocutor broke the conceptual pact by using a new referring expression to refer to a previously discussed object, subjects were slower to locate the object than they were when a new partner used a new referring expression. [Branigan *et al.*, 2000] showed that the syntactic structure of a subject's utterance was influenced by the syntactic structure of a confederate's previous utterance. [Reitter *et al.*, 2010] demonstrated that in task-oriented conversations — but not in open-ended dialogues — speakers were more likely to use a given syntactic rule soon after their interlocutor had used that rule. [Niederhoffer and Pennebaker, 2002] and [Danescu-Niculescu-Mizil *et al.*, 2011] found evidence of Linguistic Style Matching (LSM), coordination on a number of stylistic categories, in chat dialogues and Twitter conversations, respectively. [Michael and Otterbacher, 2014] found that stylistic coordination occurs even in asynchronous communication, showing that reviewers of tourist attractions were influenced by preceding reviews' use of stylistic features; they call the phenomenon "herding."

Motivated by the theories relating entrainment to pragmatic goals, some studies have explored the relationship between entrainment and extrinsic metrics. [Niederhoffer and Pennebaker, 2002] looked at the associations between Linguistic Style Matching (LSM), the degree to which participants in chat conversations match each other on the use of word categories representing a variety of psychological and linguistic dimensions, and a "click index", "the degree to which participants felt the interaction went smoothly, they felt comfortable during the interaction, and they truly got to know the other participant." In contrast to the findings of [Chartrand and Bargh, 1999] and the predictions of CAT, they did not find that LSM was correlated with the quality of the conversation as reported either by the participants or by external raters. They propose a coordination-engagement hypothesis as an alternative to the coordination-rapport hypothesis: "[T]he more that two people in a conversation are actively engaged with one another — in a positive or even negative way — the more verbal and nonverbal coordination we expect."

In a finding that would seem to contradict this hypothesis, [Ireland *et al.*, 2011] showed that the degree of Linguistic Style Matching in speed date transcripts predicted the likelihood of mutual romantic interest, and LSM in couples' instant messages was associated with relationship stability three months later. At the acoustic-prosodic level, [Lee *et al.*, 2010] demonstrated that in interactions between "seriously and chronically distressed married couples" discussing a problem in their relationship, latent measures of entrainment on pitch were significantly higher during interactions labeled as "highly positive affect" by a trained evaluator. Similar entrainment measures derived from *energy* features, however, were *not* predictive of positive or negative affect, illustrating the need to treat entrainment in different dimensions separately. [Manson *et al.*, 2013] found that partners who converged on speech rate were more likely to later cooperate in a prisoner's dilemma, although they did not evaluate each other more positively, while partners with a higher degree of LSM **did** evaluate each other more positively, but were no more likely to cooperate. In an early work, [Street, 1984] showed that interviewer-interviewee response latency *similarity* and response latency and speech rate *convergence* were positively associated with participants' judgments of their interlocutors' competence and social attractiveness.

Several studies have reported a link between entrainment and task success. [Reitter and Moore, 2007] found that lexical and syntactic entrainment in just the first five minutes of an interaction was predictive of task success in the HCRC Map Task Corpus. [Nenkova *et al.*, 2008] showed that the degree of entrainment on high-frequency words and affirmative cue words correlated significantly with game score in the Columbia Games Corpus; in addition, such entrainment was significantly associated with dialogue naturalness and smooth interaction flow. [Friedberg *et al.*, 2012] showed that high-performing student engineering groups became more similar over time in their use of project-related words, while low-performing groups tended to become less similar over time. [Thomason *et al.*, 2013] found that student entrainment to a tutoring dialogue system on several pitch features was positively correlated with learning gain, especially for students with high scores on a pretest. These results can be explained by the engagement-rapport hypothesis: a higher degree of engagement in a task (as represented by a higher degree of entrainment) is likely to lead to success.

[Danescu-Niculescu-Mizil *et al.*, 2012] looked at the relationship between linguistic style accommodation and power differentials between interlocutors in Wikipedia editor discussions and United States Supreme Court oral arguments. They found that Wikipedia users coordinated more with other users after those users were promoted to admin than before, and lawyers appearing before the Supreme Court coordinated more with justices than justices did to lawyers. Wikipedia users coordinated more with users who voted on the opposite side that to those who voted the same way, and lawyers coordinated more with justices who eventually voted against them than to justices who eventually voted in their favor, in accordance with the assumption that when Person A needs Person B's approval, B is in a position of more power if she disagrees with A than if she agrees with him. However, [Benus *et al.*, 2012; Beňuš *et al.*, 2014] found that lawyers' filled pauses ("um", "er", etc.) were more similar to those of the justices who voted for them and to those of the justices who voted against them. This analysis adds a dimension to the interaction between entrainment and speaker characteristics: entrainment can relate to speakers' relationships with their interlocutor as well as more intrinsic speaker traits.

Throughout the entrainment literature, the same term often is used to refer to divergent conceptualizations of entrainment. [Natale, 1975; Street, 1984; Coulston *et al.*, 2002; Pardo, 2006; Ward and Litman, 2007] all investigate "convergence"; Natale, Coulston et al., and Pardo each look at global similarities; Street uses time series regression techniques to measure turn by turn dynamic similarity; Ward and Litman measure user adoptions of discrete primes. [Brennan and Clark, 1996; Niederhoffer and Pennebaker, 2002; Nenkova *et al.*, 2008; Lee *et al.*, 2010] look at "entrainment"; Brennan and Clark look at global similarity, as do Nenkova et al., and Niederhoffer and Pennebaker and Lee et al. look at turn by turn dynamic similarity.

The body of literature on entrainment is prodigious, yet most studies conceptualize the phenomenon differently — globally or at the turn level, similarity by value or by direction, static or increasing similarity. In addition to introducing new dimensions to the research on entrainment, this thesis revisits dimensions and measurements that have already been explored in order to study them as an ensemble, in a consistent way.

Chapter 4

Data and Features

This chapter describes the corpus, features, and units of analysis used in this part of the thesis.

4.1 Columbia Games Corpus

The Columbia Games Corpus is a collection of 12 spontaneous dyadic conversations between native speakers of Standard American English. The corpus was collected and annotated by the Spoken Language Processing Group at Columbia University and the Department of Linguistics at Northwestern University in October 2004 as part of a project on prosodic variation in Standard American English (SAE).

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. Table 4.1 shows the subjects who participated in each session (each identified by a unique number) and their genders and approximate ages.

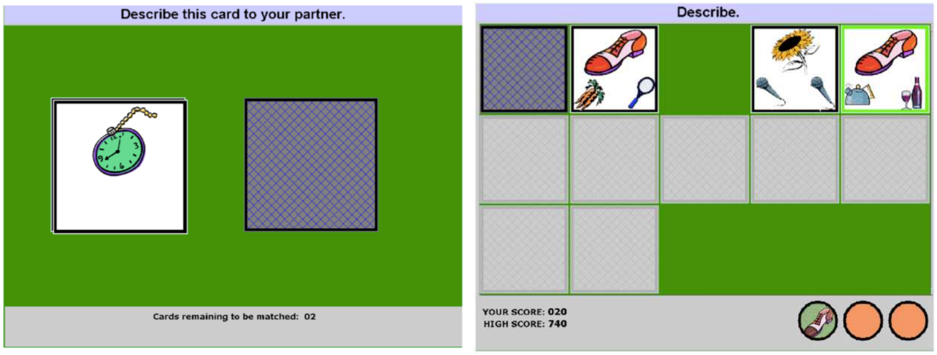
Session	*Speaker A*			*Speaker B*		
	Id	*Gender*	*Age*	*Id*	*Gender*	*Age*
1	101	male	25	102	male	25
2	103	female	25	104	male	25
3	105	female	25	106	male	30
4	107	male	30	108	male	45
5	109	female	50	101	male	25
6	108	male	45	109	female	50
7	110	female	50	111	female	20
8	102	male	25	105	female	25
9	113	male	20	112	female	20
10	111	female	20	103	female	25
11	112	female	20	110	female	50
12	106	male	30	107	male	30

Table 4.1: Participants in the Columbia Games Corpus

In order to elicit spontaneous, task-oriented speech, subjects were asked to play a series of four computer games. The games were designed to require cooperation and communication in order to achieve a high score. Participants were motivated to do well by a monetary bonus that depended on the number of points they achieved in each game. All games were played on separate laptops whose screens were not visible to the other player; the players were separated by a curtain so that all communication would be vocal. During game play, keystrokes were captured and were later synchronized with the speech recordings and game events.

The subjects first played three instances of the **Cards game**, which involved matching cards with one to four images on them, from a set of images of two sizes (small or large) and various colors. The images were selected to have descriptions that were voiced and sonorant, to improve pitch track computations. Each Cards game consisted of two parts. In the first part of each Cards game (Figure 4.1a), one player described each card in her deck while the other player searched through his own deck for the card being described. In the second part (Figure 4.1b), each player displayed a board of 12 cards, with no more than three face-up at any time; one player was the **Describer**, describing the appearance of a card on her screen, while the other was the **Searcher**, searching through his own board to find a similar card. For each card, the partners received points based on how many images the matched card had in common; they could therefore choose to “settle” for a partial match or continue searching for a better possibility. This and other complexities of the game were intended to make the game more interesting and promote discussion between the players.

After three instances of the Cards game, subjects played an **Objects game**. Each player’s screen displayed 5 to 7 objects, chosen as in the Cards game to have voiced and sonorant descriptions. The two players saw the same set of objects, arranged in the same positions on the screen, except for one, the target object, which appeared at the bottom of the **Follower’s** screen (Figure 4.2b) and in a random location among the other objects on the **Describer’s** screen (Figure 4.2a), blinking so that it could be identified as the target. The Describer was told to describe the position of the blinking object as exactly as she could so that the Follower could move the corresponding object to the identical position on his screen. Points were awarded based on how well the Follower’s target location matched the



Describe this card to your partner.

Cards remaining to be matched: 02

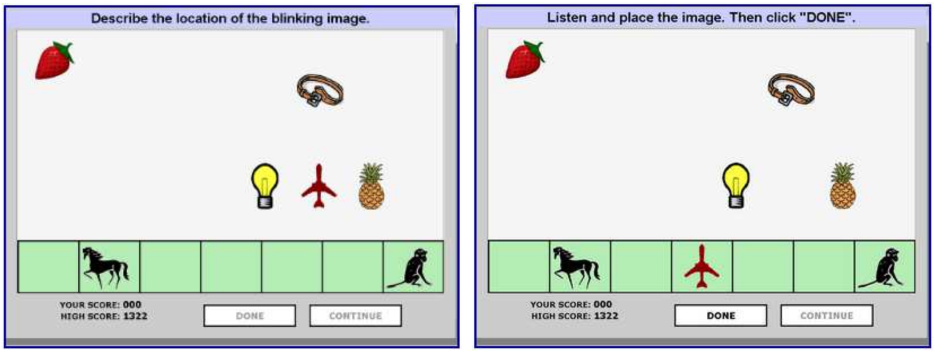
Describe.

YOUR SCORE: 020
HIGH SCORE: 740

(a) Part 1

(b) Part 2

Figure 4.1: Sample screenshots from the Cards Game.



Describe the location of the blinking image.

YOUR SCORE: 000
HIGH SCORE: 1322

Listen and place the image. Then click "DONE".

YOUR SCORE: 000
HIGH SCORE: 1322

(a) Describer's screen

(b) Follower's screen

Figure 4.2: Sample screenshots from the Objects Game.

Describer's. There were 14 tasks in the Objects game. In the first four tasks, one subject was always the Describer and the other was always the Follower; in the next four tasks, they switched roles; in the final six tasks, they alternated roles with each new task.

There are approximately 9 hours and 13 minutes of speech in the Games Corpus, of which approximately 70 minutes come from the first part of the Cards game, 207 minutes from the second part of the cards Game, and 258 minutes from the Objects game. On average, each session is approximately 46 minutes long, comprised of three Cards games of approximately 8 minutes each and one Objects game, which is approximately 22 minutes long.

The recordings took place in a double-walled soundproof booth. Each subject was recorded on a separate channel of a DAT recorder, at a sample rate of 48kHz with 16-bit precision, using a Crown head-mounted close-talking microphone.

The corpus has been orthographically transcribed and manually word-aligned by trained annotators. In addition, disfluencies and other paralinguistic events such as laughs, coughs and breaths were marked by the annotators. The corpus has also been annotated prosodically according to the ToBI framework ([Silverman *et al.*, 1992]); all turns have been labeled by type; affirmative cue words have been labeled according to their pragmatic functions; and all questions have been categorized by form and function. The annotation of the Games Corpus is described in detail in [Gravano, 2009].

4.2 Units of analysis

Throughout the thesis, we compute and compare features from the following units of analysis:

A **session** is a complete interaction between a pair of interlocutors. It consists of three Cards games and one Objects game. On average, a session is approximately 29 minutes long ($SD = 8.84$). There are 12 sessions in the Games Corpus.

An **inter-pausal unit (IPU)** is a pause-free chunk of speech from a single speaker. The threshold for pause length used here is 50 ms; this number was derived empirically from the average length of stop gaps in the corpus.

A **turn** is a consecutive series of IPUs from a single speaker. We include in our definition of "turns" utterances that are not turns in the discourse sense of the term, such as backchannels or failed attempts to take the floor. All turns in the Games Corpus have been labeled by type based on whether the speaker intends to take the floor, whether he or she successfully does so, and whether the turn overlaps with the previous turn.

4.3 Acoustic-prosodic features

In this thesis, we examine entrainment in eight acoustic-prosodic features that are typically of interest in speech research:

- Intensity mean

- Intensity max
- Pitch mean
- Pitch max
- Jitter
- Shimmer
- Noise-to-harmonics ratio (NHR)
- Syllables per second

All features except syllables per second have been computed using Praat [Boersma and Weenink, 2012], an open-source audio processing tool.

Intensity, also commonly referred to as amplitude or energy, describes the degree of energy in a sound wave. It is perceived as the volume of a sound. Intensity values were computed using Praat's native algorithm. We look at the mean intensity over a speech segment as well as the maximum.

Pitch describes the fundamental frequency of a voice, or how often the sound wave repeats itself. In general, the pitch of a voice depends on the length of the speaker's vocal tract, which is why women usually have higher-pitched voices than men; in the Games Corpus, female pitches range from about 75 to 500, while male pitches range from about 50 to 300. To allow for meaningful comparisons between male and female speakers, female pitch tracks were scaled to lie in the same range as the male pitch tracks, so that a raw female pitch value of 75 will have a scaled value of 50. Pitch tracks were computed using Praat; we look at mean and max pitch values.

Jitter, shimmer, and noise-to-harmonics ratio (NHR) are 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. However, in a study of how these features are perceived [Kreiman and Gerratt, 2005], listeners were inconsistent in their perception of jitter and shimmer, although agreement was high for NHR, suggesting that humans are not sensitive to jitter and shimmer. Jitter, shimmer and NHR were computed using Praat's "Voice report" command.

Syllables per second describes an utterance's speech rate. Syllable counts were obtained from the orthographic transcriptions using a syllable dictionary.

4.4 Significance testing

Multiple statistical tests are conducted in the course of our analysis throughout this thesis. All significance tests 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 $\frac{k \times \alpha}{n}$. We also consider a result to approach significance if its uncorrected p value is less than 0.05.

Chapter 5

Acoustic-prosodic entrainment

We begin our study of entrainment in human conversations with a broad analysis of acoustic-prosodic entrainment in the Columbia Games Corpus. By looking at whether evidence of entrainment exists according to a variety of measurements, we bring clarity to the question of how humans entrain to each other. Specifically, we address three questions:

Is entrainment global or local? *Global* entrainment implies that a feature of a speaker's speech, as represented by a single measure (such as the mean) over an entire conversational segment, is similar to that of her interlocutor. We can determine that this similarity is significant by comparing it to a baseline similarity, such as her similarity to a random other speaker, or to the same interlocutor under different conditions (e.g. after they have been speech for some time, or when the topic of the conversation has changed, or after some intervention).

Local entrainment refers to a dynamic alignment between interlocutors that occurs *within* a conversation, independent of the overall similarity between the two across the entire conversation. A pair of interlocutors may entrain globally, fluctuating around similar means, yet diverge widely at every point in the conversation (as illustrated in Figure 5.1). Alternatively, they may be globally dissimilar, but still continuously make incremental adjustments in alignment with their interlocutor at each turn. A third possibility allows for both global and local entrainment in a single interaction: the pair may be similar both globally and locally.

Do speakers match their interlocutors exactly or relatively? If entrainment

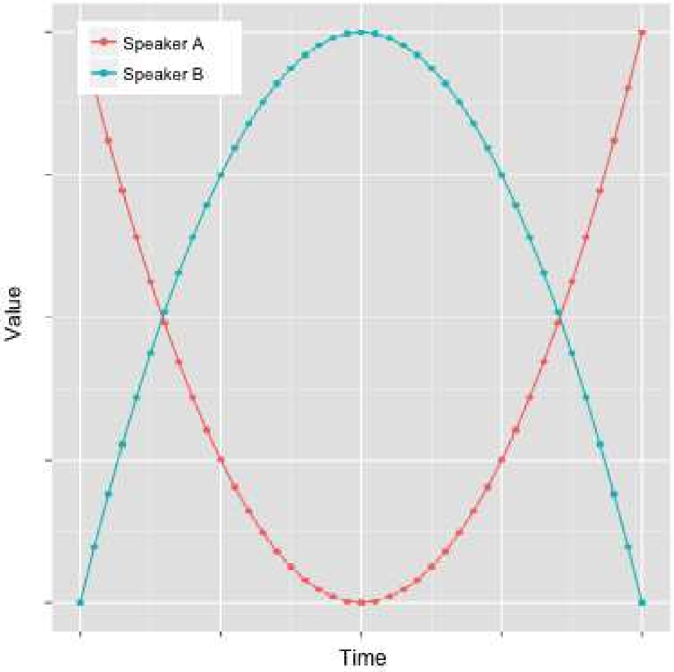


Figure 5.1: Global versus local entrainment.

entails continuous local adjustments, are these adjustments exact — to match the interlocutor's actual level — or relative? Consider a speaker who raises her pitch. An interlocutor who entrains by *value* will adjust his own pitch to match hers, while one who entrains *relatively* will adjust his pitch to a corresponding level within his own range, although this might not be similar to her actual pitch value at all. In Figure 5.2, Speaker C's feature values are similar to Speaker B's, and Speaker A's are dissimilar but are synchronized with Speaker B's.

Does entrainment improve over the course of a conversation? Is it the product of a single coordination step at the beginning of an interaction, or do the interlocutors become "better" at entraining, increasing in similarity as they continue to converse? As illustrated in Figure 5.3, such an increase in similarity, or **convergence**, can occur independently of overall similarity; a pair of interlocutors can be globally similar throughout an interaction without any increase in similarity (Speakers C and D), or they can increase in similarity throughout the interaction without ever becoming objectively similar (Speakers A and B).

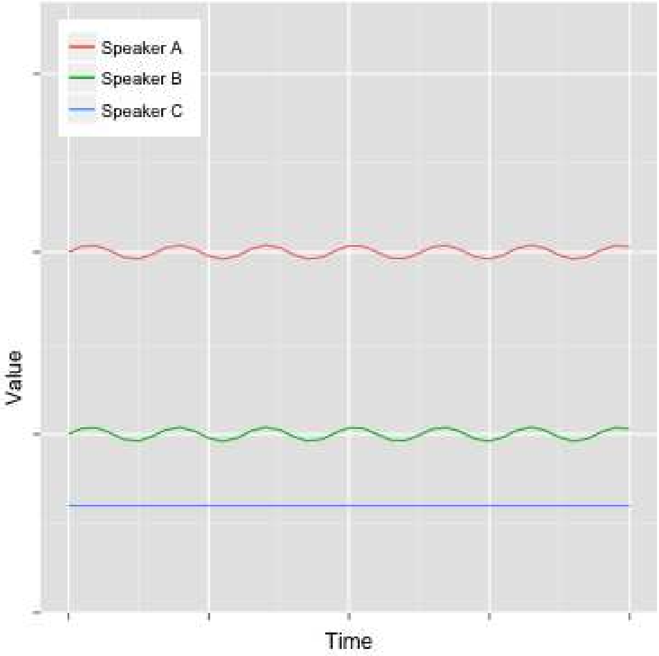


Figure 5.2: Exact versus relative entrainment.

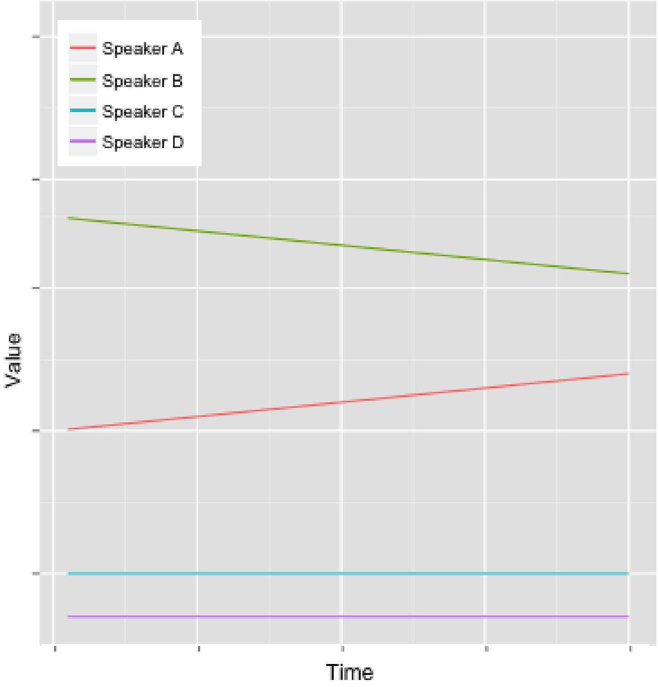


Figure 5.3: Convergence versus constant entrainment.

The answer to each of these questions may be different for each acoustic-prosodic feature, depending, among other things, on the extent to which speakers perceive their interlocutors' realization of a given feature or the extent to which they can manipulate their own realization of that feature. To answer these questions for various features of interest, we look for evidence of global and local similarity and convergence on those features in our corpus, as described in the next section.

5.1 Method

In our tests for global entrainment, we look at feature values over an entire session—that is, a complete interaction between two partners, consisting of three Cards Games and one Objects Game (see Chapter 4 for a complete description of the corpus collection). For local entrainment, we look at feature values at turn exchanges within each session—the final inter-pausal unit (IPU) from one speaker's turn and the initial IPU from the other speaker's subsequent turn. The core of our approach here is to compare the differences between partners in instances where we expect to find similarity, if entrainment exists, with baseline differences in instances where entrainment would not induce similarity. For each test, a significant result constitutes evidence of that type of entrainment for the feature in question.

Global similarity

We look for global similarity by calculating feature means over an entire session and comparing the difference between conversational partners with the differences between non-partners, all speakers in the corpus except the reference speaker, her interlocutor in this session, and her interlocutor in the other session in which she participates. For each speaker s in a session, we calculate $ENT(s,f)$ and $ENTX(s,f)$:

$$ENT(s,f) = -|s_f - s_f^i| \tag{5.1}$$

$$ENTX(s,f) = -\frac{\sum_{x=0}^{n-1} |s_f - s_f^x|}{n} \tag{5.2}$$

where s_f refers to the speaker's mean feature value for that session, s^i_f refers to her interlocutor's mean feature value for that session, n is the number of non-partners, and s^x_f is the mean feature value for one of those speakers. $ENTX(s, f)$ serves as a baseline measure of the degree of similarity we expect to see if global entrainment has no effect.

We compare $ENT(s, f)$ and $ENTX(s, f)$ with a paired t-test. If there is no effect of global entrainment, a speaker should be no more similar to her partner than to her non-partners, the speakers with whom she is never paired. If, for a given feature f, we find $ENT(s, f)$, the *partner* similarity, to be significantly greater than $ENTX(s, f)$, the non-partner similarity, we can conclude that speakers entrain globally on that feature.

An even stronger test for global entrainment is comparing a speaker to herself. Recall that eleven of the thirteen subjects who participated in the collection of the Columbia Games Corpus returned a second time for another session with a different partner.We can therefore compare the similarity between partners ($ENT(s, f)$) with $ENT_{self}(s, f)$, the similarity between speakers and themselves:

$$ENT_{self}(s, f) = -|s_f - s'_f| \tag{5.3}$$

where s'_f is speaker s's mean value for f in another session. If a speaker adheres to a consistent speaking style across conversations without aligning her speech to that of her interlocutor, $ENT_{self}(s, f)$ will be greater; if entrainment has an effect, however, we may find that $ENT(s, f)$ is greater. Again, we compare the two sets of similarities with a paired t-test.

A potential criticism of this method of finding entrainment by contrasting conversational partners with speech in separate conversations is that a greater degree of difference between non-partners may be attributed to circumstances independent of the dynamic between interlocutors: for example, when two people are having a conversation, the fact that they use similar words does not necessarily mean that they are entraining lexically; a large portion of that similarity can be attributed to the fact that they are speaking about the same things. A person having a conversation about football will have more words in common with his or her interlocutor than with someone else who is having a conversation about politics.

Similarly, external factors can affect the acoustic-prosodic features of a conversation. A simplistic example of this is ambient noise: a pair of interlocutors on a subway platform will speak more loudly than a pair talking in the back of a limo. Other factors may have a more subtle effect on the prosody of a conversation: the emotional content of the dialogue, for example, or the degree of familiarity between the interlocutors. Two speakers talking in their second language may independently speak more slowly.

The characteristics of the Columbia Games Corpus serve to address these concerns, making between-conversation comparisons relevant. All participants are native speakers of SAE; all were strangers to their interlocutors. Recordings took place in a double-walled soundproof booth, eliminating the possibility of external channel interference. Every conversation was strictly task-oriented, with participants completing identical tasks under identical conditions. These consistencies between conversations make it valid to use a cross-session comparison as a baseline for the degree of similarity we can expect between *interlocutors* if no entrainment takes place.

Global convergence

We look for global convergence by splitting a session at different points and comparing similarities before and after a split ($ENT_{first}(s, f)$ and $ENT_{last}(s, f)$). Again, if global convergence has no effect—if interlocutors do not become more globally similar over time—the similarities will be the same. If the similarities between interlocutors are smaller after a split, we can conclude that they converge.

On average, sessions in our corpus are approximately 46 minutes long. Each is composed of four games: three Cards games (approximately 8 minutes each) and one Objects game (approximately 22 minutes). We experiment with two kinds of splits, at the session level and at the game level. In addition to the question of *whether* speakers converge, we are also interested in *when* they converge.

Local similarity

For each turn in the corpus, we calculate an *adjacent* similarity ($ENT_{adj}(t, f)$) and a *non-adjacent* similarity ($ENTX_{adj}(t, f)$). The *adjacent* similarity is the similarity (negated absolute difference) between the final IPU in the turn and the initial IPU of the following turn. The *non-adjacent* similarity is between that final IPU and the first IPU in ten randomly chosen non-adjacent turns from the same session and the opposite speaker. With no effect of local entrainment, the similarity between adjacent turns should be the same as the baseline similarity between non-adjacent turns. If interlocutors entrain locally, the *adjacent* similarities should be greater.

Synchrony

An alternative view of local entrainment, proposed by [Edlund *et al.*, 2009], looks at the relationship between a pair's *relative* values: that every increase in pitch, for example, is mimicked with a corresponding pitch increase on the part of the other speaker. Following [Edlund *et al.*, 2009], we consider a significant correlation between features of adjacent IPUs at turn exchanges to constitute evidence of such synchrony. Such correlations can be positive, indicating *convergent* entrainment—that the interlocutor is adjusting her speech to become more similar to that of her interlocutor; they can also be negative, indicating *complementary* entrainment—the interlocutor is adjusting her speech in accordance with that of her interlocutor at each turn, but in the opposite direction.

Local convergence

Again following [Edlund *et al.*, 2009], we calculate local convergence as the correlation between ENT_{adj} and time: if partners converge locally, local similarities should increase over time. Like local similarity, this phenomenon can occur independently of its global counterpart, if the interlocutors' global similarities remain constant but their turn-by-turn alignment improves.

In all calculations at the local level, we exclude turns that overlap with the previous turn to avoid the possibility of cross-channel contamination.

5.2 Results

Global similarity

Entrainment is evident at the global level for intensity mean, intensity max, and speaking rate. That is, speaker averages for these features are more similar between conversational partners than between speaker pairs who were not actually speaking to each other. The same is true for all other features we examine, although the differences for these are not significant (Table 5.1).

Intensity mean and max are even more similar between conversational partners than between speakers' own speech across two occasions (Table 5.2). Interestingly, for speaking rate, speakers are more similar to themselves than they are to their interlocutors, *and* more similar to their interlocutors than they are to random other speakers. Apparently, speakers do adhere to a consistent speaking rate, but are able to modify it to some extent in order to adapt to their partner.

Pitch mean, jitter, shimmer, and NHR, meanwhile, are significantly more similar for speakers across conversations than between speakers and their interlocutors (the difference for shimmer approaches significance), and are not significantly more similar between interlocutors than between non-partners (the test for jitter approaches significance), suggesting that these dimensions of speech comprise part of an individual's consistent speaking style and are therefore less available for adaptation towards one's interlocutor.

Finally, pitch max is the only feature that does not show a consistent speaking style (it is not significantly more similar between partners and themselves as compared to their partners) *or* global entrainment (it is not more similar between partners as compared to non-partners). Pitch max is frequently associated in the literature with emotional arousal [Scherer, 1989], and it is likely to be dependent on the speaker's emotional-physiological state. Furthermore, in the ToBI system it is used as a proxy for pitch range, which can vary to convey many things such as discourse structure, conrast, or emphasis. More research is necessary to explore the factors determining a speaker's overall maximum pitch level.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-6.5	23	1.2e-06	*
Intensity max	-6.54	23	1.1e-06	*
Pitch mean	-1.89	23	0.072	
Pitch max	-1.41	23	0.17	
Jitter	-2.12	23	0.045	.
Shimmer	-1.42	23	0.17	
NHR	-1.82	23	0.083	
Speaking rate	-2.97	23	0.0069	*

Table 5.1: *T*-tests for global entrainment: Partner vs. non-partner differences.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-3.4	21	0.003	*
Intensity max	-2.1	21	0.04	.
Pitch mean	5.5	21	1.7e-05	*
Pitch max	0.2	21	N.S.	
Jitter	5.1	21	4.8e-05	*
Shimmer	2.2	21	0.04	.
NHR	2.9	21	0.009	*
Speaking rate	2.5	21	0.02	*

Table 5.2: *T*-tests for global entrainment: Partner vs. self differences.

The shaded rows indicate features for which self differences are *larger* than partner differences.

Global convergence

When conversations are split at the session level, whether after the first five minutes or after the first half, comparisons between ENT_{first} and ENT_{last} when calculated respectively before and after the split (Table 5.3) show no significant differences — that is, speakers are no more similar after the split than before (although intensity max, speaking rate and pitch mean do show tendencies). Similarly, no significant differences in similarity were found between the first five minutes and last five minutes of each session.

When we compare partner similarities before and after the midpoint of each *game*, however, speakers are significantly more similar to each other in the second half than the first for pitch max, NHR, and speaking rate (Table 5.5). For the same three features, they are more similar in the second half of each game than in the first half of the following game (pitch max: $t(71) = -2.16, p = 0.034$; NHR: $t(71) = -2.12, p = 0.037$; speaking rate: $t(71) = -2.96, p = 0.0041$). In combination with the negative results at the session level, these results suggest a “reset” effect that takes place at the start of each new game, showing that convergence may depend on contextual continuity as well as time. That is, the increased similarity does not persist across games: when the context of the conversation switches at the start of the new game, the two partners revert to the state of similarity that was in place at the start of their interaction, before they became more similar.

We have shown that when similarity between interlocutors is calculated over an entire conversation, speakers are significantly more similar to their conversational partners than to their non-partners in intensity mean, intensity max, and speaking rate. When similarity is calculated over the second half of each game alone, speakers are more similar to their partners than their non-partners for those three features as well as pitch mean, jitter, shimmer, and NHR. Again, this convergence effect is not observed at the session level.

Local similarity

Tests for entrainment at the turn level show that similarity is significantly greater between adjacent turns than between non-adjacent turns for intensity mean, intensity max, and noise-to-harmonics ratio, confirming that speakers do match their interlocutors at turn exchanges for these features. It is notable that speakers are globally similar to each other

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	1	23	0.33	
Intensity max	2.95	23	0.0071	.
Pitch mean	0.44	23	0.66	
Pitch max	-0.72	23	0.48	
Jitter	1.7	23	0.1	
Shimmer	-0.46	23	0.65	
NHR	1.01	23	0.32	
Speaking rate	2.57	23	0.017	.

Table 5.3: *T*-tests for global convergence between the first five minutes and the rest of a session.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	0.21	23	0.84	
Intensity max	3	23	0.0064	.
Pitch mean	2.3	23	0.031	.
Pitch max	1.54	23	0.14	
Jitter	0.12	23	0.9	
Shimmer	-0.56	23	0.58	
NHR	-0.36	23	0.72	
Speaking rate	-0.99	23	0.33	

Table 5.4: *T*-tests for global convergence between the first and second halves of a session.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	0.24	91	0.81	
Intensity max	0.43	91	0.67	
Pitch mean	1.04	91	0.3	
Pitch max	3.36	91	0.0012	*
Jitter	2.08	91	0.04	
Shimmer	1.23	91	0.22	
NHR	2.66	91	0.0093	*
Speaking rate	4.26	91	5e-05	*

Table 5.5: T-tests for global convergence in between the first and second half of a game.

in intensity mean and max (and in NHR, $p < 0.1$), but they are even more similar to each other at turn exchanges. In contrast, for speaking rate, speakers entrain globally — their feature means are more similar to their interlocutors' than to the mean of their non-partners — but do not dynamically match each other at turn exchanges.

The significance of the statistical tests for local similarity and synchrony may be overestimated, since the degrees of freedom relate to the number of turns rather than the number of speaker pairs. In Tables 5.7 and 5.8, we look at the evidence for local similarity and synchrony for each session. Table 5.7 shows that while adjacent similarities in intensity mean, intensity max, and NHR are more similar than non-adjacent similarities in those features for nearly every session, this difference is significant only for sessions 7, 8, 9 and 11 for intensity, and is not significant in any sessions for NHR. The true effect of exact local matching, which is significant for those features over the data as a whole, is therefore shown to be small when we look at each session individually.

Table 5.8 reveals another possible aspect of the relationship between adjacent IPUs at turn exchanges. Adjacent intensity mean values are positively and significantly correlated for six of the twelve sessions, and a subset of these sessions show positive synchrony on intensity max as well. In addition, three sessions show positive synchrony on pitch mean, pitch max, or shimmer, respectively. For most features, however, entrainment is *complementary* rather than *convergent*: speakers do indeed adjust to their interlocutors at each

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-7.56	4489	5e-14	*
Intensity max	-5.34	4489	1e-07	*
Pitch mean	-1.07	4464	0.29	
Pitch max	-1.69	4464	0.091	
Jitter	-0.93	4455	0.35	
Shimmer	-0.78	4404	0.44	
NHR	-3.05	4486	0.0023	*
Speaking rate	0.63	4533	0.53	

Table 5.6: Similarity between feature values at turn exchanges.

turn, but instead of adjusting in a manner similar to that of their interlocutor, they adjust *away* from their interlocutor. Three sessions show negative (complementary) synchrony on intensity max or pitch mean, two on pitch max, jitter, shimmer, or NHR, and one on speaking rate. For many features, some sessions show positive synchrony, while others show negative synchrony, indicating that this behavior may be dependent to some degree on speaker variation. These individual differences indicate a fruitful area of future study, since speakers clearly adjust to their partners' behavior, but do so in very different ways.

Local convergence

As Table 5.9 shows, intensity mean and max, pitch mean and max, and NHR show evidence of *local* convergence: the difference between interlocutors at turn exchanges decreases with time. The correlations between these adjacent differences and time are no higher than 0.21, indicating a moderately strong relationship that is mitigated by other factors. Local convergence is especially prevalent for intensity max and pitch mean: it is evident for those features in five or more sessions. Interestingly, most of the features that show local convergence do not show global convergence, reinforcing that entrainment at each level is a separate phenomenon.

Session	*Intensity mean*	*Intensity max*	*Pitch mean*	*Pitch max*	*Jitter*	*Shimmer*	*NHR*	*Speaking rate*
1	-0.36	-0.72	0.61	0.64	-0.87	0.56	0.18	-0.50
2	-0.74	-0.83	0.06	0.04	0.06	-1.13	-0.28	-0.99
3	-2.22	-0.82	0.57	-0.43	-0.21	-1.32	-1.97	0.18
4	-1.58	-0.58	-0.50	-0.87	0.1	0.41	-1.45	2.15
5	0.15	0.87	0.48	0.01	-0.46	1.36	-0.01	0.02
6	-0.44	-0.37	-1.62	-1.71	0.38	0.76	-0.36	-0.66
7	**-4.73**	**-5.28**	-1.76	-1.77	-1.3	-0.67	-0.95	0.40
8	**-2.64**	-1.48	-1.13	-0.93	0.36	0.18	-0.97	-0.56
9	**-6.18**	**-5.08**	-1.05	-0.33	-0.59	-1.52	-1.4	0.62
10	-0.92	-1.13	1.36	0.13	-0.31	-0.03	-0.09	0.83
11	**-3.84**	**-2.58**	-1.69	-0.63	-1.64	-0.70	-0.99	0.89
12	-1.79	-0.12	0.56	0.04	0.01	-0.25	-1.68	-0.36

Table 5.7: T statistics from paired t-tests between adjacent and non-adjacent IPUs at turn exchanges for each session.

Results in shaded cells are significant according to the FDR test with $\alpha = 0.05$.

Session	*Intensity mean*	*Intensity max*	*Pitch mean*	*Pitch max*	*Jitter*	*Shimmer*	*NHR*	*Speaking rate*
1	-0.11	-0.09	-0.12	-0.07	0.03	-0.07	-0.05	-0.16
2	0.03	0.05	**-0.27**	**-0.22**	-0.12	0.07	0.11	0.03
3	**0.14**	0.05	-0.03	0.01	**-0.20**	-0.03	**-0.16**	-0.02
4	0.00	**-0.13**	-0.05	0.07	-0.04	-0.08	0.01	**-0.20**
5	-0.04	**-0.19**	**-0.28**	-0.15	0.01	**-0.19**	-0.05	-0.01
6	0.05	0.01	0.13	**0.19**	-0.04	-0.01	0.05	0.09
7	**0.30**	**0.30**	0	0.01	0.1	**0.15**	0.04	0
8	**0.21**	0.08	0.07	0.05	-0.11	**-0.15**	**-0.15**	-0.06
9	**0.36**	**0.29**	0.06	0.03	**-0.13**	0.01	0	-0.07
10	**0.10**	0.07	**0.19**	-0.08	-0.03	0.04	0.09	-0.04
11	**0.18**	**0.13**	0.11	-0.01	0.02	0.03	0.12	0
12	-0.05	**-0.27**	**-0.53**	**-0.28**	0.05	0.08	0.11	0.04

Table 5.8: r coefficients from Pearson's correlation tests between adjacent IPUs at turn exchanges for each session.

Bolded results are significant according to the FDR test with $\alpha = 0.05$. Results in shaded cells indicate positive synchrony.

Session	*Intensity mean*	*Intensity max*	*Pitch mean*	*Pitch max*	*Jitter*	*Shimmer*	*NHR*	*Speaking rate*
1	-0.12	-0.03	-0.11	-0.06	-0.05	-0.15	-0.1	-0.01
2	0	0	**-0.15**	-0.13	-0.12	-0.09	-0.04	-0.13
3	-0.04	0.05	0.04	-0.02	-0.05	-0.03	0.01	-0.02
4	-0.06	-0.03	**-0.17**	**-0.19**	-0.02	-0.15	0.04	-0.03
5	-0.06	-0.12	-0.11	-0.13	-0.01	-0.08	-0.05	0.06
6	-0.15	**-0.21**	-0.05	-0.17	-0.08	-0.13	**-0.21**	0.05
7	-0.12	**-0.18**	**-0.16**	-0.11	0.06	-0.04	-0.06	-0.07
8	0.05	-0.1	-0.04	-0.05	0	0	0.04	0.07
9	-0.05	0.08	**-0.10**	-0.04	-0.08	-0.07	-0.07	-0.12
10	**-0.21**	**-0.16**	**-0.11**	-0.08	0.05	-0.09	**-0.13**	0.09
11	-0.08	**-0.18**	**-0.28**	**-0.21**	-0.05	-0.09	**-0.17**	-0.07
12	**-0.22**	**-0.18**	0.06	**-0.17**	0.01	-0.13	0.01	0.01

Table 5.9: r coefficients from Pearson's correlation tests between adjacent differences at turn exchanges and time for each session.

Results in shaded cells are significant according to the FDR test with $\alpha = 0.05$.

Feature	*Global similarity*	*Local similarity*	*Synchrony* (# sessions, +/–)	*Global convergence*	*Local convergence* (# sessions)
Intensity mean	✓✓	✓	6/0		2
Intensity max	✓(✓)	✓	3/3		5
Pitch mean			1/3		6
Pitch max			1/2	✓	3
Jitter			0/2		
Shimmer			1/2		
NHR			0/2	✓	3
Speaking rate	✓		0/1	✓	

Table 5.10: Summary of results on the nature of acoustic-prosodic entrainment.

5.3 Discussion

Table 5.10 summarizes our findings on how human speakers entrain to each other on the eight acoustic-prosodic features we examine. A blank cell indicates that a given feature did not show significant entrainment according to the given test; parentheses indicate that the result approaches significance; and a checkmark indicates significant entrainment. Two checkmarks in the *global similarity* column indicate that entrainment was present relative to the *self* baseline as well as the *non-partner* baseline. The numbers in the *Synchrony* and *Local convergence* columns indicate how many sessions showed significant entrainment according to each measure; in the *Synchrony* column, this number includes both complementary and convergent entrainment.

We can now return to our original questions about how people entrain:

Is entrainment global or local? With respect to *similarity*, global and local patterns of entrainment are similar: speakers entrain to each other on intensity mean and max both globally and locally; they also entrain globally on speaking rate. These results confirm that global and local entrainment can occur independently of one another: although speakers are already globally similar to each other in intensity mean and max, they nonetheless dynamically align to each other at turn exchanges throughout the interaction; conversely, interlocutor speaking rate means are similar to each other, but they are not aligned at turn exchanges.

Synchrony, however, which measures the dynamic relationship between two sets of values rather than their distance, is present and significant for all features, especially for intensity and pitch. For all features except intensity mean, speakers in some sessions exhibit *negative* synchrony, indicating *complementary* entrainment: speakers adjust their speech *away* from that of their interlocutor at each turn. This may be understood as "completing" the previous prosodic phrase: lowering one's pitch, for example, after one's interlocutor has raised it. For most features, some sessions exhibit positive synchrony, others exhibit negative synchrony, while others exhibit none at all, indicating significant speaker variation. In general, however, we can conclude that although speakers do not match each other on every feature at each turn, they do dynamically adapt to their interlocutor's behavior in a *relative* manner for nearly every feature.

These results suggest that while speakers do respond and align at each turn to their partner's behavior with respect to every feature except speaking rate, they have less ability or perhaps inclination to alter their global means for some features. This stability can be partially explained for pitch mean, jitter, shimmer and NHR by the fact that speakers are more similar to their own speech in another session than to that of their partner; these features can be said to be determined by a speaker's consistent speaking style and are therefore less accessible to accommodation. Pitch max, however, is neither more similar between speakers and themselves or between interlocutors. The primary factors determining speaker variation for this feature is an interesting question for future work. Speaking rate, meanwhile, is *globally* similar between interlocutors, but not *locally* similar: speakers fluctuate around similar means but do not match each other at turn exchanges.

Synchrony for each feature is significant but only moderate; partner feature values at turn exchanges are undoubtedly related, but the effect is clearly mitigated by other factors. One such factor may be the pragmatics of turn-ending intonation: for example, a turn ending in a rising intonation to indicate a question may be more difficult or less appropriate for the interlocutor to match than a turn ending in falling intonation. These pragmatics are likely to be particularly important in the Games Corpus, since the tasks used for eliciting speech during the collection of the corpus required much question answering.

In future work, a discussion of local entrainment would benefit from differentiation by turn type. In [Heldner *et al.*, 2010], the authors, having shown that backchannels match their preceding turn's pitch more closely than other turn types, suggest that this serves the purpose of "backgrounding" the backchannel, making it more unobtrusive. Similar predictions could be made for the degree of entrainment expected for other turn types: interruptions, for example, can be predicted to *diverge* from the preceding turn in order to make the turn *more* obtrusive, helping the speaker take the floor; divergence in this case can also serve as an assertion of dominance.

Is entrainment exact or relative? We find evidence of both exact and relative entrainment in human-human conversations. Interlocutors have similar feature means for some features, and match each other's feature values at turn exchanges for for a subset of those features, showing that they respond to the **value** of their partner's speech features. In addition, speakers entrain **relatively** on every feature, converging to or complementing each of their interlocutor's turns in a manner either similar or opposite relative to their own feature mean. Overall, relative entrainment is more prevalent — speakers respond to a greater number of features in a relative way.

Does entrainment improve over the course of a conversation? Globally, we find no evidence for convergence at the *session* level: features are no more similar between interlocutors after the first five minutes of a session, after the first half, or when comparing the first and last five minutes. At the *game* level, however, speakers converge on pitch max, NHR and speaking rate, becoming more similar after the first half of each game, and becoming less similar again at the start of the following game. These results show that convergence does occur for some features — although entrainment on others remains static throughout the interaction — but it depends on context as well as time, with the dynamic being "reset" at the start of each new game.

Our results conflict with those reported by [Natale, 1975], who found not only that interlocutors do become more similar in intensity, but that this effect is apparent over multiple interactions. The subjects in this study met once a week for three consecutive weeks, for sessions of 60 minutes each. Natale found that they became more similar in intensity over the course of a single interaction *and* over multiple interactions: the difference in mean vocal intensity was smaller in the third week than in the first. However, the analysis did not involve the same period from each session: it compared the first ten minutes of the first dialogue, the middle ten minutes of the second dialogue, and the last ten minutes of the final dialogue. It is therefore possible that the effect captured is simply the within-session convergence already observed. The fact that we observe no convergence of intensity (although, like Natale, we do show evidence of similarity) may be attributed to the difference in collection paradigms: while speakers in the Columbia Games Corpus engaged in strictly task-oriented conversation, Natale's subjects were instructed to speak freely with their partner. Furthermore, Natale's sessions are longer, and not divided into multiple contexts, as the Games Corpus sessions are.

At the game level, we look only at midpoint splits. It is possible that with further splits occurring earlier or later in the conversation, other features might show evidence of convergence. Future research should explore the possibility of modeling the dynamics of global similarity according to a greater variety of temporal splits.

Locally, the effect of convergence is more consistent: the differences between partners at turn exchanges for intensity, pitch, and NHR decrease with time. Our measurement of local convergence is over an entire session, although we did not observe global convergence at this level. It is possible that local convergence at the game level would be stronger that the moderate correlations we observe.

Intensity mean and max. Intensity is notable in our results as the speech dimension with the strongest and most consistent evidence of entrainment. Intensity mean and max, in addition to being two of only four features to exhibit global similarity *or* local similarity, are the only two features to exhibit global similarity *and* local similarity. They are the only features for which partner similarity is stronger than self similarity, indicating that entrainment plays a greater role in determining a speaker's intensity level than that speaker's usual speaking style. In addition, while all features show significant turn-by-turn synchrony, intensity mean and max show synchrony in the greatest number of sessions. It is possible that the strength and prevalence of entrainment on intensity are related to the prominence of entrainment in human perception, its ease of manipulation by the speaker, and the range of values appropriate in multiple contexts.

The diversity of the results for each feature attest to the value of a broad, multidimensional study that allows for the comparison of results across multiple features and multiple definitions of entrainment. Doing so allows us to draw conclusions about which features are or are not entrained on, and in which ways, and is one of the most significant contributions of this thesis.

Chapter 6

Entrainment on Outliers

In this chapter, we raise the question of how entrainment behavior might change in the cases of an *outlier* speech feature — that is, when a speaker's acoustic-prosodic feature deviates significantly from the norm. Outlier behavior can be expected to affect entrainment based on two influential theories about the cognitive underpinnings of entrainment. [Chartrand and Bargh, 1999] posit that humans entrain because of the *perception-behavior link*, the theory that since the processes of perception and production are linked, perceiving a behavior makes one more likely to engage in that behavior. If we assume that an outlier speech feature is more likely to be perceived, it is therefore, according to this theory, more likely to be imitated.

The same effect would be predicted according to a more social view of entrainment. Natale's *communication model* [Natale, 1975] proposes that a speaker converges to her interlocutor's behavior on the assumption that the interlocutor is using the speech behavior that he would prefer for her to use. An outlier speech feature is a stronger signal that this particular behavior is most optimal for its speaker, as in the case of an interlocutor who speaks unusually slowly (faster speech is difficult for her to understand), softly (he doesn't want to be overheard), or quickly (she is in a hurry).

In our investigation of entrainment on outliers, we consider both global and local entrainment, as discussed in Chapter 5. We predict:

Hypothesis 1 *Sessions in which one interlocutor's average value for a feature is an outlier* *will exhibit greater relative mutual* global *entrainment on that feature.*

Hypothesis 2 *Speakers will respond to turns ending with an outlier feature value with greater relative* local *entrainment on that feature.*

6.1 Global outlier entrainment

We first consider entrainment in the case of a speaker whose session mean for a given feature is an outlier. We consider speaker s to be an "outlier speaker" for feature f if his or her session mean for f is in the 10th or 90th percentile of all session means for that feature. Table 6.1 shows which speakers are outliers for each feature.

A first observation from Table 6.1 is that few of the outlier labels are grouped in row-wise pairs: that is, the partner of an outlier speaker is not usually an outlier for that feature as well. Such a pattern would be a red flag indicating that perhaps outside conditions, such as noise in the room, were likely to be the cause of the unusual feature values for a given session. Some features are grouped column-wise: of the six speakers who are outliers for intensity mean, five are also outliers for intensity max; of the six who are outliers for pitch mean, four are outliers for pitch max; of the six who are outliers for jitter, five are also outliers for shimmer.

If outlier labels were distributed randomly, we would expect each speaker to be an outlier for two features; instead, three speakers (109 in Session 6, 113 in Session 9, and 112 in Session 11) exhibit no outlier speech behavior, while Speaker 107 in Session 12 is an outlier for six features and Speaker 103 in Session 2 is an outlier for five. In a few cases, outlier behavior is consistent across a speaker's two conversations: Speaker 101 is loud; Speaker 102 speaks quickly; Speaker 103 has a rough, high-pitched voice; Speaker 106 sounds crackly and rough; and Speaker 107's voice is high-pitched. Other instances of outlier behavior are not explained by consistent speaker variation.

As in Chapter 5, we calculate ENT (Equation 5.1), the similarity between a pair of interlocutors' session averages, and a baseline non-partner similarity, which we call here $ENTX_{gender}$. While in Chapter 5 this equation measured the averaged similarity between a speaker and *all* other speakers in the corpus with whom he or she is never paired, here

Session	*Speaker*	*Intensity mean*	*Intensity max*	*Pitch mean*	*Pitch max*	*Jitter*	*Shimmer*	*NHR*	*Speaking rate*
1	101	✓						✓	
1	102	✓	✓		✓				✓
2	103		✓	✓	✓	✓	✓		
2	104	✓	✓		✓				✓
3	105								
3	106					✓	✓	✓	
4	107			✓	✓				
4	108								
5	109			✓			✓		
5	101	✓	✓						✓
6	108								✓
6	109								
7	110			✓	✓				
7	111								
8	102								✓
8	105								✓
9	113								
9	112					✓	✓	✓	
10	111							✓	
10	103			✓		✓		✓	
11	112								
11	110	✓	✓						
12	106					✓	✓	✓	
12	107	✓	✓	✓	✓	✓	✓		

Table 6.1: Global outlier labels.

we restrict the set of non-partners to speakers of the same gender as the reference speaker's partner, to control for the possibility that some features may be more similar between speakers of the same gender.

$$ENT(s,f) = -|s_f - s_f^i|$$

$$ENTX_{gender}(s,f) = -\frac{\sum_{x=0}^{n-1} |s_f - s_f^x|}{n}$$

We can consider $ENTX_{gender}$, the similarity between a speaker and the speakers of her partner's gender, the "baseline" value for the similarity between interlocutors when entrainment has no effect. $ENT - ENTX_{gender}$, the difference between the similarity existing between partners and the baseline similarity, is then a measure of how much entrainment exists relative to baseline.

We compare $ENT - ENTX_{gender}$ for "normal" versus "outlier" speakers. ENT should be smaller for outlier speakers, since their interlocutors are not likely to be similarly unusual. However, $ENTX_{gender}$ should also be lower for outlier speakers, since they are by definition divergent from the norm. It is therefore reasonable to expect $ENT - ENTX_{gender}$ to be the same for outlier speakers and normal speakers.

If $ENT - ENTX_{gender}$ is higher for outlier speakers, that means that ENT is higher than we expect, and entrainment is *greater* relative to baseline for outlier speakers. If $ENT - ENTX_{gender}$ is lower for outlier speakers, that means that ENT is lower than we expect, and interlocutors of outliers speakers entrain *less* than the interlocutors of normal speakers do (or the outlier speakers entrain less to them), even allowing for the fact that their usual values should be further apart to begin with.

Our results for global entrainment (Table 6.2) show that outlier speakers have higher relative entrainment than do normal speakers for intensity max. This means that speakers confronted with an interlocutor who diverges significantly from the norm for intensity max make a *larger* adjustment to their speech in order to converge to that interlocutor, in accordance with our hypothesis that entrainment to an outlier interlocutor should be greater. However, the other features we examine do not show this relationship, indicating that there is no straightforward relationship between the increased salience of a feature, at

Feature	t	df	p	Sig.
Intensity mean	-2.31	6.97	0.054	
Intensity max	-3.6	10.15	0.0047	*
Pitch mean	0.28	9.03	0.79	
Pitch max	1.14	6.86	0.29	
Jitter	-1.47	6.71	0.19	
Shimmer	-1.61	6.29	0.16	
NHR	-1.65	10.52	0.13	
Speaking rate	-0.42	7.26	0.69	

Table 6.2: T-tests for relative entrainment for outlier vs. normal interlocutor pairs.

least globally, and a greater degree of entrainment. In addition, because our definition of global entrainment is mutual, it is possible that the greater degree of entrainment can be attributed to adjustments made by the outlier speaker.

6.2 Local outlier entrainment

We next look at the effect of feature values that are outliers within a speaker's range. We hypothesize that turn-by-turn entrainment may be stronger after a feature value that is an outlier – unusually high or low – for the given speaker.

We test for the effect of outlier feature values on *local* entrainment by considering the final IPU of each speaker turn. We consider a feature value from such an IPU to be an outlier if it is more than one standard deviation away from the individual speaker's mean for that feature over all turn-final IPUs in the session. Table 6.3 shows the percentage of session turns that are outliers for each feature.

For each IPU, we calculate $ENT_{adj} - ENTX_{adj}$, where ENT_{adj} is the similarity between the reference turn-final IPU and its adjacent turn-intial IPU from the opposite speaker, and $ENTX_{adj}$ is the baseline similarity between the reference IPU and 50 turn-initial IPUs drawn randomly from the other speaker's turns throughout the session. The inferences to be drawn from comparing this metric between outlier and normal IPUs are analogous to

Session	*Speaker*	*Intensity mean*	*Intensity max*	*Pitch mean*	*Pitch max*	*Jitter*	*Shimmer*	*NHR*	*Syllables /sec*
1	101	0.16	0.22	0.22	0.16	0.09	0.18	0.23	0.27
1	102	0.21	0.26	0.22	0.27	0.15	0.20	0.25	0.35
2	103	0.17	0.08	0.13	0.20	0.20	0.20	0.12	0.33
2	104	0.31	0.23	0.28	0.08	0.15	0.18	0.28	0.33
3	105	0.31	0.29	0.22	0.23	0.13	0.20	0.21	0.28
3	106	0.28	0.25	0.29	0.19	0.23	0.17	0.31	0.30
4	107	0.25	0.20	0.04	0.13	0.11	0.22	0.22	0.34
4	108	0.18	0.19	0.14	0.11	0.11	0.27	0.27	0.32
5	109	0.23	0.20	0.20	0.09	0.22	0.15	0.22	0.34
5	101	0.23	0.26	0.21	0.19	0.12	0.21	0.22	0.28
6	108	0.33	0.24	0.15	0.17	0.23	0.29	0.27	0.30
6	109	0.21	0.22	0.09	0.12	0.19	0.18	0.10	0.32
7	110	0.25	0.25	0.27	0.27	0.19	0.22	0.29	0.26
7	111	0.26	0.30	0.19	0.19	0.15	0.16	0.16	0.23
8	102	0.29	0.31	0.19	0.29	0.18	0.26	0.27	0.26
8	105	0.30	0.27	0.21	0.18	0.13	0.23	0.11	0.17
9	113	0.30	0.31	0.24	0.17	0.15	0.25	0.24	0.31
9	112	0.27	0.30	0.19	0.16	0.18	0.24	0.18	0.34
10	111	0.23	0.31	0.24	0.25	0.22	0.18	0.16	0.29
10	103	0.26	0.20	0.18	0.20	0.27	0.22	0.18	0.30
11	112	0.28	0.27	0.17	0.16	0.11	0.17	0.17	0.35
11	110	0.29	0.27	0.30	0.27	0.09	0.11	0.19	0.24
12	106	0.30	0.32	0.23	0.17	0.21	0.27	0.31	0.31
12	107	0.29	0.27	0.16	0.09	0.17	0.23	0.25	0.26

Table 6.3: Local outlier percentages.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-6.99	1948.66	3.7e-12	*
Intensity max	-6.18	1939.03	7.6e-10	*
Pitch mean	0.63	1342.24	0.53	
Pitch max	-1	1216.93	0.32	
Jitter	-0.79	1075.8	0.43	
Shimmer	-0.15	1296.4	0.88	
NHR	-3.24	1710.36	0.0012	*
Speaking rate	0.87	2056.36	0.38	

Table 6.4: T-tests for relative entrainment for outlier vs. normal turn-final IPUs.

those at the global level: since both the adjacent and baseline similarities should be lower for outlier IPUs, we expect the metric to be the same for both outlier and normal IPUs. If it is in fact higher for outlier IPUs, we can conclude that speakers entrain *more* locally to outlier feature values. Unlike the global metric, $ENT_{adj} - ENTX_{adj}$ is not symmetric, and any greater similarity after outlier IPUs may be attributed unambiguously to the speaker responding to the outlier. As in Chapter 5, we restrict these comparisons to turns that do not overlap with each other, to reduce the possibility of similarity as an artifact of cross-channel contamination.

Table 6.4 shows the results of t-tests for $ENT_{adj}(f) - ENTX_{adj}(f)$ between turn exchanges whose turn-final IPUs are normal or outlier for f. Speakers are more similar to their interlocutors at the beginning of turns for which the previous turn ended with an outlier for intensity mean, intensity max, and NHR. Although speakers align to each other at turn exchanges for *all* features, they make even larger adjustments towards their partner for intensity mean, intensity max, and NHR when those features are outliers for him or her. In addition, we repeat these t-tests for $ENT_{adj}(f) - ENTX_{adj}(f)$ between turn exchanges whose turn-final IPUs are normal or outlier for any feature **except** f, and find that speakers entrain *more* on speaking rate (and intensity mean; $p < 0.05$) when the previous turn ends in an IPU that is an outlier for any other feature.

These results relate to the difference between interlocutors' raw feature values at turn

Feature	*(Outlier, outlier)*	*(Outlier, normal)*	*(Normal, outlier)*	*(Normal, normal)*
Intensity mean	381 (345.2)	798 (833.8)	931 (966.8)	2371 (2335.2)
Intensity max	368 (317.72)	782 (832.28)	870 (920.28)	2461 (2410.72)
Pitch mean	204 (190.44)	720 (733.56)	714 (727.56)	2816 (2802.44)
Pitch max	161 (160.12)	679 (679.88)	688 (688.88)	2926 (2925.12)
Jitter	161 (141.8)	594 (613.2)	674 (693.2)	3017 (2997.8)
Shimmer	229 (201.31)	647 (674.69)	781 (808.69)	2738 (2710.31)
NHR	242 (225.29)	738 (754.71)	787 (803.71)	2709 (2692.29)
Speaking rate	389 (375.99)	954 (967.01)	877 (890.01)	2302 (2288.99)

Table 6.5: Observed and expected counts for outlier and normal feature values at turn exchanges. Expected counts are in parentheses.

exchanges. It is also possible that entrainment has an effect on the *relative* expression of a turn beginning that follows an outlier turn ending. Table 6.5 shows observed counts for outlier and normal feature values at turn exchanges, as well as the counts to be expected if the outlier statuses of turn endings and beginnings are independent. Turn exchanges at which both the turn ending and beginning are outliers occur more often than expected for all features, and outliers followed by non-outliers and outliers preceded by non-outliers occur less often than expected, also for all features. Chi-squared tests (Table 6.6) show that these differences are significant for intensity mean, intensity max, and shimmer. Although speakers' raw feature values are not necessarily more similar after outliers, they are more likely to respond to an outlier with an outlier of their own for these features. Furthermore, as Table 6.7 shows, turn beginnings that are outliers in speaking rate (and intensity mean and max, $p < 0.05$) are also more likely to occur following turn endings that are outliers for any *other* feature.

6.3 Discussion

Our results support the hypothesis that entrainment is relatively greater in the presence of outlier speech behavior.

Feature	χ^2	*df*	*p*	*Sig.*
Intensity mean	6.93	1	0.0085	*
Intensity max	14.5	1	0.00014	*
Pitch mean	1.42	1	0.23	
Pitch max	0	1	0.97	
Jitter	3.66	1	0.056	
Shimmer	5.95	1	0.015	*
NHR	1.94	1	0.16	
Speaking rate	0.82	1	0.36	

Table 6.6: Chi-squared tests for the increased likelihood of an outlier turn beginning after an turn ending that is outlier for that feature.

Feature	χ^2	*df*	*p*	*Sig.*
Intensity mean	5.25	1	0.022	.
Intensity max	5.08	1	0.024	.
Pitch mean	1.23	1	0.27	
Pitch max	0.64	1	0.42	
Jitter	3.72	1	0.054	
Shimmer	1.96	1	0.16	
NHR	0.27	1	0.6	
Speaking rate	9.53	1	0.002	*

Table 6.7: Chi-squared tests for the increased likelihood of an outlier turn beginning after an turn ending that is outlier for any *other* feature.

Globally, speaker pairs of which at least one speaker's mean for **intensity max** is an outlier (in the 10th or 90th percentile) are relatively more similar to each other in intensity max than non-outlier pairs: that is, speakers confronted with an interlocutor whose volume peaks are unusually high is likely make *more* of an adjustment to converge to that feature than speakers whose interlocutor's intensity max is not an outlier. Other features, however, show no difference in relative entrainment between outlier and non-outlier pairs. Clearly, the relationship between perception and mimcry is not straightforward.

Locally, speakers entrain *more* on outlier feature values for intensity mean, intensity max, and NHR; in addition, speakers are more likely to respond to an outlier feature value with an outlier of their own for intensity mean, intensity max, and shimmer. Furthermore, speakers entrain *more* on intensity mean and speaking rate, and they are more likely to begin their turn with an outlier on intensity mean, intensity max, or speaking rate, when the previous turn ends with an outlier on any *other* feature.

The prominence of intensity in these results recalls its prominence in our tests for entrainment on general speech features; perhaps the perception-behavior link is stronger for intensity than for other features (c.f. [Chartrand and Bargh, 1999]), or perhaps it serves as a stronger communicative social signal (c.f. [Natale, 1975]).

Although our hypothesis is not supported for all features, we can conclude that a feature's perceptual prominence plays a role in degree of entrainment for intensity, voice quality and speaking rate. Pitch, however, shows no connection between outlier status and entrainment, according to any of our tests.

A close read of other studies of entrainment reveals other findings that support our hypothesis. In a study of vowel accommodation, [Babel, 2012] found that participants who accommodated the most to /a/ were those from the Upper Midwest, whose customary fronted /a/ phonemes were most distinct from the speaker's Californian /a/; the author attributed this finding to the fact that greater phonetic distance offers more acoustic-phonetic space in which to accommodate. When looking at entrainment on reference attributes, [Vullinghs *et al.*, 2013] found that speakers adapted to their partners' use of both size and color; however, the effect was stronger for size, the dispreferred attribute (35% relative difference versus 18% for color). The idea of a "dispreferred" attribute seems to be the conceptual analog of the acoustic-prosodic outliers we look at here; it is interesting to see that this effect is present at a much higher level of communication.

[Goldinger, 1998] suggests that low-frequency words are more likely to encourage imitation than high-frequency words, because there are fewer competing representations in the listener's experience. It is unclear how this idea would translate to acoustic-prosodic features, which do not exist in discrete categories, but it is possible that an outlier speech value can trigger the perception of a categorical "atypical" speech, which has fewer competitors than "typical" speech. This interpretation is supported by our finding that turn beginnings that are outliers for intensity mean, intensity max, and speaking rate are more likely to occur, and relative entrainment on intensity mean and speaking rate is greater, following turn endings that are outliers for any *other* feature.

The concept that outlier speech features may promote entrainment has useful application for a spoken dialogue system that attempts to induce a user to entrain to it, so that he or she will abandon a speech behavior that is difficult for the system to process (such as shouting or hyperarticulation [Fandrianto and Eskenazi, 2012]). The system should have a greater chance of success in inducing the user to adopt its target intensity (a quieter tone) or speaking rate (a faster rate) if the utterance it uses to prime the user is unusually soft or quick; implementing and testing this idea is a useful direction for future work. Another interesting direction is to extend the cognitive implications of this work by attempting to separate the social signal of an outlier from its perceptual salience, in order to measure the degree to which each aspect of an outlier contributes to the increased entrainment effect.

Chapter 7

Entrainment on Backchannel-Inviting Cues

In this chapter, we take a first look at entrainment in a novel domain: backchannel-inviting cues. Entrainment research thus far has focused on what is said and how it is said; to our knowledge, this is the first study to investigate entrainment in turn-taking. 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 are typically phatic expressions, serving a social purpose rather than conveying information, and usually go unacknowledged by the other speaker. In the following example from the Columbia Games Corpus, the word "okay" is a backchannel:

Speaker A: All right so I have a- a a nail on top

Speaker B: ***okay***

Speaker A: with an owl in the lower left.

In this example, "okay" is not said in response to any question or request, and Speaker A continues without a break. Approximately 11% of the 5641 turns in the Columbia Games Corpus are labeled as backchannels. (Although backchannels are not pragmatic turns, throughout the thesis we use "turn" to refer to any uninterrupted speaker utterance.) Backchannels in the Columbia Games Corpus were identified in the course of its annotation for affirmative cue words [Gravano, 2009]. Annotators were provided with the following definition:

> *Backchannel: The function of 'okay' [or 'alright', 'mm-hm', 'yeah', etc.] in response to another speaker's utterance that indicates only "I'm still here / I hear you and please continue."*

When at least two out of three annotators considered an utterance to be a backchannel, it was labeled "BC" if it did not overlap with the preceding utterance, and "BC_O" (backchannel with overlap) if it did. There are 553 speech segments labeled "BC" in the Games Corpus, and 105 labeled "BC_O." Since we are interested in the segments of speech preceding backchannels, we restrict our analysis to non-overlapping backchannels.

Backchannels play an important role in facilitating communication and are very frequent in human task-oriented dialogues. Several studies have demonstrated the existence of *backchannel-inviting cues*, ways in which a speaker may indicate that a backchannel at a certain point would be appropriate [Ward, 1996; Ward and Tsukahara, 2000; Gravano and Hirschberg, 2009]. When a user pauses, a spoken dialogue system that can identify such cues will know whether it is expected to produce a backchannel, remain silent, or take the floor. A spoken dialogue system that can produce such cues has a better chance of eliciting a backchannel from the user, enabling it to confirm that the user is keeping up when the system is conveying large amounts of information.

[Gravano and Hirschberg, 2009] identified five acoustic-prosodic cues that tend to be present before backchannels: final rising intonation, higher intensity level, higher pitch level, a change in voice quality, and longer duration of pause-free speech. These cues, and the features that model them, are listed in Table 7.1. They found that the likelihood of speech being followed by a backchannel from the other speaker increases quadratically with the number of cues. Looking at the speakers in the Games Corpus, they found that no single acoustic-prosodic cue was used by all the speakers; each appeared to use an individual combination of cues. We hypothesize that entrainment may be one of the factors influencing a speaker's backchannel-inviting behavior.

We make no claims here regarding the intentionality of producing or responding to backchannel-inviting cues, or whether they are in fact perceived by either speakers or listeners. For convenience, we will occasionally speak of an interlocutor "choosing" to use or respond to a cue; this refers to the patterns perceived in the data and not the underlying cognitive state.

Throughout this chapter, we refer to a speaker "using" a cue when speaking of the differences observed in that speaker's speech immediately before her interlocutor produces a backchannel. We can also say that a speaker "uses" a cue when referring to the differences observed in his interlocutor's speech immediately before he produces a backchannel. For consistency (if not clarity), we refer to the backchannel-inviting speaker as "using" a cue and the backchannel-producing speaker as "responding" to a cue; however, all interpretations should hold if these conventions are inverted.

In our analysis, we distinguish between two main aspects of backchannel-inviting cue behavior: cue **use** (*whether* a feature is varied before backchannels) and cue **realization** (*how* a feature is varied). We discuss entrainment on cue use in Section 7.1 and on cue realization in Section 7.2.

Part of this work was done in collaboration with Agustín Gravano of Universidad de Buenos Aires.

Cue	*Feature*
Intonation	pitch slope over the IPU-final 200 and 300 ms
Pitch	mean pitch over the final 500 and 1000 ms[1]
Intensity	mean intensity over the final 500 and 1000 ms
Duration	IPU duration in seconds and word count
Voice quality	NHR over the final 500 and 1000 ms

Table 7.1: Features modeling backchannel-preceding cues.

7.1 Entrainment on cue use

In a study of backchannel-inviting cues in the Games Corpus, [Gravano and Hirschberg, 2009] found that no single acoustic-prosodic cue was used by all speakers. They concluded

[1] Pitch features are gender-normalized as described in Chapter 4.

that significant speaker variation exists in the use of backchannel-inviting cues, with individual speakers using different combinations of cues. We hypothesize that some of this variation might be explained by entrainment: that is, that speakers may "choose" which cues to use with reference to their partner's use of backchannel-inviting cues.

The unit of analysis employed here is the *inter-pausal unit*, or IPU, defined in Chapter 4 as a pause-free segment of speech from a single speaker. Consecutive pairs of IPUs from a single speaker are termed *holds*. Following [Gravano, 2009], we contrast hold-preceding IPUs with backchannel-preceding IPUs to isolate a speaker's backchannel-inviting cues; a speaker is considered to use a certain backchannel-inviting cue if, for either of the features modeling that cue, the difference between backchannel-preceding IPUs and hold-preceding IPUs is significant (ANOVA, $p < 0.05$). Unlike that study, which pooled each speaker's data over both conversations in which he or she participated in the corpus, we consider each speaker-interlocutor combination individually. Table 7.2 shows which cues are used by each speaker in a session.

Similarity of cue combinations

We measure the similarity of two speakers' cue *combinations* by counting the number of cues they have in common over an entire conversation. The speakers in the Columbia Games Corpus each displayed 0 to 5 of the backchannel-inviting cues described in Table 7.1 ($M = 2.08, SD = 1.56$) (Table 7.2). The number of cues speaker pairs had in common ranged from 0 to 4 (out of a possible maximum of 5, $M = 1.33, SD = 1.13$). We can conclude that speakers entrain on their choice of backchannel-inviting cues if the number of cues they have in common with their partner is greater than the average number of cues they have in common with all other speakers in the corpus with whom they are never paired, according to a paired t-test.

For a more fine-grained look at entrainment on the use of backchannel-inviting cues, we can consider individual cue features instead of collapsing them into the cues they model. This approach gives us a stricter test of cue use similarity: speakers must significantly vary the same acoustic-prosodic speech characteristic *over the same interval*. From the ten cue features, speakers used as few as 0 cues or as many as 9 ($M = 3.58, SD = 2.86$). Again, we

Session	Speaker	Intonation	Pitch	Intensity	Duration	Voice quality
1	101					
1	102					
2	103				✓	
2	104		✓		✓	
3	105	✓			✓	
3	106			✓	✓	✓
4	107				✓	
4	108				✓	
5	109					
5	101					
6	108				✓	
6	109				✓	
7	110	✓	✓			
7	111		✓	✓	✓	✓
8	102	✓				
8	105	✓			✓	✓
9	113	✓	✓	✓	✓	✓
9	112	✓	✓	✓	✓	
10	111		✓	✓	✓	✓
10	103	✓		✓	✓	
11	112	✓	✓		✓	✓
11	110	✓			✓	✓
12	106	✓		✓	✓	✓
12	107				✓	

Table 7.2: Use of backchannel-inviting cues by individual session-speakers.

compare the number of cue *features* speakers have in common with their interlocutors with the average number of cue features they have in common with their non-partners.

The results of both paired t-tests are significant: on average, speakers have significantly more cues in common with their interlocutors than with other speakers in the corpus ($t(23) = 2.55, p = 0.018$), and the same is true for our stricter test of cue use similarity, considering cue features individually ($t(23) = 2.45, p = 0.022$). We can conclude that overall, speakers entrain on the combination of backchannel-inviting cues that they use.

Similarity of individual cue feature use

We can measure entrainment on the use of each *individual* backchannel-inviting cue feature as follows:

$$ENT_u(s, f) = \frac{C(U_{s,f}, U_{s^i,f})}{C(U_{s,f})} \tag{7.1}$$

where $U_{s,f}$ indicates that feature f is used as a cue by Speaker s, and s^i refers to Speaker s's interlocutor. That is, out of all the sessions in which feature f is used as a cue by at least one interlocutor, in what percentage do both interlocutors use that feature as a cue? We compare this metric with $ENTX_u$, which calculates the same metric over all permutations of non-interlocutor speaker pairs; this serves as a baseline measure.

Measurements of entrainment on cue presence for each individual cue feature are displayed in Table 7.3. Overall, the entrainment on backchannel cue presence between conversational partners is strictly greater than the degree of entrainment between non-partners. This difference is significant according to a paired t-test ($t(16.92) = 2.21, p = 0.042$).

With respect to individual cue features, the third column of Table 7.3 is most informative. $ENT_u/ENTX_u$ is an adjusted entrainment metric that normalizes the similarity between partners by the baseline similarity between all non-partners in the corpus. Using this adjusted metric, we can see that entrainment on the use of pitch as a cue is relatively high, while entrainment on the use of duration is relatively low. A particularly interesting comparison is between pairs of features that model the same cue: relative entrainment on intensity in the last **1000** ms of a backchannel-preceding IPU, for example, is nearly 60% greater than relative entrainment on intensity in the last **500** ms; similarly, the adjusted

Feature	ENT_u	$ENT_u/ENTX_u$
Pitch slope 200	0.2	1.6
Pitch slope 300	0.5	2.96
Pitch 500	0.4	4
Pitch 1000	0.25	3.67
Intensity 500	0.4	2.85
Intensity 1000	0.2	1.8
Duration	0.7	1.35
# Words	0.5	1.21
NHR 500	0.2	2.24
NHR 1000	0.14	0.9

Table 7.3: Entrainment on individual cue use.

entrainment measure for pitch slope in the last **200** ms is 85% greater than the same measure for pitch slope in the last **300** ms; and 150% greater for NHR in the last 500 ms than in the last 1000. To our knowledge, there is no cognitive literature to explain why this is so; one hypothesis would be that the cues entrained on to a higher degree are somehow more salient to the interlocutor.

Entrainment on cue use and backchannel responses

The analysis so far deals with feature variations that result in — or at least are followed by — backchannels. However, backchannel-inviting cues are highly optional: a speaker may vary a feature, and his interlocutor may or may not respond with a backchannel. Here, we look at cues that are *not* followed by a backchannel to investigate how entrainment may affect an invitee's "choice" of which cues to respond to.

[Gravano, 2009] found that the likelihood that a speaker would utter a backchannel increased quadratically with the number of backchannel-inviting cues present in the preceding IPU. We hypothesize that a speaker will be more likely to respond with a backchannel to an IPU displaying a specific cue if he or she generally uses that cue to invite backchannels as well. This hypothesis is based on the intuition that speakers may entrain on the cues that are more salient to them (c.f. [Chartrand and Bargh, 1999]), and therefore be more likely to respond to those cues.

We say that a **speaker** generally uses a cue if either of the features modeling that cue is significantly different in the speaker's IPUs that immediately precede backchannels from her interlocutor (as described above). Following [Gravano, 2009], we consider a given **IPU** to use a backchannel-inviting cue if for either of the features modeling that cue, its value for that feature is closer to the global average for IPUs before backchannels than to the global average for IPUs before holds (another IPU from the same speaker). Feature values used in this calculation were speaker-normalized using z-scores.

For each speaker in each session, we then calculate the percentage of IPUs using each cue that are in fact followed by a backchannel. For example, when Speaker 111 in Session 7 ends a turn with an intensity level in the last 500 ms that is more similar to the global backchannel-preceding mean than to the global hold-preceding mean, her interlocutor, Speaker 110, responds with a backchannel 14% of the time. When Speaker 101 in Session 5 does the same thing, his partner, Speaker 109, responds with a backchannel only 0.3% of the time. The considerable speaker variation observed here highlights the inherent optionality of backchannel-inviting cues.

We then conduct an ANOVA for each feature with these session-speaker percentages as the dependent variable and cue labels as the independent variable. The cue labels indicate whether the speaker and her interlocutor use that feature as a cue, as defined earlier; a session-speaker is labeled "both" if both the speaker and her interlocutor use that feature, "inviter" if the backchannel-inviting speaker uses the feature but her interlocutor does not, "responder" if the backchannel-producing speaker does not and her interlocutor does, and "none" if neither speaker uses the feature as a backchannel-inviting cue. Figure 7.1 shows plots of response percentages for each cue.

The ANOVA for each feature shows a significant relationship between the cue label for a session speaker and what percentage of the session speaker's IPUs displaying that cue are followed by a backchannel ($p < 0.005$ for all features). A post-hoc Tukey test reveals that for duration and voice quality, IPUs using a feature as a backchannel-inviting cue are significantly more likely to be followed by a backchannel when their speaker alone uses that feature as a cue ("inviter") than when neither the speaker nor her interlocutor uses that feature as a cue ("none"). This association between two largely orthogonal measures of backchannel-inviting cue use validates the use of the original measure for determining whether a speaker uses a cue, at least for these features.

We are particularly interested in the cases of "both," when both the speaker of the backchannel and the speaker of the backchannel-inviting IPU use a feature as a cue — that is, when they entrain on the use of that cue. The strongest test of our hypothesis that speakers are more likely to respond to cues on which they entrain fails: backchannels are no more likely to occur in "both" cases (when both speakers use the feature as a cue) than in "inviter" cases (when the backchannel-inviter alone does so), or than in "responder" cases (when the backchannel-speaker alone uses the cue). However, the comparison of "both" and "none" is significant for intonation and pitch. That is, speakers are no more likely to respond to one of those cues with a backchannel if the backchannel-inviting speaker *alone* generally uses that cue, but they are more likely to do so if they *both* use that cue — if they entrain on the use of that cue. (In addition to intonation and pitch, "both" is greater than "none" for voice quality, but this does not relate to our hypothesis, since "inviter" is greater than "none" for voice quality as well, and it is not less then "both." That is, the increased likelihood of a backchannel in response to a voice quality cue can be attributed to the responder's "learning" of the inviter's usual cues; the responder is no more likely to produce a backchannel if she entrains on it.)

Interestingly, speakers are no more likely to respond to a cue with a backchannel when they themselves use that cue than when neither speaker does so. Assuming that the cues that a speaker responds to are the ones she perceives, it appears that the processes of perceiving and producing backchannel-inviting cues are not linked in a straightforward way, as they might be for lower-level speech functions. However, the idea that entrainment is related to perception (c.f. [Chartrand and Bargh, 1999]) is supported by our results for intonation and pitch: speakers are more likely to respond to (perceive) the cues that they entrain on.

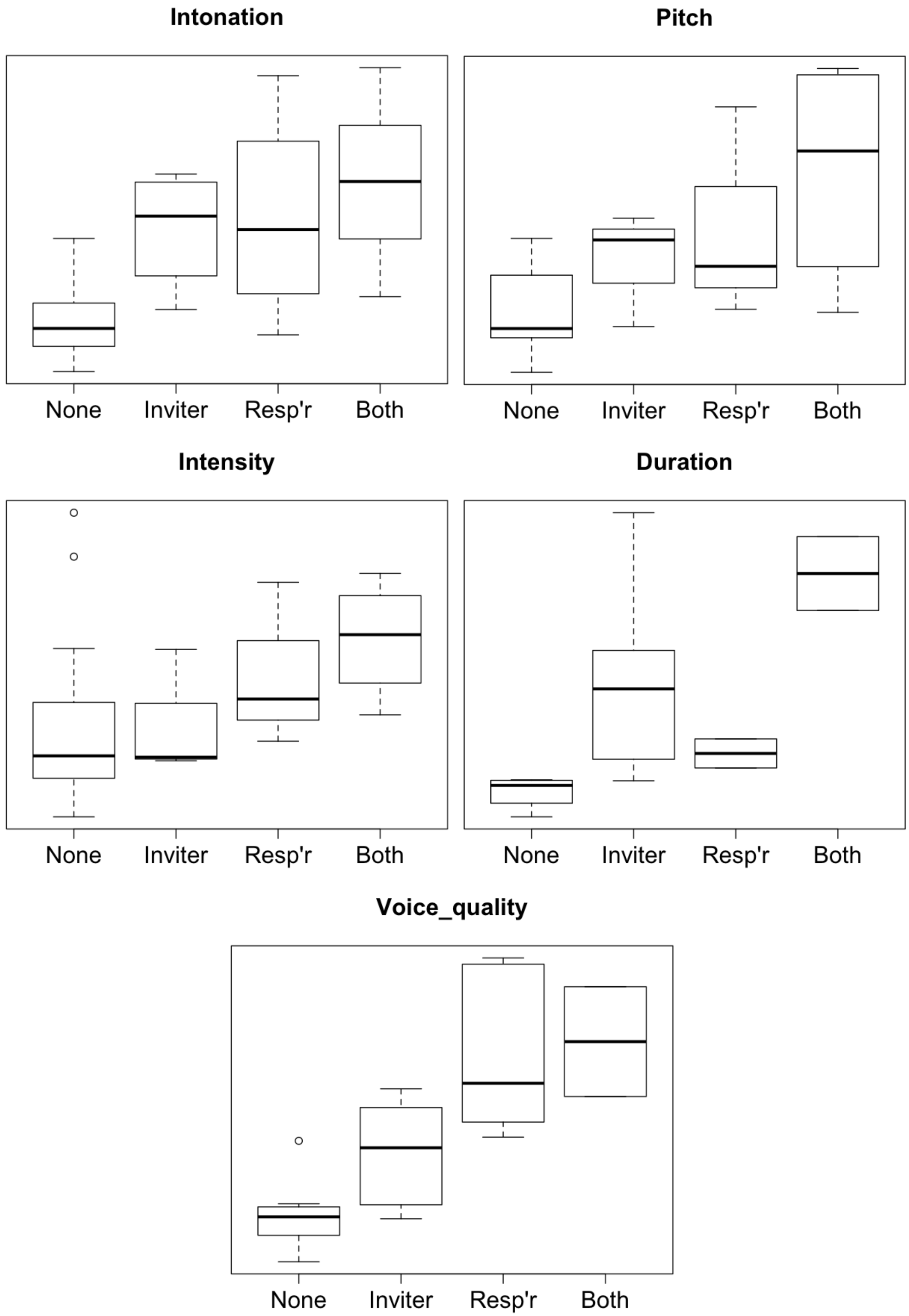


Figure 7.1: Percentages of IPUs with cue that are followed by a backchannel.

7.2 Entrainment on cue realization

Our definition for cue use requires only that a speaker consistently vary a feature value in IPUs before backchannels from his or her interlocutor; it does not address *how* the speaker varies that feature. Two speakers who both use a given cue may express it in two different ways: Speaker A may speak 50% more loudly before her interlocutor produces a backchannel, for example, while Speaker B may speak only 20% more loudly. In this section, we test the hypothesis that not only do interlocutor pairs entrain on *which* cues to use, they also entrain on *how* to realize those cues.

In order for entrainment at the global level to be defined, variation must exist in the realization of backchannel-inviting cues, as [Gravano, 2009] showed it does for the choice of backchannel-inviting cues. As Table 7.4 shows, the degree of variation in speaker-normalized feature values in backchannel-preceding IPUs is rather low for most features; only pitch mean in the last 500 and 1000 ms before a backchannel have a coefficient of variation[1] (CV) greater than 1, which is the "rule of thumb" cutoff for significantly dispersed data. Pitch mean in the last 1000 ms, in particular, has a very high degree of variation, with a CV of 8.12, and therefore has the greatest potential for partner entrainment; however, more variation is observed for all features before backchannels than overall (overall CV is displayed for comparison in the last column of Table 7.4).

We test for entrainment on backchannel-inviting cue realization by comparing $ENT_{BIC}(s,f)$ and $ENTX_{BIC}(s,f)$ for each speaker in a session.

$$ENT_{BIC}(s,f) = -|s_f - s^i_f| \tag{7.2}$$

$$ENTX_{BIC}(s,f) = -\frac{\sum_{x=0}^{n-1} |s_f - s^x_f|}{n} \tag{7.3}$$

where s_f refers to the speaker's feature value in IPUs immediately preceding backchannels in that session, s^i_f refers to her interlocutor's feature value in backchannel-preceding IPUs in that session, n is the number of speakers with whom she is never paired (usually 10), and s^x_f is the feature value in backchannel-preceding IPUs for one of those speakers.

[1] SD/M

Feature	Mean	SD	**CV**	Overall CV
Pitch slope 200	0.3867	0.3033	**0.7843**	0.6270
Pitch slope 300	0.3894	0.2781	**0.7142**	0.6962
Pitch 500	0.2165	0.4975	**2.2985**	0.1186
Pitch 1000	0.0455	0.3699	**8.1237**	0.1238
Intensity 500	0.6787	0.3667	**0.5402**	0.0674
Intensity 1000	0.6728	0.3853	**0.5726**	0.0660
Duration	0.6146	0.2873	**0.4674**	0.1936
# Words	0.5149	0.3534	**0.6864**	0.1786
NHR 500	-0.5114	0.3056	**-0.5976**	0.3173
NHR 1000	-0.4925	0.2458	**-0.4991**	0.3276

Table 7.4: Variation in realization of backchannel-inviting cues.

All feature values in these calculations have been speaker-normalized using z-scores, in accordance with our definition of a backchannel-inviting cue: how a speaker tends to *vary* her voice before a backchannel from her interlocutor. This also serves to differentiate this analysis from the overall presence of entrainment on acoustic-prosodic features that we found in Chapter 5.

As shown in Table 7.5, the results here for entrainment on the expression of backchannel-inviting cues are negative: we do not see significant evidence that speakers are more similar to each other than to their non-partners in their relative feature values before backchannels. This negative result is interesting in the light of our observation that a great deal of variation does exist in the realization of backchannel-inviting cues. This variation cannot be attributed to individual speaker differences: when comparing speaker behavior across the two sessions in which eleven of the thirteen speakers in the corpus participate, we do not find that speakers are more similar to themselves than to any of the other speakers in the corpus (according to paired t-tests between $ENTX_{BIC}$ and $ENTself_{BIC}$, which is defined analogously to Equation 5.3). Here, we have shown that acceptable realizations are not negotiated between partners. This suggests the need for future research into what factors may affect the *expression* of backchannel-inviting cues.

Feature	*t*	*df*	*p*	*Sig.*
Pitch slope 200	-1.01	23	0.32	
Pitch slope 300	-0.52	23	0.61	
Pitch 500	0.64	23	0.53	
Pitch 1000	-0.16	23	0.87	
Intensity 500	0.77	23	0.45	
Intensity 1000	0.39	23	0.70	
Duration	-0.58	23	0.57	
# Words	0.02	23	0.98	
NHR 500	0.30	23	0.76	
NHR 1000	0.28	23	0.78	

Table 7.5: Entrainment on realization of backchannel-inviting cues.

In Chapter 5, we showed that for some features, interlocutors *converge*, increase in similarity, in addition to the overall similarity observed throughout the conversation. Here, although we do not see overall similarity on expression of backchannel-inviting cues, we look at whether speakers do become more similar as the conversation progresses. We compare ENT_{BIC} in the first and second half of each session to see if differences between interlocutors are smaller in the second half. The results of paired t-tests for this comparison are shown in Table 7.6.

While most features do not show convergence, the differences between interlocutors in mean pitch and intensity in the last second of a backchannel-preceding IPU and in total duration of a backchannel-preceding IPU are significantly smaller in the second half of a conversation than in the first. Interestingly, neither mean pitch nor mean intensity show convergence in our study of general global entrainment (Chapter 5; IPU duration is not considered).

However, although speakers do become significantly more similar in their pitch, intensity and IPU duration before backchannels, they are still not significantly more similar to their interlocutors than to their non-partners in the second half of the conversation (pitch: $t(23) = -0.78, p = 0.44$; intensity: $t(23) = 0.54, p = 0.60$; duration: $t(23) = 1.66, p = 0.11$).

Feature	*t*	*df*	*p*	Sig.
Pitch slope 200	1.54	19	0.14	
Pitch slope 300	-0.61	19	0.55	
Pitch 500	1.83	19	0.083	
Pitch 1000	3.34	19	0.0035	*
Intensity 500	-0.53	19	0.6	
Intensity 1000	-2.69	19	0.014	*
Duration	-3.38	19	0.0032	*
# Words	0	19	1	
NHR 500	-2.37	19	0.029	
NHR 1000	-1.65	19	0.11	

Table 7.6: Convergence on realization of backchannel-inviting cues.

In Chapter 5, we found that speakers did not entrain on any of a standard set of acoustic-prosodic features at the session level, although they did at the game level. Here, we do find that speakers entrain on the expression of backchannel-inviting cues at the session level, suggesting that the contextual shift that occurs at the beginning of each new game does not "reset" each speaker's representation of her partner's backchannel-inviting behavior, which appears to occur for lower level features. This recalls [Brennan and Clark, 1996], who found that after completing a task with a partner and then repeating a task with the same partner, speakers reused the conceptualizations for referring expressions they had negotiated in the original conversation. Entrainment on complex, cooperative features such as backchannel-inviting cues and referring expressions may be more durable than entrainment on lower-level features because of the cognitive effort invested in producing and decoding them.

Convergence on backchannel-inviting cue expression was *not* significant at the game level, most probably because of a lack of data; many game halves contain fewer than three backchannels per interlocutor.

7.3 Discussion

Our results provide the first evidence of entrainment on turn-taking cues. Backchannel-inviting cues are particularly interesting from an entrainment perspective because unlike the standard acoustic, prosodic, phonetic, lexical and syntactic features for which entrainment is well-documented in the literature, they exist, like entrainment, only in the context of dialogue.

The hypothesis that interlocutors will entrain on backchannel-inviting cues was motivated by the variation observed in this domain, both in the use and realization of backchannel-inviting cues. We do find substantial evidence of entrainment on the *use* of backchannel-inviting cues, both individually — a speaker is more likely to use a cue if her interlocutor uses it, and in combination — a speaker is likely to have more cues in common with her interlocutor than with a random other speaker. We can conclude that part of the variation observed in the use of backchannel-inviting cues can be explained by entrainment. However, speakers do not entrain on the *realization* of backchannel-inviting cues, although they do become more similar for some features as the conversation progresses. Future research is necessary to determine what factors affect this aspect of backchannel-inviting behavior.

Our results also yield some insight into the relationship between perception and entrainment. We find that speakers are more likely to respond to intonation and pitch cues with a backchannel if they entrain on that cue — if both they and their interlocutor use that cue. This is consistent with the theory that entrainment is driven by the perception-behavior link [Chartrand and Bargh, 1999], because in this case we show that the cues that people entrain on are more likely to be perceived. This behavior is also consistent with Natale's communication model [Natale, 1975], since the speakers entrain on – adopt – the cues that the interlocutor has signaled are most intelligible to her (by responding to those cues). However, both these interpretations imply a causality that has not been proven.

This work presents an analysis of entrainment in a novel domain and suggests promising directions for future work in entrainment on other complex, pragmatic features of dialogue, such as other turn-taking cues, backchannel responses, questions, discourse markers, or repairs. Future work should also focus on incorporating these findings into a model for generating a system's backchannel-inviting cues (by modeling cues shown by the human user) or predicting whether the user is inviting a backchannel (by looking for cues shown by the system).

Chapter 8

Acoustic-Prosodic Entrainment in Mandarin Chinese

Most studies in this thesis have been carried out on the Columbia Games Corpus, a collection of task-oriented, Standard American English conversations between strangers. A critical question for assessing the impact of these studies is whether their conclusions will generalize beyond that data. While the domain of task-oriented conversations between strangers is deliberately chosen to parallel typical interactions with a spoken dialogue system, it is very desirable to be able to reason about entrainment in languages other than English.

In this chapter, we replicate our study of acoustic-prosodic entrainment (Chapter 5) on a comparable corpus in Mandarin Chinese. The comparison of entrainment in Standard American English (SAE) and Mandarin Chinese (MC) is interesting for two reasons. Firstly, the prosody of the two languages is fundamentally different: Chinese is a tonal language, with a speaker's pitch conveying lexical as well as paralinguistic information. Secondly, contemporary Chinese culture is quite different from its American counterpart, meaning that the social factors that mediate entrainment may operate in different ways for the two languages. Any consistencies we find between two such dissimilar languages can encourage us to think that they may hold true for other languages more similar to English.

In addition to the value of a cross-linguistic comparison, this consitutes (to our knowledge) a first look at entrainment of any kind in Mandarin Chinese.

This was joint work with Zhihua Xia of Jiangsu Normal University and Tongji University, who collected the Chinese corpus and collaborated equally on the analysis of the results and comparisons with SAE.

8.1 Tongji Games Corpus

The Tongji Games Corpus [Xia *et al.*, 2014] is a corpus of spontaneous, task-oriented conversations in Mandarin Chinese. The corpus contains approximately 12 hours of speech, comprising 99 conversations between 84 unique speakers (57 female, 27 male), some of whom participated in more than one conversation with a different partner. Conversations average 6 minutes in length. Participants in the corpus were randomly selected from university students who had a National Mandarin Test Certificate level 2 with a grade of A or above. This restriction enforced that the elicited speech would be standard Mandarin, with minimal effect of regional dialect. As in the collection of the Columbia Games Corpus, recordings were made in a sound-proof booth on laptops with a curtain between participants so that neither could see the other's screen and so that all communication would be verbal.

Two games were used to elicit spontaneous speech in the collection of the corpus. In the **Picture Ordering** game, one subject, the information *giver*, gave the other, the *follower*, instructions for ordering a set of 18 cards. When the task was completed, the same pair switched roles and repeated the task. In the **Picture Classifying** game, each pair worked together to classify 18 pictures into appropriate categories by discussing each picture. Seventeen pairs played the Picture Ordering game, 39 pairs played the Picture Classification game, and 14 pairs played both games (each time with the same partner).

The corpus was segmented automatically using SPPAS (SPeech Phonetization Alignment and Syllabification) [Bigi and Hirst, 2012], a tool for automatic prosody analysis. The automatic segments were manually checked and orthographically transcribed. Turns were identified by two PhD students specializing in Conversation Analysis.

8.2 Method

Our experiments on Mandarin Chinese closely parallel the line of analysis described in Chapter 5. Replicating these experiments on a parallel corpus allows us to compare and contrast patterns of entrainment behavior between the two languages. We briefly review the methods described in Chapter 5 and specify certain differences in how the analysis was conducted on the Chinese data.

8.2.1 Features and units of analysis

As in the SAE analysis, the smallest unit of analysis here is the *inter-pausal unit*, or IPU, defined as a pause-free segment of speech from a single speaker; a *turn* is then a maximal sequence of IPUs from a single speaker. While 50ms was the threshold for pause length in the Columbia Games Corpus, the threshold used here is 80ms. This number was derived empirically from the average length of stop gaps in each corpus.

For our analysis, we include one randomly chosen conversation from each of the 70 speaker pairs, for a total of 70 conversations between 84 speakers, since some speakers participated more than once with different partners.

We compare seven acoustic features: intensity min, mean, and max; pitch min, mean, and max, and speaking rate (syllables per second). All features were extracted from each IPU using Praat [Boersma and Weenink, 2012]. We compare results from the MC subjects with our previous experiments on SAE speakers, in which we looked at intensity mean and max, pitch mean and max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate. In this study, we did not examine the voice quality features, choosing to focus on the three main aspects of prosody: intensity, pitch and duration. Unlike the SAE pitch features, which were scaled so that male and female pitch values would lie within the same range, all features used in the MC calculations are raw (unscaled). Instead, we control for gender in the global calculations by restricting the non-partner baseline to include speakers of the same gender as the reference speaker's partner. Gender does not affect the local calculations, since baselines are drawn only from the reference session. Results from the two studies are therefore comparable.

8.2.2 Measuring entrainment

Similarity, synchrony, and convergence at the global and local level are calculated similarly to the methods in Chapter 5, with several refinements.

To measure **global similarity**, we compare partner differences, the absolute difference between a pair of interlocutors' feature averages over a session, with non-partner differences, the average absolute differences between each speaker's session average and the session averages of all her non-partners (see Equations 5.1 and 5.2). In the SAE study, "non-partners" included all speakers in the corpus with whom the reference speaker was never paired; here, we further restrict this group to speakers of the same gender and task role as the reference speaker's partner to control for the possible effect of increased baseline similarity between speakers of the same gender and role.

In the Tongji Games Corpus, each conversation consists of 18 sections, each of which involves the placement or classification of a single card. To measure global convergence, we compare partner differences over the first nine sections with those in the second nine. In addition, we compare partner differences in the first section with those in the last. This analysis covers only 66 conversations; four were omitted because they were missing speech from one of the interlocutors for one or more of the 18 sections. Recall that in the SAE results (Chapter 5), all evidence of global convergence was at the game level, and subjects did not globally converge over sessions; in the Tongji Games Corpus, each session consists of only one game, with no significant internal contextual shifts.

We test for **local similarity** by comparing the similarity at turn exchanges between *adjacent* and *non-adjacent* IPUs (ENT_{adj} and $ENTX_{adj}$). If, for a given feature, turn-final IPUs are more similar to the subsequent turn-initial IPUs than to ten other turn-initial IPUs chosen randomly from the same session, we conclude that speakers entrain *locally* on that feature, irrespective of their *global* similarity. For **synchrony**, we measure the Pearson's correlation between the two sets of values, as proposed by [Edlund *et al.*, 2009], to test whether interlocutors' respective acoustic features vary in synchrony even if they are not similar. As discussed in Chapter 5, local entrainment can be convergent (positive) or complementary (negative): speakers can adjust their speech *towards* that of their interlocutor, or they can respond by adjusting it in the *opposite* direction, as if to complement their interlocutor's speech.

To test for local **convergence**, an increase in similarity between interlocutors at turn exchanges over time, we correlate the difference between adjacent IPUs at turn exchanges with turn index; a negative correlation is evidence of local convergence. To reduce the degree of computation for synchrony and local convergence, we computed the correlations over only 30 conversations out of the 70 for which we examined global entrainment, randomly selecting ten conversations each from female, male, and mixed-gender pairs. Like the local calculations for the SAE conversations, the MC local calculations exclude overlapping turns to reduce the possibility of cross-channel contamination.

8.3 Results

We present the results of our tests for entrainment in Mandarin Chinese (MC), with reference to the findings of entrainment in Standard American English (SAE) in Chapter 5, followed by discussion in Section 8.4.

8.3.1 Global entrainment

Our comparison of global partner and non-partner similarities shows (Table 8.1) that MC speakers are significantly more similar to their partners than to their non-partners in intensity mean, intensity max, pitch max, and speaking rate.

Checkmarks in the final two columns indicate features that show significant global similarity in MC and SAE, respectively. It is apparent that the patterns for these features in the two languages are very similar. For all three features that show evidence of entrainment in SAE, MC speakers show evidence of entrainment as well. Unlike SAE speakers, however, MC speakers do entrain on pitch max; like SAE speakers, they do not entrain on pitch mean. The difference cannot be attributed to the fact that the SAE pitch measures are gender normalized, while the MC pitch measures are unscaled, since the MC entrainment calculation normalizes for gender by restricting the baseline comparison to speakers of the same gender as the speaker's interlocutor. The MC subjects also do not entrain on intensity min or pitch min, which the SAE study did not consider.

Feature	*t*	*df*	*p*	*MC*	*SAE*
Intensity mean	-5.05	98	0.0	✓	✓
Intensity max	-5.13	98	0.0	✓	✓
Intensity min	-1.16	98	0.25		–
Pitch mean	0.67	98	0.51		
Pitch max	-3.44	98	0.001	✓	
Pitch min	0.45	98	0.65		–
Speaking rate	-7.99	98	0.0	✓	✓

Table 8.1: *T*-tests for global similarity in Mandarin Chinese.

✓: Significant evidence of entrainment; (blank cell) : No evidence of entrainment; – : Feature not tested.

In contrast to the SAE study, however, we are unable to conclude that MC speakers globally converge on any of the features we examine here. When comparing interlocutor differences at the beginning or end of conversations, we found that intensity mean and max were significantly *less* similar in the second halves; no other significant differences were found. This is in contrast to our findings in SAE, where we found evidence of global convergence on pitch max and speaking rate, as well as NHR, which is not considered here.

8.3.2 Local entrainment

Table 8.2 shows the *t* statistics from paired *t*-tests comparing adjacent and non-adjacent differences for each of 30 MC sessions. It is apparent from the table that negative—or complementary—local entrainment is much more pronounced in the MC data than in SAE. Adjacent differences are significantly *larger* than non-adjacent differences in multiple sessions for nearly every feature. As in SAE, we observe positive, or convergent, local entrainment (indicated by negative *t*-statistics and shaded cells in Table 8.2) in several sessions for intensity mean and max. Unlike in SAE, however, multiple sessions show negative entrainment on these features as well, and nearly all sessions show strong negative entrainment on the pitch features. In addition, one session out of 30 shows negative entrainment on speaking rate; in SAE, no session shows positive or negative entrainment on speaking rate.

This pattern is repeated in the results of our tests for *synchrony*, which are more consis-

Session	*Intensity min*	*Intensity mean*	*Intensity max*	*Pitch min*	*Pitch mean*	*Pitch max*	*Speaking rate*
1	0.64	0.12	-0.45	**3.62**	**7.16**	**6.07**	2.31
2	-0.34	1.59	1.67	**9.89**	**11.22**	**2.41**	-1.21
3	0.58	-2.13	-1.84	-1.46	**2.38**	**3.97**	-0.27
4	2.36	0.96	-1.33	2.02	**7.19**	1.82	1.54
5	2.15	1.12	1.01	0.31	0.93	**2.33**	0.76
6	-1.43	-2.3	-2.11	1.19	**6.01**	0.47	0.03
7	1.12	**8.26**	**3.11**	**2.73**	1.11	0.45	-0.13
8	0.57	-2.01	-0.77	0.04	1.01	1.62	2.31
9	**3.15**	**3.38**	1.84	1.21	-0.72	-2.05	-0.71
10	-0.24	-0.7	0.34	2.15	2.16	1.79	-0.57
11	2.26	**8.30**	**8.11**	**13.83**	**20.44**	**10.44**	1.93
12	1.87	**2.66**	**2.89**	**22.66**	**46.57**	**17.78**	1.57
13	1.34	1.99	2.24	**26.40**	**42.03**	**23.04**	-0.54
14	0.87	**7.46**	**3.43**	**36.24**	**62.05**	**32.21**	1.04
15	-0.34	-1.12	-0.16	**40.96**	**40.52**	**23.37**	2.84
16	-0.54	**-2.96**	0.26	**32.20**	**24.28**	**10.28**	-0.56
17	2.27	-2	-0.73	**31.71**	**49.59**	**31.71**	0.6
18	2.67	-0.49	0.87	**20.07**	**26.36**	**8.03**	-0.45
19	**4.01**	-2.33	-0.82	**35.57**	**47.57**	**28.01**	1.28
20	1	-0.4	-0.82	**14.63**	**20.97**	**16.97**	1.45
21	1.17	**5.48**	**4.72**	**7.60**	0.75	**3.64**	0.19
22	0.19	-0.95	-1.17	**11.3**	1.97	-0.28	**4.43**
23	2.58	-0.79	-1.06	**11.18**	**17.73**	**8.52**	0.87
24	0.44	**-4.53**	**-3.49**	**4.17**	**6.08**	**2.46**	1.58
25	1.27	1.92	1.61	0.95	**3.30**	**5.33**	0.22
26	0.19	-0.53	-1.12	**9.05**	-0.11	0.55	1.6
27	0.97	-0.45	1.29	**2.68**	0.31	0.42	1.23
28	0.37	**-3.16**	**-2.57**	0.37	-0.5	-1.37	0.97
29	0.86	0.67	-0.47	**18.36**	**22.61**	**12.04**	1.22
30	1.25	1.5	1.36	**8.59**	**8.98**	2.12	-1.14

Table 8.2: T statistics from paired t-tests between adjacent and non-adjacent IPUs at turn exchanges for each session.

Bolded results are significant according to the FDR test with $\alpha = 0.05$.

Results in shaded cells are evidence of convergent entrainment.

tent with their corresponding SAE results. Six out of 30 sessions show positive synchrony for intensity mean, and three for intensity max, while five show *negative* synchrony for intensity mean, and four for intensity max. For the pitch features, negative synchrony is very strong and prevalent: 24 sessions show negative synchrony on at least one feature, and nearly a third of those correlations are stronger than -0.90. Three sessions show negative synchrony on speaking rate.

The absolute values of the MC correlations for intensity mean and max, which range from 0.29 to 0.55, are somewhat stronger than their corresponding correlations in SAE, which range from 0.10 to 0.36. Synchrony on pitch mean and max, however, is much stronger in MC: the strongest SAE correlation on pitch is -0.53, and the others are around -0.25, while in MC most correlations are stronger than -0.50, and many are even stronger than -0.90. In both languages, most positive or convergent synchrony is on intensity mean or max, and most negative or complementary synchrony is on pitch mean or max, while speaking rate displays negative synchrony in just a few sessions.

There is little evidence of local convergence on any feature in MC; one session exhibits local convergence on intensity mean, and another on pitch min and max. Another session in fact shows *divergence*, an increase in adjacent difference over time, on pitch max. This is in contrast to the SAE sessions, of which a substantial percentage exhibit local convergence on intensity and pitch features. Neither language shows any evidence of local convergence on speaking rate.

8.4 Discussion

Table 8.5 summarizes our findings on how human speakers entrain to each other in Standard American English and Mandarin Chinese. A blank cell indicates that a given feature did not show significant entrainment according to the given test; a checkmark indicates significant entrainment; a dash indicates that the relevant feature was not studied for the given language. The numbers in the Synchrony and Local convergence columns indicate how many sessions showed significant entrainment according to each measure; in the Synchrony column, this number includes both complementary and convergent entrainment.

Session	*Intensity min*	*Intensity mean*	*Intensity max*	*Pitch min*	*Pitch mean*	*Pitch max*	*Speaking rate*
1	-0.09	-0.01	0.06	**-0.31**	**-0.52**	**-0.47**	-0.18
2	0.14	-0.03	-0.1	**-0.45**	**-0.46**	-0.08	0.17
3	-0.01	**0.41**	0.32	**0.30**	-0.18	**-0.42**	-0.05
4	-0.2	0.02	0.15	-0.09	**-0.52**	-0.16	-0.07
5	-0.15	0.04	-0.13	-0.17	-0.11	-0.28	-0.03
6	0.12	**0.37**	**0.36**	-0.21	**-0.37**	0.04	-0.05
7	-0.06	**-0.35**	-0.07	**-0.27**	0.05	0.1	-0.01
8	-0.08	0.2	0.03	-0.09	-0.11	-0.21	**-0.39**
9	-0.1	-0.1	-0.04	-0.07	0.16	**0.25**	-0.04
10	0.04	0.1	0.06	-0.14	-0.25	-0.07	-0.11
11	-0.14	**-0.55**	**-0.54**	**-0.69**	**-0.77**	**-0.57**	-0.06
12	-0.04	**-0.21**	**-0.27**	**-0.79**	**-0.93**	**-0.77**	-0.06
13	-0.19	-0.15	-0.12	**-0.91**	**-0.94**	**-0.85**	0.08
14	0.04	**-0.51**	**-0.29**	**-0.94**	**-0.96**	**-0.89**	-0.15
15	0.05	0.16	0.11	**-0.92**	**-0.94**	**-0.90**	-0.14
16	0.14	**0.41**	0.22	**-0.96**	**-0.94**	**-0.77**	-0.09
17	-0.09	0.28	0.23	**-0.91**	**-0.97**	**-0.94**	-0.08
18	-0.22	0.19	0.05	**-0.81**	**-0.84**	**-0.62**	0.02
19	-0.22	**0.31**	0.24	**-0.92**	**-0.94**	**-0.81**	-0.25
20	-0.13	0	0.04	**-0.95**	**-0.95**	**-0.88**	-0.4
21	-0.1	**-0.31**	**-0.25**	**-0.43**	-0.05	-0.15	-0.05
22	-0.04	0.08	0.05	**-0.36**	-0.01	0.06	**-0.29**
23	-0.01	0.12	0.14	**-0.60**	**-0.68**	**-0.47**	-0.01
24	0.02	**0.30**	**0.26**	**-0.17**	**-0.27**	-0.08	-0.1
25	-0.05	-0.19	-0.13	-0.03	**-0.24**	**-0.35**	0.01
26	0.08	0.11	0.1	**-0.48**	0.1	0.06	**-0.30**
27	-0.2	0.15	-0.06	-0.21	0.01	-0.09	-0.22
28	-0.03	**0.29**	**0.30**	0.04	0.07	0.19	-0.13
29	-0.07	-0.07	0.02	**-0.89**	**-0.89**	**-0.76**	-0.08
30	-0.16	0.01	-0.02	**-0.51**	**-0.52**	-0.12	0.06

Table 8.3: r coefficients from Pearson's correlation tests between adjacent IPUs at turn exchanges for each session.

Bolded results are significant according to the FDR test with $\alpha = 0.05$.

Results in shaded cells are positive.

Session	*Intensity min*	*Intensity mean*	*Intensity max*	*Pitch min*	*Pitch mean*	*Pitch max*	*Speaking rate*
1	0.01	0.19	0.17	0.11	0.33	**0.35**	-0.06
2	0	-0.07	-0.01	0.23	0.07	-0.1	-0.01
3	0.03	0.14	0.13	0.13	-0.05	-0.1	0.25
4	-0.03	-0.29	-0.02	-0.04	-0.09	-0.17	0.03
5	-0.01	0.15	0.11	-0.06	0.01	-0.17	0.14
6	-0.09	0.06	-0.09	0.29	0.21	-0.03	0
7	0.15	-0.12	-0.1	-0.11	0.15	-0.09	-0.35
8	0.31	0.02	-0.12	-0.17	-0.17	-0.22	-0.05
9	-0.03	**-0.31**	-0.31	0.07	-0.12	-0.21	-0.07
10	0.08	-0.06	0	0.29	0.33	0.04	0.11
11	0.02	0.32	0.29	-0.16	-0.03	0.04	0.11
12	0.1	0.03	-0.05	0.03	-0.24	-0.18	-0.01
13	0.07	0.22	0.1	-0.02	-0.12	-0.12	0.10
14	-0.06	-0.09	-0.02	-0.05	-0.19	-0.19	0.04
15	0.1	-0.09	-0.11	0.2	0.18	0.19	-0.14
16	-0.04	-0.26	-0.31	-0.26	-0.35	-0.21	-0.38
17	-0.18	0	0.02	0.24	0.22	-0.08	0
18	0.04	0.16	0.19	0.04	-0.14	-0.25	0.23
19	-0.13	-0.06	0.06	0.07	-0.27	-0.38	-0.13
20	0.03	0.03	-0.16	0.36	0.32	-0.09	-0.27
21	-0.04	0.14	0.1	**-0.30**	-0.16	**-0.29**	0.02
22	-0.06	0.01	-0.03	0.01	0.09	0.04	0.14
23	-0.07	-0.01	0.01	-0.08	-0.15	-0.15	-0.11
24	0	0.13	0.14	0.01	0.05	0.07	0.04
25	0.12	-0.11	-0.12	-0.02	-0.1	0.03	-0.09
26	-0.09	0.01	-0.02	-0.12	-0.06	-0.07	-0.08
27	-0.17	-0.12	-0.06	-0.11	-0.24	-0.19	-0.17
28	-0.06	0.11	0.22	-0.16	-0.16	-0.18	0.04
29	-0.17	-0.22	-0.05	0.26	0.15	0.17	-0.25
30	-0.18	0.07	0.02	0.08	-0.15	-0.25	-0.04

Table 8.4: r coefficients from Pearson's correlation tests for local convergence.

Bolded results are significant according to the FDR test with $\alpha = 0.05$.

Results in shaded cells indicate convergence.

Feature	*Global similarity*		*Local similarity* (% sessions, +/-)		*Synchrony* (% sessions, +/-)		*Global convergence*		*Local convergence* (% sessions)	
	SAE	MC	SAE	MC	SAE	MC	SAE	MC	SAE	MC
Intensity min	–		–	0/7	–		–		–	
Intensity mean	✓	✓	33/0	10/20	50/0	20/17			17	3
Intensity max	✓	✓	25/0	7/17	25/25	10/13			42	
Pitch min	–		–	0/70	–	1/67			–	3
Pitch mean				0/67	8/25	0/63			50	
Pitch max		✓		0/63	8/17	3/50	✓		25	3
Jitter		–		–	0/17	–		–		–
Shimmer		–		–	8/17	–		–		–
NHR		–		–	0/17	–	✓	–	25	–
Speaking rate	✓	✓		0/3	0/8	0/10	✓			

Table 8.5: Cross-linguistic summary of results on the nature of acoustic-prosodic entrainment.

✓: significant entrainment; –: feature not tested

The truly striking finding of this study is that entrainment on pitch, intensity and speaking rate appears to be very similar in Standard American English and in Mandarin Chinese. We have presented evidence that MC speakers entrain globally on the three main aspects of prosody: duration, pitch and intensity. However, unlike SAE speakers, they show no evidence of global convergence on any feature. Locally, they entrain positively and negatively on intensity mean and max, and negatively on pitch features and speaking rate. They converge locally on intensity mean, pitch min, and pitch max. The prominence of intensity among these results — it is the only feature for which there is evidence of entrainment for global similarity and positive local entrainment in a substantial number of sessions according to both local measures — is something we observed in SAE as well.

Globally, MC and SAE speakers entrain on a similar subset of the features examined here: both sets of speakers entrain on intensity mean, intensity max, and speaking rate, and MC speakers entrain on pitch max as well. Locally, the subset of features they entrain on is similar as well: both MC and SAE speakers entrain locally on intensity mean and max; MC speakers also entrain locally on speaking rate, and SAE speakers also entrain locally on NHR. Both sets of speakers show evidence of synchrony, positive and/or negative, on almost all features examined, but the correlations are significantly stronger for MC, especially for the pitch features. The disparity between correlation strengths for pitch features in particular (moderate for SAE; very strong for MC) may reflect the additional role pitch plays in Mandarin Chinese, conveying lexical information (through tones) as well as paralinguistic information. The increased importance of pitch may increase its salience to MC speakers, which in turn may increase the likelihood that they will entrain on those features (c.f. [Chartrand and Bargh, 1999]).

Unlike SAE speakers, MC speakers show no evidence of becoming globally more similar at the game level. This is true even though we test a much wider variety of temporal splits than are considered in the SAE study. Similarly, only a couple of sessions show local convergence on intensity and pitch, while SAE speakers converge locally on intensity mean and max, pitch mean and max, and NHR, in a substantial percentage of the sessions.

The first study of entrainment in Mandarin Chinese, this work contributes a comprehensive analysis of the phenomenon. We show that speakers of Mandarin Chinese entrain globally and locally, that they synchronize their prosody to that of their interlocutor at turn exchanges in both a convergent and complementary fashion, and that they converge locally — but not globally, according to any of 18 temporal splits distributed throughout each session. They entrain on all three aspects of prosody, but most prominently and consistently on intensity features, which was true for SAE speakers as well.

In addition, this work is the first cross-linguistic comparison of entrainment patterns in multiple dimensions. This comparison allows us to find consistencies and discrepancies that may shed light on the ways in which language and social dynamics interact with entrainment. We observe consistencies between entrainment patterns in American English and Mandarin Chinese that may generalize to other languages as well; specifically, we find that for both languages, interlocutors entrain globally on intensity and speaking rate, entrain locally on intensity, entrain in synchrony for all features except speaking rate, and converge locally on pitch; for both languages, the most consistent evidence of entrainment can be found for intensity features. We find differences as well, most notably in the correlation strengths for synchrony. These patterns suggest that for some features — such as intensity and speaking rate — entrainment may be more automatic, while for others — such as pitch — it might depend more on language-specific prosody and culture-specific social processes. Future work should replicate this analysis on data from other languages in order to strengthen — or find discrepancies with — these patterns; whether entrainment patterns in other languages are consistent or inconsistent with the results of the languages we have examined, they can potentially further expose the processes underlying entrainment behavior.

Chapter 9

Entrainment and Gender

Since prior research on entrainment has sometimes observed differences in the degree of entrainment between female-female, male-male, and mixed-gender pairs, we examine our data for variation by gender pair. Previous findings are mixed with regard to the differences in entrainment behavior between male and female interlocutors, but based on the theoretical literature on entrainment, which associates females with tentative, distance-minimizing speech behavior such as entrainment, we might expect female-female pairs to entrain to a greater degree than male-male pairs and female partners in mixed-gender pairs to entrain more than their male counterparts.

We look at differences in entrainment behavior between gender pairs speaking Standard American English (SAE) or Mandarin Chinese (MC). Since the two languages are extremely different prosodically, and the two cultures have different social norms, we may hypothesize that consistencies found between the two groups will extend to other, more closely related languages.

The Mandarin Chinese portion of this work was done in collaboration with Zhihua Xia of Jiangsu Normal University and Tongji University.

9.1 Related work

An early work that inspired much research addressing gender differences in language was Lakoff's *Language and Woman's Place* (1972), which proposed that women use a distinct language of their own, characterized by the increased use of, among other forms, empty adjectives (such as “lovely” or “divine”), tag questions (“John is here, isn’t he?”), and compound requests (“Won’t you please close the door?”). Lakoff argued that “woman’s language” reflects females’ subordinate status by conveying powerlessness and need for approval.

Results of studies addressing whether women are in fact more likely to use these forms have been extremely mixed. A meta-analysis of 29 such studies [Leaper and Robnett, 2011] found that significant differences do exist, but that the overlaps between male and female frequency distributions are very large. Some psychologists have argued that “women’s language” is not necessarily specifically characteristic of women; rather, it is used whenever the speaker is in a subordinate position, regardless of gender [LaFrance, 2001]. Women’s greater use of tentative language can therefore be attributed to the fact that they are more likely to be in powerless positions. Communication Accommodation Theory [Giles *et al.*, 1991], predicts that when power imbalance exists between interlocutors, the less dominant or powerful speaker will converge more; this association has been found in several domains [Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil *et al.*, 2012]. Since entrainment is often associated with a lack of power [Giles *et al.*, 1991; Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil *et al.*, 2012], this theory would predict that women will entrain more than men, especially in mixed-gender interactions.

However, in their meta-analysis, [Leaper and Robnett, 2011] find that women are no more likely to use tentative language in mixed-gender conversations. They suggest that women’s use of tentative language may be attributed to an alternative interpretation of tentative speech: that it functions to express interpersonal sensitivity by eliciting the interlocutor’s involvement [McMillan *et al.*, 1977]. Women are socialized from an early age to be “nice,” and are penalized if they violate this norm (e.g. [LaFrance, 1992]). Many of the forms associated with women’s language have been shown to be used in this way. For example, [Brown and Levinson, 1987] reported that hedges do not necessarily convey uncertainty; they are also used to avoid shaming one’s interlocutor when expressing disagreement. Similarly, [McLemore, 1991] reported that sorority leaders used uptalk, a speaking style usually associated with powerlessness and uncertainty, to build consensus in group meetings. Again, CAT predicts that women will entrain more when speaking to both males and females, since entrainment is a tool for creating social goodwill.

Studies of “women’s language” in Chinese are rather less numerous. In one review, [Chan, 1998] cites a number of observed characteristics of female Chinese speech that are consistent with Lakoff’s original formulation. For example, girls in Beijing were observed to produce palatal phonemes [tɕ, tɕ’, ɕ] as dental sibilants [ts, ts’, s] or as more fronted palatals [Hu, 1991]. Hu explains that “fronted palatals and dentals sound more ‘fragile’...and so more ‘feminine’, while alveolars tend to be more ‘blunt’ and ‘masculine’,” and also mentions “the requirement for girls and young women to display good manners by avoiding laughing and talking with their mouths open.” In other words, the social expectation for females to appear fragile and lady-like results in a nonstandard “feminine accent.” [Farris, 1995] describes *sajiao*, a communication style marked as feminine, as “the adorable petulance of a spoiled child or young woman who seeks material or immaterial benefit from an unwilling listener.” Similarly, *niangniangqiang*, a derisive label meaning “womanish accent”, is associated with a speaking style characterized by high pitch and thin voice quality, as in whining [Shen, 1995]. Using *renjia* as a first-person pronoun is also *niangniangqiang*/“womanish”; it has been associated with society’s expectation that women be uncertain and indirect [Chao, 1995]. Finally, as proposed by Lakoff, Chinese women’s speech is characterized by the use of sentence-final particles, which have the effect of conveying “intentional politeness and nonassertiveness” [Light, 1982]. In a more recent study, a focus group of both male and female Chinese college students preferred women to use“lady talk”, described as “reserved”, “polite”, “gentle”, and “small movement of lips.” “Baby talk” was described as frequently used by young women in everyday life, albeit not preferred by young women or men [Zhang and Kramarae, 2012]. In all, these observations form a picture consistent with how English “women’s language” has been described, suggesting that gender patterns in entrainment may well be similar in both languages.

Like the rest of the literature on women’s language, empirical studies of gender differences in entrainment have mixed results. [Namy *et al.*, 2002] found that female speakers were perceived to accommodate more in a shadowing task than male speakers. [Street, 1984], on the other hand, found that males converged to a male interviewer on turn duration, while females diverged; [Pardo, 2006] found that female pairs were less phonetically similar to each other than male pairs were; and [Thomason *et al.*, 2013] found that males entrain more than females on vocal intensity features. In a study of behavioral mimcry, [Chartrand and Bargh, 1999] found no effect of gender. [Bilous and Krauss, 1988], using a within-subjects design that compared subjects' convergence behavior in mixed-gender and same-gender groups, found that females converged on some features (total words uttered, utterance length, interruptions, pauses) and diverged on others (frequency of backchannels, laughter), while males converged on utterance length, pauses, backchannels and laughter). The lack of consistency of the results within a single study, as well as across multiple studies, argues against a straightforward relationship between gender and entrainment, or suggests that some results may in fact be spurious. Since the theories of women's language relate to the social differences between men and women, differences in behavior are likely to be strongly mediated by the social context of an interaction, including each participant's age, their relative social status, the topic being discussed, the degree of familiarity between them, and the emotional register of the conversation, to list just a few factors likely to be relevant.

Like the studies reported by [Pardo, 2006] and [Thomason *et al.*, 2013], our data consists of task-oriented conversations between strangers. In spite of the associated theories, we might therefore tentatively expect to find, as they do, that entrainment is more prevalent among males.

9.2 Method

We define a partner similarity, ENT, and a non-partner similarity, $ENTX$, as in Chapter 5 (Equations 5.1 and 5.2). Recall that ENT is the negated distance between a pair of interlocutors' feature averages over a session, and $ENTX$ is the averaged negated distance between a speaker and all her *non-partners*. In Chapter 5 non-partners included all speakers in the corpus with whom the reference speaker is never paired; here, as in Chapter 8, we additionally restrict the set of non-partners to speakers of the same gender as the reference speaker's partner. This additional restriction controls for the possibility of an increased baseline similarity between speakers of the same gender and role. Controlling for baseline gender similarity is especially important in this study, since we compare entrainment behavior across gender groups; allowing the baseline similarity to include speakers whose gender is different from the partner's may artificially inflate the relative partner similarity in same-gender pairs, and artificially deprecate the same measure in mixed-gender pairs.

In Chapters 5 and 8, we found that speakers of Standard American English (SAE) were globally more similar to their partners than their non-partners in intensity mean, intensity max and speaking rate, while speakers of Mandarin Chinese (MC) were globally similar in those features as well as pitch max.

Here, we repeat our comparisons between partner and non-partner similarities, broken down by gender group (female-female, male-male, or mixed-gender). We say that a gender group shows evidence of entrainment on feature f if ENT, the partner similarities, are significantly greater than $ENTX$, the similarities between non-partners, according to a paired t-test.

To look more closely at the entrainment behavior of males and females in mixed-gender groups, we define ENT_{adj} as in Chapter 5:

$$ENT_{adj} = -\frac{\sum_i |P_{i,f} - T_{i,f}|}{|T|}$$

where T is the set of inter-pausal units, or IPUs (pause-free chunks of speech from a single speaker) that begin a speaker's turns, and P is the corresponding set of IPUs that end the interlocutor's preceding turns. This measures local entrainment, or how well interlocutors match each other at turn exchanges. It is useful here because unlike ENT, it is asymmetric, allowing us to consider each member of a dyad separately and compare the degree of entrainment between the female and male members of mixed-gender pairs.

In addition to the *prevalence* of entrainment for each gender group, we are interested in whether the *strength* of the entrainment effect varies between gender groups. We can discover this using an ANOVA with $\frac{ENT}{ENTX}$ as the dependent variable and gender group as the independent variable. Normalizing ENT by $ENTX$ allows us to compare the degree of entrainment across gender pairs, since different gender pairs will have different levels of baseline similarity between non-interlocutors.

We compare entrainment behavior between different gender groups of Standard American English and Mandarin Chinese speakers. The corpora used here, the Columbia Games

Corpus (SAE) and the Tongji Games Corpus (MC) are described in Chapters 5 (English) and 8 (Chinese). For the SAE speakers, we look at intensity mean and max, pitch mean and max, jitter, shimmer, noise-to-harmonics ratio (NHR), and syllables per second. For the MC speakers, we look at intensity min, mean and max; pitch min, mean and max; and syllables per second. The pitch tracks of female SAE speakers were scaled to lie within the same range as the male SAE speakers; the pitch tracks of all MC speakers are raw (unscaled). However, since all comparisons are conducted within gender groups, scaling should not affect the relative differences between gender groups.

Recall that the Columbia Games Corpus includes 3 female-female conversations, 3 male-male, and 6 mixed-gender; the Tongji Games Corpus includes 23 female-female, 17 male-male, and 30 mixed-gender.

9.3 Standard American English results

The results of a series of paired *t*-tests comparing partner and non-partner similarities for Standard American English speakers in female-female, male-male, and mixed-gender dyads are displayed in Tables 9.1, 9.2, and 9.3. We find that **female-female** pairs entrain on, in descending order of significance, jitter, intensity max, and intensity mean, with evidence for syllables per second, and shimmer approaching significance ($p < 0.05$). They do not entrain on pitch mean or max or NHR. **Male-male** pairs show the least prevalence of entrainment, entraining only on intensity mean and max. **Female-male** pairs entrain on, again in descending order of significance, intensity mean, intensity max, jitter, syllables per second, pitch mean, NHR, shimmer, and pitch max — in fact, on every feature we examine.

We compute ENT_{adj} for each feature for males and females of mixed-gender pairs. Contrary to the hypothesis that females will entrain more to a male interlocutor, we find no effect of partner gender. Females in mixed-gender pairs do not match their interlocutor's previous turn any more than males do. This finding contradicts the prediction made by the male dominance hypothesis.

However, entrainment is least prevalent among male pairs, as predicted by the theory that tentative speech behavior — such as entrainment — expresses interpersonal sensitivity,

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-2.93	83	0.0044	*
Intensity max	-3.22	83	0.0019	*
Pitch mean	-1.44	83	0.15	
Pitch max	-0.64	83	0.52	
Jitter	-4.18	83	7.1e-05	*
Shimmer	-2.03	83	0.045	.
NHR	-0.6	83	0.55	
Speaking rate	-2.18	83	0.032	.

Table 9.1: *T*-tests for global entrainment for SAE female-female pairs.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-4.22	83	6.3e-05	*
Intensity max	-3.66	83	0.00044	*
Pitch mean	-1.42	83	0.16	
Pitch max	-0.37	83	0.71	
Jitter	-1.17	83	0.25	
Shimmer	-1.23	83	0.22	
NHR	-0.16	83	0.87	
Speaking rate	1.21	83	0.23	

Table 9.2: *T*-tests for global entrainment for SAE male-male pairs.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-9.2	157	2.1e-16	*
Intensity max	-8.65	157	5.7e-15	*
Pitch mean	-3.15	157	0.002	*
Pitch max	-2.88	157	0.0046	*
Jitter	-5.18	157	6.8e-07	*
Shimmer	-2.92	157	0.004	*
NHR	-3.01	157	0.003	*
Speaking rate	-4.14	157	5.7e-05	*

Table 9.3: T-tests for global entrainment for SAE female-male pairs.

which is more characteristic of females. Although this would predict female pairs to exhibit the highest prevalence of entrainment, they do not show evidence of entrainment on pitch mean, pitch max, or NHR, while mixed-gender pairs entrain on every feature. In fact, although ENT, the partner similarity, is not significantly smaller for these features between female pairs than between mixed-gender pairs, $ENTX$, the baseline similarity between non-partners for these features, is significantly larger among females than between females and males, even though pitch features have been scaled so that female and male pitch values lie in the same range. Despite the sensitivity of this metric to the overall similarity among the reference group, it is valid to compare it between groups, since we are interested in the amount of *adjustment* speakers can be said to have made to each other. If all members in a group have the same pitch value, a pair within that group cannot be said to entrain to each other on pitch.

All three gender groups exhibit entrainment on intensity mean and max. We look more closely into the gender-based differences in entrainment behavior with an ANOVA with the ratio of ENT to $ENTX$ as the dependent variable and gender group as the independent variable. Normalizing ENT by $ENTX$ allows us to compare the degree of entrainment across gender groups with different baseline similarities. We find that mixed-gender pairs have the highest *degree* of entrainment, as well as the highest *prevalence*: they are more similar to each other than female pairs ($p < 0.05^1$) and than male pairs are ($p < 0.1^1$), for both intensity mean and intensity max. The differences in entrainment strength between male and female pairs are not significant.

9.4 Mandarin Chinese results

The results of t-tests comparing partner and non-partner similarities for each gender group among the MC speakers are presented in Tables 9.4, 9.5 and 9.6. **Female-female** pairs entrain globally on speaking rate, intensity mean, and intensity max; **male-male** pairs entrain only on speaking rate; and **female-male** pairs entrain on speaking rate, intensity mean, intensity max, and pitch max.

Again, we compare each group's *partner* similarities, normalized by the *non-partner* similarities to control for the overall within-group similarity, for intensity mean, intensity max, and speaking rate, the three features that show the most evidence of entrainment among all three gender groups. For MC, the differences in entrainment strengh are significant between all three groups for all three features (intensity mean: $F = 3.13, p = 0.048$; intensity max: $F = 3.73, p = 0.028$; speaking rate: $F = 5.10, p = 0.008$). A post-hoc test revealed that entrainment on intensity mean and max was weakest for male pairs, while entrainment on speaking rate was weakest for mixed gender pairs. While for SAE we concluded that entrainment is both strongest and most prevalent in mixed-gender pairs, for MC we can only conclude that it is most prevalent in mixed-gender pairs, but not necessarily strongest.

9.5 Discussion

Table 9.7, which summarizes our findings for SAE and MC, shows striking similarities in entrainment behavior patterns between speakers of the two languages. For both SAE and MC, mixed-gender dyads entrain on the greatest number of features (all for SAE; all except intensity and pitch minima for MC) and male-male pairs entrain on the fewest (only intensity mean and max for SAE; only speaking rate for MC). In addition, of the three main

[1] p values are adjusted for multiple comparisons

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-5.08	38	0.00	*
Intensity max	-4.77	38	0.00	*
Intensity min	-1.73	38	0.092	
Pitch mean	-0.52	38	0.60	
Pitch max	-0.24	38	0.81	
Pitch min	-1.28	38	0.21	
Speaking rate	-5.79	38	0.00	*

Table 9.4: *T*-tests for global entrainment for MC female-female pairs.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	0.18	19	0.87	
Intensity max	0.099	19	0.92	
Intensity min	-1.87	19	0.077	
Pitch mean	0.23	19	0.82	
Pitch max	0.60	19	0.56	
Pitch min	0.45	19	0.66	
Speaking rate	-6.80	19	0.00	*

Table 9.5: *T*-tests for global entrainment for MC male-male pairs.

Feature	*t*	*df*	*p*	*Sig.*
Intensity mean	-3.38	39	0.002	*
Intensity max	-3.23	39	0.003	*
Intensity min	0.68	39	0.50	
Pitch mean	0.97	39	0.34	
Pitch max	-4.28	39	0.00	*
Pitch min	1.03	39	0.31	
Speaking rate	-3.02	39	0.004	*

Table 9.6: *T*-tests for global entrainment for MC female-male pairs.

Feature	FF		MM		FM	
	MC	SAE	MC	SAE	MC	SAE
Intensity mean	✓	✓		✓	✓	✓
Intensity max	✓	✓		✓	✓	✓
Intensity min		–		–		–
Pitch mean					✓	✓
Pitch max					✓	✓
Pitch min		–		–		–
Jitter	–	✓	–		–	✓
Shimmer	–	(✓)	–		–	✓
NHR	–		–		–	✓
Speaking rate	✓	(✓)	✓		✓	✓

Table 9.7: Summary of global entrainment by gender group for SAE and MC speakers. ✓: Significant evidence of entrainment; (blank cell) : No evidence of entrainment; – : Feature not tested.

components of prosody — intensity, pitch, and duration — intensity and duration are the only ones on which same-gender pairs entrain: SAE and MC female-female pairs entrain on both, SAE male-male pairs entrain on intensity, and MC male-male pairs entrain on speaking rate. In both SAE and MC, mixed-gender pairs are the only ones to entrain on pitch features.

In SAE, in addition to entraining on the most features, mixed-gender pairs entrain the most on some features: for intensity mean and max, partner similarity normalized by baseline similarity is greatest for the female-male pairs. In MC, however, they are no stronger than female pairs for intensity mean and max, and weaker than both male and female pairs for speaking rate.

The greater prevalence of entrainment between mixed-gender pairs may possibly be attributed to the greater within-group speaker variation as compared to the homogenous gender groups, which results in a lower baseline similarity. However, entrainment — adjusting one's speech towards one's interlocutor — cannot be said to exist without a baseline of comparison, and the variation between non-interlocutor members of a group is as important to the measurement of entrainment as the interlocutor pair's actual observed similarity.

Our results join a very mixed collection of studies addressing "women's language," and contribute significantly to its mixedness. Unlike [Pardo, 2006] and [Thomason *et al.*, 2013], we do not find that males entrain more than females; unlike [Bilous and Krauss, 1988], we find that entrainment is more prevalent among females. Regarding [Thomason *et al.*, 2013], the only study to address entrainment on acoustic-prosodic features, we should note that their data consists of conversations between male or female humans and a prerecorded or synthesized system voice rather than between two humans; the difference in results is likely to be due to the different data collection method.

Contrary to the male dominance hypothesis, we do not find that females entrain more than males do in mixed-gender SAE dialogues. We do find some support for the theory that females are more likely to use language that reflects interpersonal sensitivity, since entrainment is more prevalent in female-female dialogues than in male-male dialogues. However, it is strongest and most prevalent in mixed-gender dialogues, which does not quite accord with that theory.

This study constitutes an experimental analysis of a hypothesized aspect of "women's language" (and a complementary "men's language"), including what is to our knowledge the first cross-linguistic comparison relating to this topic. Our results are quite consistent between SAE and MC dialogues, which suggests that they accurately capture an aspect of the relationship between gender and entrainment that is shared by both cultures. In combination, they contribute valuable empirical perspective to the theories relating to gender, tentative speech, and entrainment. Future work can extend this analysis by exploring how gender groups entrain differently on other aspects of entrainment, such as local similarity, synchrony, and global and local convergence, and by replicating these experiments on data from additional languages to enrich the cross-linguistic comparisons.

Chapter 10

Entrainment and Social Behavior

One of the most interesting aspects of entrainment is its association in many dimensions with multiple aspects of dialogue quality. For example, entrainment on high-frequency words has been associated with smoother dialogue flow and increased task score [Nenkova *et al.*, 2008]; entrainment on syntactic constructions was predictive of score on the Map Task [Reitter and Moore, 2007]; entrainment on pitch features (but not intensity features) was predictive of the degree of positivity of interactions between married couples in therapy [Lee *et al.*, 2010]; and confederates who mimicked a subject's mannerisms and posture were seen as more likable [Chartrand and Bargh, 1999].

These studies have been motivated by theoretical models such as Giles' Communication Accommodation Theory [Giles *et al.*, 1991], which proposes that speakers promote social closeness 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.

Motivated by this body of theoretical proposals and empirical findings, we hypothesize that entrainment on acoustic-prosodic dimensions such as pitch, intensity, voice quality and speaking rate might also be correlated with positively perceived social behaviors and other characteristics of efficient, well-coordinated conversations. In this chapter we describe a series of experiments investigating the relationship between entrainment on a variety of acoustic-prosodic features and a manually-annotated set of social variables designed to capture important aspects of conversational partners' perceived social behaviors. Since prior research has sometimes observed differences in entrainment behavior between genders [Bilous and Krauss, 1988; Pardo, 2006; Namy *et al.*, 2002], a finding we explored in Chapter 9, and since it is reasonable to assume that the differing social dynamics between various speaker-gender combinations should affect the interaction of entrainment and social behavior, we examine our data for variation by gender pair, considering female-female, male-male, and female-male speaker pairs separately.

This chapter addresses two main research questions:

- *What kinds of entrainment are most important in human conversations?* Humans entrain on numerous speech features. Furthermore, as discussed in Chapter 5, entrainment has many aspects. Which ones are the most valuable for a system to model? We can begin to answer this question by looking at which entrainment measures are correlated most strongly with desirable dialogue characteristics.

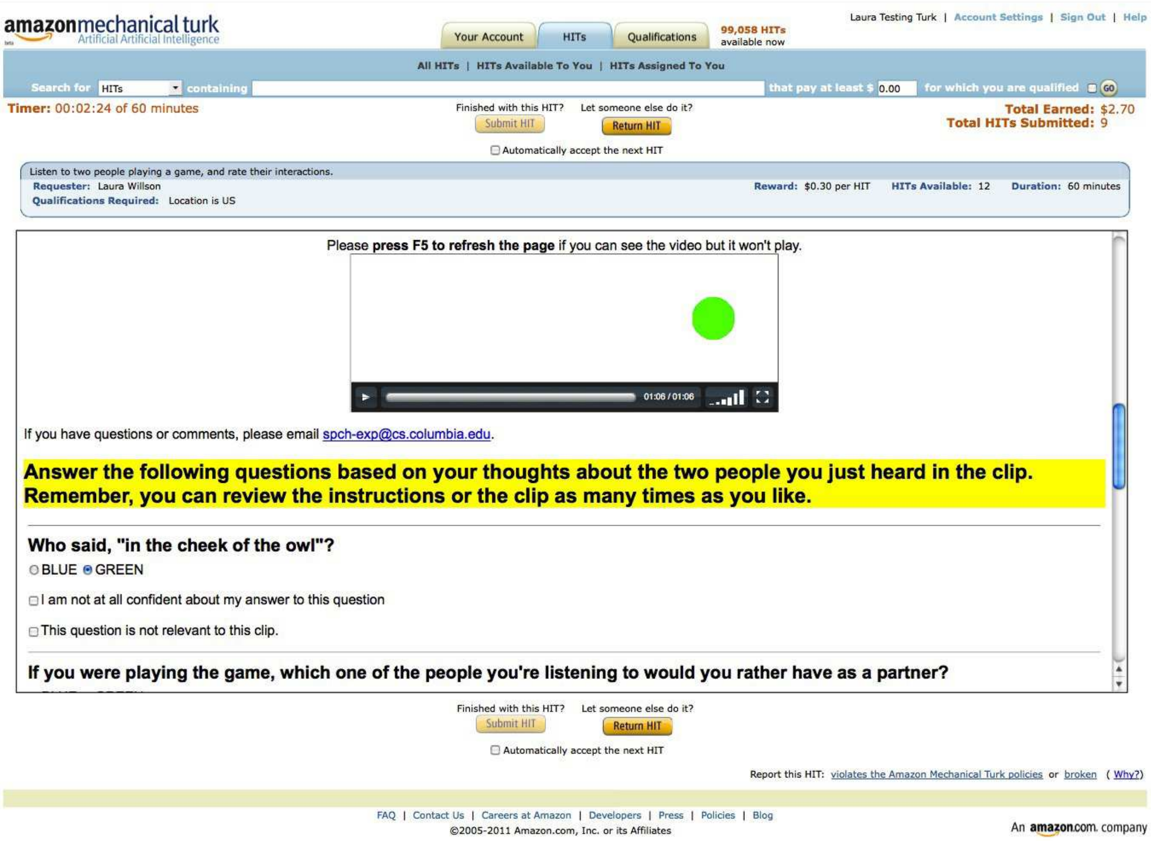
- *Can entrainment be used as a feature for evaluating conversations or interlocutor behavior?* This question, which we address only tangentially, is complementary to the first. When we know what dialogue or speaker characteristics are associated with each kind of entrainment, it may be possible to make predictions about a new dialogue based on its entrainment measures, as in [Reitter and Moore, 2007; Lee *et al.*, 2010]. While we do not directly address this possibility, the findings in this chapter may be helpful in the area of automatic dialogue assessment.

This work was done in collaboration with Agustín Gravano, Laura Willson, Štefan Beňuš, and Ani Nenkova.

10.1 Annotation of Social Variables

In order to study how entrainment in various dimensions is correlated with the perceived social behavior of the interlocustors, we asked Amazon Mechanical Turk (AMT)[1] annotators

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


amazonmechanical turk
Artificial Artificial Intelligence
Your Account
HITs
Qualifications
Laura Testing Turk | Account Settings | Sign Out | Help
All HITs | HITs Available To You | HITs Assigned To You
Search for
containing
that pay at least $
for which you are qualified
GO
Submit HIT
Return HIT
FAQ | Contact Us | Careers at Amazon | Developers | Press | Policies | Blog
An amazon.com. company

99,058 HITs available now

Search for HITs containing that pay at least $ 0.00 for which you are qualified □

Timer: 00:02:24 of 60 minutes

Finished with this HIT? Let someone else do it?

Total Earned: $2.70
Total HITs Submitted: 9

□ Automatically accept the next HIT

Listen to two people playing a game, and rate their interactions.
Requester: Laura Willson
Qualifications Required: Location is US
Reward: $0.30 per HIT HITs Available: 12 Duration: 60 minutes

Please **press F5 to refresh the page** if you can see the video but it won't play.

01:06 / 01:06

If you have questions or comments, please email spch-exp@cs.columbia.edu.

**Answer the following questions based on your thoughts about the two people you just heard in the clip.
Remember, you can review the instructions or the clip as many times as you like.**

Who said, "in the cheek of the owl"?
○ BLUE ● GREEN

□ I am not at all confident about my answer to this question

□ This question is not relevant to this clip.

If you were playing the game, which one of the people you're listening to would you rather have as a partner?

Finished with this HIT? Let someone else do it?

□ Automatically accept the next HIT

Report this HIT: violates the Amazon Mechanical Turk policies or broken (Why?)

©2005-2011 Amazon.com, Inc. or its Affiliates

Figure 10.1: Screenshot from a sample HIT.

to label the 168 Objects game tasks in the Columbia Games Corpus for an array of social behaviors perceived for each of the speakers, which we term here "social variables."

Each Human Intelligence Task (HIT) presented to the AMT workers for annotation consisted of a single Objects game task. (Each Objects game consisted of 14 tasks, each of which involved placing a single object; please refer to Chapter 4 for a more detailed description of the Objects game.) To be eligible for this annotation work, annotators had to have a 95% success rate on prevous AMT HITs and to be located in the United States. They also had to complete a survey establishing that they were native American English speakers with no hearing impairments. The annotators were paid $0.30 for each HIT they completed. Over half the annotators completed fewer than five HITS, and only four completed more than twenty. Figure 10.1 shows a screenshot from a sample HIT.

The AMT workers ("Turkers") listened to an audio clip of the task, which was accompanied by an animation that displayed a blue square or a green circle depending on which speaker was currently talking. The animation made it possible for the listener to differentiate between speakers and associate each one with a visual mnemonic; to further aid differentiation, each speaker's channel was played in a different ear. (Turkers were instructed to wear headsets, but this was impossible to enforce.) The audio clip was available for replay throughout the task. Annotators were then asked to answer a series of questions about each speaker:

- Does Person A/B believe s/he is better than his/her partner?
- Does Person A/B make it difficult for his/her partner to speak?
- Does Person A/B seem engaged in the game?
- Does Person A/B seem to dislike his/her partner?
- Is Person A/B bored with the game?
- Is Person A/B directing the conversation?
- Is Person A/B frustrated with his/her partner?
- Is Person A/B encouraging his/her partner?
- Is Person A/B making him/herself clear?
- Is Person A/B planning what s/he is going to say?
- Is Person A/B polite?
- Is Person A/B trying to be liked?
- Is Person A/B trying to dominate the conversation?

Part of each question was displayed in the color (blue or green) that was associated with the reference speaker in the animation that accompanied the task audio.

The questionnaire also included items regarding the dialogue as a whole:

- Does the dialogue flow naturally?
- Are the participants having trouble understanding each other?
- Which person do you like more?
- Who would you rather have as a partner?

A few questions with objective answers (e.g. "Which speaker is the Describer?") were included among the target questions to ensure that the annotators had paid attention to the audio and questions. HITs for which the annotator failed to answer the check questions correctly were disqualified.

Each task was rated by five unique annotators who answered "yes" or "no" to each question, yielding a score ranging from 0 to 5 for each social variable, representing the number of annotators who answered "yes." We chose to retain this information instead of enforcing annotator agreement because the questions addressed subjective features. The resulting scores are measures of the degree to which a given speaker was *perceived* as exhibiting a particular behavior.

The annotated social variables and their association with acoustic-prosodic speech features are more extensively described in [Gravano *et al.*, 2011].

10.2 Entrainment and social behavior

In this study, we focus our analysis on annotations of four social variables:

- Is the speaker trying to be liked?
- Is the speaker trying to dominate the conversation?
- Is the speaker giving encouragement to his/her partner?
- Is the conversation awkward?

These social variables were chosen for analysis according to two criteria. Firstly, more than half of the annotated tasks had an inter-annotator agreement for the reference variable of at least 4. This eliminated the variables for which the majority of task scores were 2 or

3, which would indicate that the annotators had difficulty understanding or identifying the target behavior. Secondly, task scores for the reference variable were sufficiently dispersed to allow for meaningful correlational analysis. This eliminated variables for which most tasks had the same answer, such as “Does the speaker believe that s/he is better than his/her partner?” (no) or “Is the speaker polite?” (yes).

We correlate task scores for these variables with measures of entrainment on eight acoustic-prosodic features. Based on Communication Accommodation Theory and the finding in [Natale, 1975] that entrainment on vocal intensity is correlated with social desirability, we can expect **trying to be liked**, a social behavior aimed at *decreasing* social distance, to be positively associated with entrainment; we expect **giving encouragement** to correlate with entrainment for the same reason.

Trying to dominate can be viewed as a behavior whose goal is to *increase* social distance between oneself and one’s inferior, and might therefore be expected to correlate negatively with with entrainment. On the other hand, convergence may be a tool for asserting dominance in a situation of *overaccommodation* [Turner and West, 2004]. Such convergence places the interlocutor in a lower-status role by making him or her appear dependent on the speaker. In a small-scale study of the Watergate transcripts [Niederhoffer and Pennebaker, 2002], and in a study of Supreme Court oral arguments and discussions among Wikipedia editors [Danescu-Niculescu-Mizil *et al.*, 2012], speakers in a position of dependence converged more on linguistic style towards the individuals who were in a position of power over them.

The perception of **conversation awkward** may stem from multiple sources. At its core, it can be said to imply a failure of the interlocutors to come together, from lack of either ability or interest. Again, in this scenario, the interlocutors are increasing social distance, or at least failing to decrease it, which should imply a negative correlation with entrainment.

Results of the correlations between ENT (Equation 5.1) and each of the social variables, broken down by gender group, are displayed in Table 10.1. Only significant[2] results are

[2] All significance tests correct for family-wise Type I error by controlling the false discovery rate (FDR) at 0.05 (described in Chapter 4).

Group	*Social*	*ENT*	*r*	*p*
FM	Trying to be liked	Pitch mean	0.26	0.001
		Pitch max	0.27	7e-04
	Gives encouragement	Intensity mean	0.36	2.8e-06
		Intensity max	0.31	7.7e-05
		Pitch mean	0.23	0.003
MM	Gives encouragement	Intensity mean	0.39	3e-04

Table 10.1: Correlations between entrainment and social variables.

shown.

For mixed-gender pairs, *trying to be liked* is positively correlated with the degree of similarity between partners in pitch mean and max. This result recalls the finding [Lee *et al.*, 2010] that the positivity of therapy interactions between troubled married couples was correlated with measures of entrainment on pitch; no such relationship was found for entrainment on intensity. The reason for this association is unknown, but grounds for speculation may be found in a study on flirting [Ranganath *et al.*, 2009], which found that both men and women who were flirting used higher pitch. Perhaps listeners rating a mixed-gender conversation in which each speaker's pitch was high perceived those speakers as flirting, and therefore trying to be liked.

Gives encouragement is correlated with entrainment on intensity mean, intensity max, and pitch mean for mixed-gender pairs, and with entrainment on intensity mean for male pairs. The association with entrainment on intensity features, which is not present for *trying to be liked*, shows that conclusions about associations between a specific social behavior and entrainment on a specific feature will not necessarily generalize to other, similar social behaviors or entrainment on other features. Furthermore, the relationship between entrainment and social behavior is dependent on the interlocutors' respective genders, indicating that it is mediated by the social context of the interaction.

Overall, the prediction that social variables associated with decreasing social distance would be positively correlated with entrainment is supported by our results. This association is found only with entrainment on measures of pitch and intensity. The strongest evidence of such a relationship is found for mixed gender pairs, which show multiple correlations with both *trying to be liked* and *gives encouragement*; male pairs show only a correlation between *gives encouragement* and entrainment on intensity mean; while female pairs show no correlation between social variables and entrainment. Entrainment on speaking rate or voice quality is not correlated with any social variable.

Surprisingly, *trying to dominate* and *conversation awkward*, which seem to be clear examples of increasing social distance, show no correlation with entrainment. It is possible that the raters' characterization of *trying to dominate* was confused by the opposing concepts of **trying** to dominate, which implies a position of weakness, and actually behaving in a dominant way, which implies power. This may explain the discrepancy between our results and prior work [Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil *et al.*, 2012], which used hierarchical roles as an indication of dominance. In the next section, we look at the relationship between entrainment and automatically derived measures of dialogue coordination, which may reveal aspects of *conversation awkward* that were not perceived by human raters.

10.3 Entrainment and dialogue coordination

We now examine the relationship between acoustic-prosodic entrainment and social behavior according to four measures of dialogue coordination that can be automatically derived from the turn annotations and timing information in the corpus: the number of turns in a task, the mean latency between turns, the percentage of turns that are interruptions, and the percentage of turns that are overlaps. Together, these features represent the "flow" of a conversation, the degree to which the give-and-take is smooth and well coordinated, and can be interpreted as a measure of rapport between the interlocutors.

The **number of turns** in a task can be interpreted positively or negatively. Most dialogue systems try to minimize the number of turns in an interaction because a high number is a sign of an inefficient dialogue, one which takes many turn exchanges to accomplish the objective. However, it may also be a sign of easy, flowing dialogue between the partners. In our domain, it may also be the sign of a high-achieving pair who are placing the object meticulously in order to secure every single point; recall that in the Objects Game, the partners cooperated to place a target object in an exact location, and points were awarded based on how well the location they chose matched the actual coordinates, so a high number of turns reflects a high degree of cooperation. The expected relationship between number of turns and entrainment is therefore unclear.

Latency is the length of the period of silence between the end of one speaker's turn and the beginning of the next. Latency can have a value of zero, in cases where the second speaker begins speaking immediately after the first falls silent, or even a negative value, in cases where the second speaker begins speaking before the end of the first speaker's turn. High latency, or long pauses between turns, can be considered a sign of an awkward and unsuccessful conversation, with badly coordinated turn-taking behavior indicating a possible lack of rapport and difficulty in communication between the partners, and can therefore be expected to correlate negatively with entrainment. Previously, [Nenkova *et al.*, 2008] found that mean latency in the Columbia Games Corpus correlated negatively with entrainment on high frequency words.

Interruptions, another example of poor turn-taking behavior, are defined in the guidelines for the turn-taking annotation of the Games Corpus [Gravano, 2009] as cases in which a speaker breaks in, leaving his or her interlocutor's turn incomplete. They are distinct from **overlaps**, which are cases in which the beginning of a speaker's turn overlaps with the completion of his or her interlocutor's turn, in that in the case of overlaps the first speaker's utterance is complete[3]. A high percentage of interruptions may be a symptom or cause of hostility or awkwardness between partners; however, in our task-oriented domain, interruptions may in fact be collaborative completions, cases in which the second speaker "helps" the first by completing his or her turn. A high percentage of overlaps may be a sign of a well-coordinated conversation that is flowing easily. Overlaps require the successful reading of turn-taking cues and by definition preclude awkward pauses. [Nenkova *et al.*, 2008] found that the proportion of interruptions in a session was *negatively* correlated

[3] In the annotation of the Games Corpus, Beattie's informal definition of completeness was used: "Completeness was judged intuitively, taking into account the intonation, syntax, and meaning of the utterance." [Beattie, 1982]

with entrainment on high frequency words, while the proportion of overlaps was *positively* correlated.

Table 10.2 shows significant correlations between entrainment and coordination measures. It is immediately noticeable that mixed-gender pairs, for whom social variables were highly associated with entrainment, show the least evidence of a relationship between entrainment and dialogue coordination, while female pairs, who showed no association between social variables and entrainment, show correlations of entrainment with every coordination measure.

In general, the correlations are consistent with our predictions. The *number of turns* in a task is positively correlated with entrainment for all three gender groups. *Latency* is negatively correlated with entrainment for male-male and female-female pairs. These two variables are most consistently associated with entrainment, displaying correlations with entrainment on multiple acoustic-prosodic variables for both male-male and female-female pairs. In addition, the percentage of *overlaps* is positively correlated with entrainment for all three gender groups, and the percentage of *interruptions*, contrary to our predictions, is positively correlated with entrainment for female-female and male-male groups.

The majority of these findings accord with [Nenkova *et al.*, 2008], which showed that entrainment on high-frequency words was associated with overlaps and negatively associated with mean latency; however, they also found negative associations with interruptions, which we predicted but do not find here. The positive correlations between interruptions and entrainment on shimmer (for female-female pairs) and NHR (for male-male pairs) is difficult to interpret in the framework of Communication Accommodation Theory. One possible explanation may be that in the context of a dialogue between strangers, for which politeness is required (nearly every speaker in the corpus was described as "polite" by the annotators), interrupters may mitigate the rudeness of their behavior by minimizing social distance in other ways.

For female-female pairs, the features for which entrainment is associated with dialogue coordination are intensity max (# turns and latency), speaking rate (# turns and latency), and shimmer (all four coordination measures). The correlations between entrainment on shimmer and every coordination measure are intriguing; it is not correlated with any coor-

Group	*Coordination*	*ENT*	*r*	*p*
FF	# Turns	Intensity max	0.3	0.006
		Shimmer	0.34	0.002
		Speaking rate	0.28	0.01
	Mean latency	Intensity max	-0.31	0.005
		Jitter	-0.29	0.007
		Shimmer	-0.33	0.002
		Speaking rate	-0.39	2e-04
	% Interruptions	Shimmer	0.33	0.003
	% Overlaps	Shimmer	0.3	0.005
FM	# Turns	Intensity mean	0.24	0.003
	% Overlaps	Shimmer	0.26	0.001
MM	# Turns	Intensity mean	0.39	2e-04
		Pitch mean	0.32	0.003
		Pitch max	0.29	0.007
		NHR	0.47	7.9e-06
	Mean latency	Intensity mean	-0.57	8.8e-08
		Intensity max	-0.43	1e-04
		Pitch mean	-0.52	2.4e-06
		Pitch max	-0.61	5.7e-09
		Jitter	-0.65	4.5e-10
		NHR	-0.4	4e-04
	% Interruptions	NHR	0.33	0.002
	% Overlaps	Intensity mean	0.37	5e-04
		Intensity max	0.39	2e-04

Table 10.2: Correlations between entrainment and measures of dialogue coordination.

dination measure for male-male pairs. Speech with vocal fry, a register frequently used by young female speakers of SAE [Wolk *et al.*, 2012], has been shown to have significantly higher shimmer [Blomgren *et al.*, 1998] and is considered, among other things, non-aggressive [Yuasa, 2010]. Pairs of females using vocal fry may be cooperating and communicating particularly well.

Speaking rate is another feature for which entrainment is associated with dialogue coordination for female-female pairs and not male-male pairs. Male-male pairs, in contrast, show correlations for NHR and pitch features, which do not show up among female-female and mixed-gender pairs.

Global entrainment on intensity (mean or max) is correlated with most coordination measures for every gender group. This is consistent with our finding that intensity shows the strongest and most consistent evidence of global entrainment: it is the only speech dimension to be more similar between partners than between speakers and themselves and to show evidence of entrainment between pairs of all three gender groups. Without assigning undue intentionality to entrainment, it is reasonable to hypothesize that speakers entrain on intensity *because* it is so important to dialogue coordination.

10.4 Discussion

In general, the hypothesis that acoustic-prosodic entrainment will be correlated with pro-social behavior and dialogue coordination is upheld by our findings. For mixed-gender and male pairs, *trying to be liked* and *gives encouragement* are correlated with entrainment measures, although the reverse is not true for the anti-social behaviors, *trying to dominate* and *conversation awkward*. For all three gender groups, especially for female-female and male-male pairs, entrainment measures are correlated positively with measures representing good dialogue flow and negatively with mean latency; interruptions, however, are also positively correlated with entrainment.

An interesting observation emerging from this study is that entrainment seems to be more important to social behavior for mixed-gender pairs, and more important to dialogue coordination for same-gender pairs. This may be related to the finding (Chapter 9) that entrainment is most prevalent, and strongest for some features, for mixed-gender pairs. Further support for this speculation comes from the observation that intensity is the feature for which entrainment is most consistently correlated with dialogue coordination for all three gender groups, and for which the strongest and most consistent evidence of entrainment can be found.

Different gender groups show associations between dialogue coordination and entrainment on different features. Entrainment on shimmer is correlated with all four coordination measures for female-female pairs and with none for male-male pairs; entrainment on speaking rate is correlated with number of turns and mean latency for female-female pairs, but not for male-male pairs. Male-male pairs, meanwhile, show associations between coordination and entrainment on pitch features and NHR, which are not associated with coordination for female-female pairs. There is no clear reason why entrainment on certain features should be associated with dialogue flow for one gender group and not another, but this is likely to be related to alignment on other levels of communication (c.f. [Pickering and Garrod, 2004]). Shimmer, in particular, has been shown to be indescernible by human listeners [Kreiman and Gerratt, 2005], and entrainment on this feature is most probably the result of alignment on a higher-level speech phenomenon such as vocal fry.

To return to the research questions posed at the beginning of this chapter:

- *What kinds of entrainment are most important in human conversations?* Our results show that this question should be answered differently depending on the genders of the conversation's participants. For mixed-gender conversations, the perception of *trying to be liked* is associated with entrainment on pitch features. Entrainment on shimmer is correlated with multiple measures of dialogue flow for female-female pairs, as is entrainment on pitch features and NHR for male-male pairs. A spoken dialogue system that can implement entrainment on only one feature should choose intensity; it is associated with *gives encouragement* for male-male and mixed-gender pairs and with coordination measures for all gender groups. It may also entrain on shimmer if both the human user and its own persona are female, on NHR if they both are male, or on pitch if one or both are male.

- *Can entrainment be used as a feature for evaluating conversations or interlocutor behavior?* The degree of acoustic-prosodic entrainment can be a valuable unsupervised feature for detecting pro-social behavior in mixed-gender conversations, and for measuring “click” or rapport in same-gender conversations.

This study looks only at correlations between social behavior and global entrainment. Future work should examine associations with other aspects of entrainment, such as local entrainment and convergence. In addition, the social variables here are annotated by outside judges; self-reports of rapport and social behavior by the conversation’s participants would be better indicators of a dialogue’s quality. The most valuable improvement to this study would be a paradigm that allows for the establishment of a causal link between entrainment and positive social behavior, by manipulating the degree of entrainment in a systematic way.

The correlations found here are not strong — none is higher than 0.65, and most are around 0.30. However, they are highly significant, and provide solid support for the link between entrainment and successful conversations. This link, whether correlative or causative, motivates the second half of this thesis, the implementation of a virtual conversational agent that entrains to a human interlocutor.

Chapter 11

Conclusions and Future Work

Our studies on acoustic-prosodic entrainment in human-human conversations build a comprehensive understanding of the phenomenon and provide a framework for analysis that can influence the design and interpretation of future studies of entrainment. Previous research tends to make global statements about the characteristics of entrainment; our results show that each combination of feature, entrainment metric, and social context must be considered on its own.

We make use of this framework to analyze how entrainment behavior differs between interlocutor pairs of different genders and speakers of different languages. Our comparison of entrainment in Mandarin Chinese and Standard American English is the first study of entrainment in Chinese and the first cross-linguistic comparison of entrainment patterns. We show that acoustic-prosodic entrainment takes place at the global and local level, by value and by direction, and that it improves with time, for conversations in each language. Consistencies and discrepancies in entrainment behavior between speakers of different genders and languages lead us to suggest that entrainment on some features (such as intensity and speaking rate) is more automatic, while for others (such as pitch) it is more strongly mediated by the interactional context.

Future work should pursue this direction, which has the potential to shed light on the cognitive underpinnings of entrainment. Three particular directions would be immediately useful. Firstly, further cross-linguistic studies should be conducted on a greater array of languages, to increase the number of data points and make patterns more apparent. Secondly, the experiments for analyzing local similarity, synchrony and convergence should be replicated for differentation by gender; our study looks only at global similarity. Finally, we propose one entrainment metric that allows for direct comparisons in Chapter 9 but rely for the most part on indirect comparisons; metrics that can be directly compared across different conditions would allow for a more rigorous exploration of interactions between entrainment and social context.

Our study of entrainment in outliers contributes further insight into the relationship between entrainment, perception, and social signals. We show that for some features, speakers entrain more to outlier realizations of those features, possibly because the outlier realizations are more perceptually salient than normal realizations. This work has implications for systems that attempt to induce the user to entrain; future work should test the hypothesis that system utterances with outlier feature values will be more readily entrained to by a human user.

With our analysis of entrainment on backchannel-inviting cues, we introduce the idea of entrainment on a complex, pragmatic speech feature. Our finding that such entrainment does exist suggests interesting directions for future research on other such features, such as other turn-taking cues, the form and function of questions, or the realization of a distinction between given and new information. Future work can also attempt to incorporate these findings into a turn-taking model to improve prediction of whether the user is inviting a backchannel (by looking at the system's backchannel-inviting behavior) and generation of natural and effective backchannel-inviting cues (by modeling cues used by the human user).

Finally, we show that global entrainment on certain acoustic-prosodic features is correlated with externally annotated measures of social behavior and automatically derived measures of rapport and dialogue flow. We observe that entrainment seems more important to social behavior for mixed-gender pairs, and more important to dialogue coordination for same-gender pairs, and that entrainment on intensity is most consistently associated with dialogue coordination for all three gender groups. Extending this analysis to other conceptualizations of entrainment (local, synchrony, and convergence) as well as other languages is an interesting direction for future work.

Since entrainment has been shown to correlate with dialogue quality, an important goal of these studies, aside from the theoretical contributions of this work, is to facilitate the implementation of entrainment models in a virtual conversational agent. We have shown that depending on the acoustic-prosodic feature, language, and speaker and interlocutor gender, an entraining system should match the user globally, at each turn, and/or with increasing precision as the interaction progresses; for some features, it should begin turns with a feature value from a point in its range *opposite* to the one with which the user has ended her previous turn. Furthermore, the system should design its realization of turn-taking cues and other pragmatic speech features with reference to the user's output.

In the second part of this thesis, we present a method for implementing some of these entrainment behaviors in a spoken dialogue system, and show that an entraining avatar improves the user's opinion of the system's quality.

Part II

Entrainment in a Spoken Dialogue System

Chapter 12

Motivation and Research Goals

Many factors point to the desirability of implementing the capability for acoustic-prosodic entrainment in a spoken dialogue system. Firstly, we have shown that entrainment is highly prevalent in human dialogues. Entrainment occurs at all levels of communication, under diverse conditions, on many features and in many ways, showing that it is a key component of natural human communication. A virtual speaker that does not entrain to its human interlocutor cannot be perceived as wholly natural.

In addition, as we and others have shown, entrainment in human conversations is highly associated with dialogue quality. As shown in Chapter 10, acoustic-prosodic entrainment is correlated with the perception of pro-social behavior and with automatically derived measures of dialogue coordination and flow. While the associations in the literature are mainly correlative, in conversation with a virtual speaker we hypothesize that entrainment will be causative of pro-social outcomes, since humans are accustomed to the association of entrainment and dialogue quality.

Trust is a social behavior that is of paramount importance in human-computer interactions, as computers — in the form of robots or self-driving cars — are assigned increasingly autonomous agency. Whether in combat, a rescue situation, an assembly line, or on the road, in order to cooperate effectively with a robot, a human must trust that the robot will behave the way it is supposed to. We propose that entrainment can be a mechanism for promoting trust in a virtual conversational agent. Several studies have shown that assigning human-like qualities to an autonomous agent can create trust in that agent ([Waytz *et al.*, 2014;

Hancock *et al.*, 2011]). An entraining agent will appear more humanlike, and therefore more trustworthy. In addition, entrainment is a social, other-directed activity that may make a speaker seem more worthy of trust.

In this part of the thesis, we describe an experimental system that implements acoustic-prosodic entrainment in a virtual speaker; to our knowledge, this is the first system to do so in a flexible, largely unsupervised way. Our method is lightweight, adding little latency to the interaction; it relies on existing technologies, treating ASR and TTS as black boxes. In addition, we describe a study that shows that human users prefer an entraining agent and rely on it to a greater extent.

This work constitutes a valuable contribution to research both on entrainment and on spoken dialogue systems. It presents a method for incorporating the considerable body of knowledge of acoustic-prosodic entrainment into a virtual conversational agent: studies that have been observational can become blueprints for a system design. Furthermore, an entraining system can be a valuable mechanism for testing the associations between entrainment and dialogue quality in the absence of confounding factors. With respect to spoken dialogue systems, this work introduces new possibilities for improving the user experience in a way that is orthogonal to improvements that can be made to a system's core components.

Chapter 13

Related Work

Most research on entrainment has involved human-human conversations. However, several studies have looked at whether humans will entrain to computers. [Bell *et al.*, 2000] found that humans entrain to a dialogue system with both a graphical and spoken interface on the modality of deictic references. Prosodically, [Bell *et al.*, 2003] found that humans entrain to a system's speaking rate, and [Coulston *et al.*, 2002] found that children entrain to the volume of an animated persona. [Brennan, 1991; Brennan, 1996] found that humans lexically converge to a speech interface to a degree comparable to interactions with a human partner, but suggests that the reasons for converging may be different; they may be navigating around the system's limitations, rather than negotiating common ground. When comparing student interactions with a virtual tutor, [Thomason *et al.*, 2013] showed that users entrained more on pitch to a prerecorded tutor voice than to a synthesized one. This finding has implications for systems that attempt to induce the user to entrain.

Several such systems have attempted to leverage entrainment by designing the system's behavior to elicit the kind of user behavior that would be optimal for its performance. [Gustafson *et al.*, 1997] showed that subjects using a simulated dialogue system adapted their lexical choices to the lexical content of the system's questions. [Stoyanchev and Stent, 2009] showed that in a live system with actual users, changing the system's behavior affected the user's behavior: changing the system's primes to use certain syntactic constructions made the user more likely to use those constructions. Following this finding, [Lopes *et al.*, 2011] proposed an algorithm for prime selection whereby terms that failed to be recognized by ASR are replaced in the system's prompts by synonymous terms; successful primes will be adopted by the users and recognized correctly, while unsuccessful primes will be replaced in the next iteration of the loop. Using this method, they found that users did adopt the new vocabulary introduced in the system prompts, and ASR accuracy improved. In a subsequent paper, [Lopes *et al.*, 2013] supplemented this method to allow for what they call "two-way entrainment," which accounts for both user preferences and system performance; the use of a prime is upweighted if the user indicates a preference for that term by using it and downweighted if the user fails to entrain to it or if its use is followed by an ASR failure. Using this method, error rate is reduced by 10% and session length is reduced by 6%. This line of work shows that the lexical convergence of humans to computers demonstrated in [Brennan, 1996] can and should be incorporated into system design to ensure that the terms the system uses are the ones it wants the user to use.

[Fandrianto and Eskenazi, 2012] attempt to use entrainment to influence users to stop shouting and hyperarticulating, two speaking styles that are associated with word error rate and are more frequent during failing dialogues. Their method relies on two classifiers trained to detect each of the target speaking styles; when one of the styles is detected with greater than 55% probability, the system randomly chooses one of three strategies to induce the user to revert to unmarked speech. The first strategy, *explicit*, prompts the user to speak more quietly (for shouting) or "normally" (for hyperarticulation). The second strategy, *implicit*, is the entrainment strategy; the volume of the TTS voice is lowered in the case of shouting, and the speaking rate is increased in the case of hyperarticulation. A third strategy is *backoff*; the system changes the subject, with the reasoning that the user is likely to adopt a new behavior for the new context. The study found that all three strategies were successful in influencing the user to abandon the target behavior. The explicit strategy was most successful for both shouting and hyperarticulation, followed by implicit and then backoff for shouting; both the implicit and backoff strategies performed similarly for hyperarticulation. While the "entrainment" strategy does not outperform the explicit strategy, it is a more elegant solution; it is transparent to the user and obviates the need for an extra "instruction" turn. Since there is no reason for a human user to logically assume that the speaking style a system is using is the one easiest for it to understand, the success of the implicit strategy suggests that in this regard, entrainment is automatic rather than strategic: humans are relating to the system as though it were human.

The idea that humans apply typical social responses to technology is sometimes called the Computers as Social Actors paradigm; [Carlson *et al.*, 2006] call it the "human metaphor." This paradigm is extremely relevant to our hypothesis that users will prefer a virtual interlocutor that entrains to them, just as they prefer human interlocutors who entrain. [Nass *et al.*, 1994] showed that even experienced computer users applied social norms to computers: when interacting with computers, they behaved in accordance with politeness norms, applied concepts of "self" and "other" to the computer, and applied gender stereotypes. The study concluded that "findings in social psychology are relevant to human responses to computers."

In another study conducted under this paradigm, [Nass and Lee, 2001] tested whetherthe similarity-attraction principle, which posits that individuals will prefer other people who are similar to themselves, will apply when the "other people" are in fact a computer. They found that subjects interacting with a spoken interface to an online bookstore preferred an interface whose personality — as manifested by prosodic cues in the synthesized voice — matched their own. When the interface's personality matched their own, they liked the voice more, found it more credible, and were more likely to buy the book.

Several researchers have designed natural language generators that can dynamically entrain to a human user's language at each turn. [Brockmann *et al.*, 2005] model alignment by interpolating a default language model with a local model trained on recent dialogue context. This model allows for controlling the strength of the entrainment effect by manipulating the weights of the local model. [Buschmeier *et al.*, 2009] introduce a microplanner for natural language generation that uses a priming-based model of entrainment. Influenced by psycholinguistic and cognitive literature, their model accounts for both *recency* and *frequency* of use, as proposed in [Brennan and Clark, 1996] and further demonstrated by [Reitter and Moore, 2007]. The model is extremely flexible; it can be parameterized to account for different empirical findings about how humans entrain or to convey different speaker "personalities."

[de Jong *et al.*, 2008] move beyond strictly linguistic alignment to *affective alignment*, matching the user's affective style. They describe a model for dynamically adapting a virtual agent's linguistic style to the level of politeness and formality displayed by a user. The virtual agent will respond to a polite question with a polite answer, and to a rude question with a "less polite" answer; as in [Brockmann *et al.*, 2005], the level of alignment can be determined by a parameter setting. External raters judging generated conversations did not exhibit a clear trend of overall preference for any specific level of alignment; personal preference seemed to play a strong role. The authors note that alignment on formality often entails lexical and syntactic alignment, since the choices of lexical and syntactic constructions are constrained by the chosen affective style.

In a similar vein, [Bateman, 2006] proposes a trickle-down approach to implementing alignment in which lexical and syntactic choices emerge from the selection of an appropriate *micro-register*, the contextual configuration of an interaction. This approach is motivated by Pickering and Garrod's Interactive Alignment Model [Pickering and Garrod, 2004], which states that alignment at one level of communication entails alignment at other levels.

In a recent study, [Hu *et al.*, 2014] describe a natural language generator that can dynamically entrain to a range of features of the human user's utterances, including referring expressions, tense-modality selection, verb and noun lexical selection, hedges and cue words, syntactic constructions, and any combination of these. They report the results of an experiment in which they asked naive raters to rank a number of sentences containing all possible entrainment combinations in response to a particular utterance for *friendliness* and *naturalness*. They found that human raters perceived utterances that entrained on hedges as more friendly but less natural. Utterances that entrained to a user's referring expressions and syntactic and tense/modal choices were perceived as both more natural and more friendly.

A large body of work deals with the issue of promoting trust in robots. [Laursen, 2013] describes scenarios in which human operators "micromanage" their robots, failing to concentrate on their own tasks because they feel unable to rely on the robot's autonomy. He quotes Geert-Jan Kruijiff, an artificial intelligence expert at the German Research Center for Artificial Intelligence, as saying that "At a disaster scene with lives on the line, human rescuers need to learn to trust their robotic teammates or they'll have their own meltdowns." One approach holds that in order for a human user to trust a robot, the human must have an effective and complete understanding of the robot's capabilities. Humanizing the robot provides a paradigm for interaction that the human already knows: human-human interaction [Bruemmer *et al.*, 2004].

[Hancock *et al.*, 2011] compile a meta-analysis of the literature on factors affecting trust in human-robot interactions. They find that of the robot attributes affecting trust, the most important factors relate to the robot's performance, such as its reliability, false alarm rate, and failure rate. However, the robot's personality also affected trust: robot anthropomorphism was positively associated with trust. Several more recent studies have confirmed the benefits of anthropomorphizing robots. [Broadbent *et al.*, 2013] find that a robot with a more humanlike face display was perceived as having the most mind and as being most sociable and amiable. [Waytz *et al.*, 2014] show that user trust in an autonomous vehicle was predicted by the number of its anthropomorphic features (name, gender, and voice). These results motivate our hypothesis that entrainment can be a mechanism for promoting trust in a virtual agent.

The only example in the literature of a system that entrains to *how* the user speaks rather than *what* he or she says is [Ward and Nakagawa, 2004], which uses multiple regression to derive a formula of how the speech rate of human operators dictating numbers to human users in a corpus of simulated directory assistance dialogues relates to the speech rate of the user utterances to which they are responding. They then use this formula to obtain a target operator speech rate for a response to a given user utterance, and synthesize an operator utterance with the desired rate. This study is notable as a first attempt to enable a computer to entrain; however, their approach is designed for the specific task of entraining the speech rate of a number-giving operator and does not extend to other acoustic-prosodic features or kinds of dialogues.

Chapter 14

A Method for Implementing Acoustic-Prosodic Entrainment

This chapter presents a method for dynamically entraining a system's TTS output to a human user's prosody. This work is complementary to the methods (described in Chapter 13) for dynamically aligning to a human user's *words* — the lexical, syntactic, and/or stylistic characteristics of the input. To our knowledge, there has been no implementation of an interactive voice system that can dynamically entrain to multiple acoustic-prosodic characteristics of any input.

The primary motivation for implementing acoustic-prosodic entrainment is its extraordinary prevalence in human dialogue. Humans are used to dynamic, responsive interlocutors; a virtual conversational agent that does not adapt to its human interlocutor's prosody can never be perceived as wholly natural.

Additionally, numerous studies have shown that acoustic-prosodic entrainment is associated with desirable dialogue characteristics in human-human dialogues, while others suggest that the benefits of entrainment should apply in human-computer interactions as well. We will address this motivation more fully in Chapter 15.

In the following sections, we describe the method we use to entrain to a user's voice in the acoustic-prosodic domain and then discuss its performance on a small test corpus.

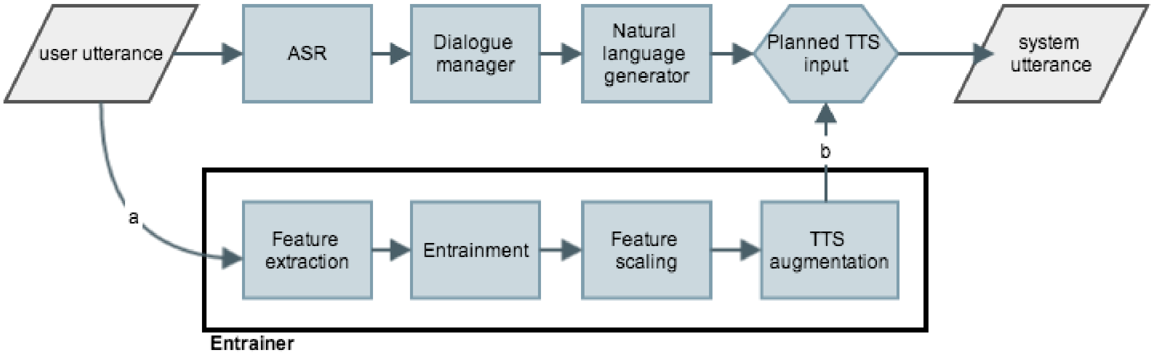


Figure 14.1: The entrainer integrated in an existing system.

14.1 Method

Entrainment relies on two processes: (1) Perceiving features of the interlocutor's speech, and (2) reproducing those features in the speech output. For (1), we use a Praat [Boersma and Weenink, 2012] script to extract acoustic-prosodic features from a user utterance. We can then accomplish (2) by transforming those features into SSML markup for a TTS output utterance. SSML markup enables a system designer to manipulate TTS output without knowledge of or access to the component's inner workings. Our method can therefore be easily integrated into any existing system.

The method has four steps, illustrated in Figure 14.1.

(1) Feature extraction

The entrainment loop is triggered when the system captures a user utterance. While the ASR translates the audio to text, a Praat script extracts a set of acoustic-prosodic features from the signal. Praat is an open-source audio processing tool; its accompanying scripting language makes it a flexible and powerful tool for conducting audio analysis at scale. The script uses Praat's native algorithm to calculate intensity mean (in decibels), and a Prosogram Praat script [Mertens, 2004] to calculate speaking rate (syllables per second). It is simple to use Praat to extract other features, such as pitch, jitter, shimmer, or NHR.

(2) Entrainment

In this preliminary implementation, the entrainment step is naive: we simply have the system match the user's acoustic-prosodic features as closely as possible at each turn. This entrainment is local, since it strives for similarity at defined points in the interaction; since the system matches the user at every point, their overall averages will be similar, so entrainment is global as well. Global entrainment alone might be implemented by adapting the system to the prosody of the user's first few turns and keeping it at that level. Local entrainment alone would involve adapting to each user turn, but within a different feature range. In a single system, different strategies can be used for each feature, as we have shown to occur in human behavior (Chapter 5).

Here, we implement both global and local entrainment for all features supported by the TTS, including speaking rate, for which we have *not* in fact found in our data that humans entrain locally, to demonstrate the flexibility and responsiveness of the method. In addition, exact matching is not necessary and may not even be desirable, since entrainment in human-human conversations is not exact. In an actual system, a designer may wish to emulate human behavior more closely, or test different strategies for their respective impact on interaction quality, real or perceived.

(3) Scaling

The `prosody` element of the SSML specification[1] allows for control of the three main aspects of prosody: volume, pitch, and speaking rate. SSML is not intended to allow a designer to enforce specific acoustic-prosodic targets; some of the implementation is processor-dependent, and not all the attribute values use standard units.

The `volume` attribute modifies the output's loudness on a scale from 0 (silent) to 100 (the processor's maximum volume). Throughout this chapter, we use the terms "volume" and "intensity" interchangeably. To translate between the decibel measures output by Praat and this relative scale, we generate utterances at 10-unit intervals from 0 to 100 and extract their intensity measures in decibels. For a given decibel target, we can then interpolate into

[1] http://www.w3.org/TR/speech-synthesis/

this series to find the appropriate attribute value. This translation is processor dependent, requiring a configuration step at time of system integration.

Pitch can be modified by three attributes: `pitch`, which sets baseline pitch relative to default or in Hz, `contour`, which sets a pitch contour for the text, and `range`, which sets the pitch range relative to default or in Hz. We do not entrain on pitch features in this study, but doing so should be straightforward, since the attributes use standard units.

Speech rate (referred to throughout this chapter as simply "rate"), calculated as words per minute, can be modified relatively using the `rate` attribute. For greater accuracy, we use the `duration` attribute to specify how long the output utterance should be. The Praat script gives us the target rate in syllables per second. With knowledge of the syllable count of the output utterance (obtained from a syllable dictionary), it is straightforward to calculate the duration necessary to obtain the target rate.

(4) Generate output

As a sanity check, attribute values are compared against the maximum and minimum values that have been determined to result in utterances that sound normative — not, for example, unusually loud, unusually slow, or distorted. Generating these ranges requires another configuration step, in which the range of values available for each attribute is explored and the extremes of the range that result in utterances perceived as normal are identified.

Before synthesizing the TTS output with the entrained parameters, each attribute value is compared with the range of acceptable values. If it lies outside this range, it is replaced with the extreme to which it is closest. The output string is then augmented with a `prosody` tag incorporating the scaled, translated and vetted attribute-feature pairs.

Enforcing that output feature values lie within a normal range also addresses a potential pitfall of an entraining system, which is the possibility of a vicious cycle of entrainment in which the system entrains to the user, the user entrains to the system, the system entrains to the user's new value, and so on until both the user and the system are shouting, for example, or speaking unacceptably slowly. Instead, the system will stop entraining when it reaches the bounds of normal speech behavior.

14.2 Performance

To test the performance and accuracy of this method, we make use of 19 user interactions that were collected in the course of an experiment that will be described in the following chapter. Each session consists of approximately 47 user utterances, with 932 utterances in total. For each user utterance, we generate three TTS utterances using Cepstral 6, a unit selection based text-to-speech engine, with the "David" voice.

- **Entrained.** The entrained utterance is generated using the method described above.
- **Random.** A second utterance is generated with volume and rate values chosen randomly from a range of values previously determined to sound normative.
- **Default.** The third utterance is generated without markup, using the TTS's default values for volume and rate.

We then use statistical tests (as in Chapter 5) to verify the entrainment accuracy of our method. "Entrainment accuracy" refers to the degree to which the **entrained** utterances match their corresponding user utterances, globally and locally, compared to the **default** or **random** baseline.

14.2.1 Timing

In order for this method to be usable in a real-world system, it must not add a prohibitive amount of latency to the system response time. On average, Cepstral's `swift` command completes in 1.04 seconds for the **default** utterances. The **entrained** utterances, on average, take 1.73 seconds to complete. The difference in times is primarily taken up by the Praat feature extraction script, which takes an average of 0.70 seconds. This difference, while small, is substantial in this context. However, since the script runs in parallel with the automatic speech recognizer (ASR), the extra latency can be subsumed by the time the ASR takes to complete. We experimented with Pocketsphinx [Huggins-Daines *et al.*, 2006], which took an average of 2.19 seconds per utterance, meaning that in an actual interaction the user would experience no difference in latencies. A state-of-the-art dialogue system is likely to have much lower ASR response latencies, but it is also likely to have more complex dialogue management, which may add latency. Finding strategies for increasing the efficiency of the feature extraction is a useful direction for future work, but much of the latency is likely to be subsumed by the core workings of the dialogue system, since it works in parallel.

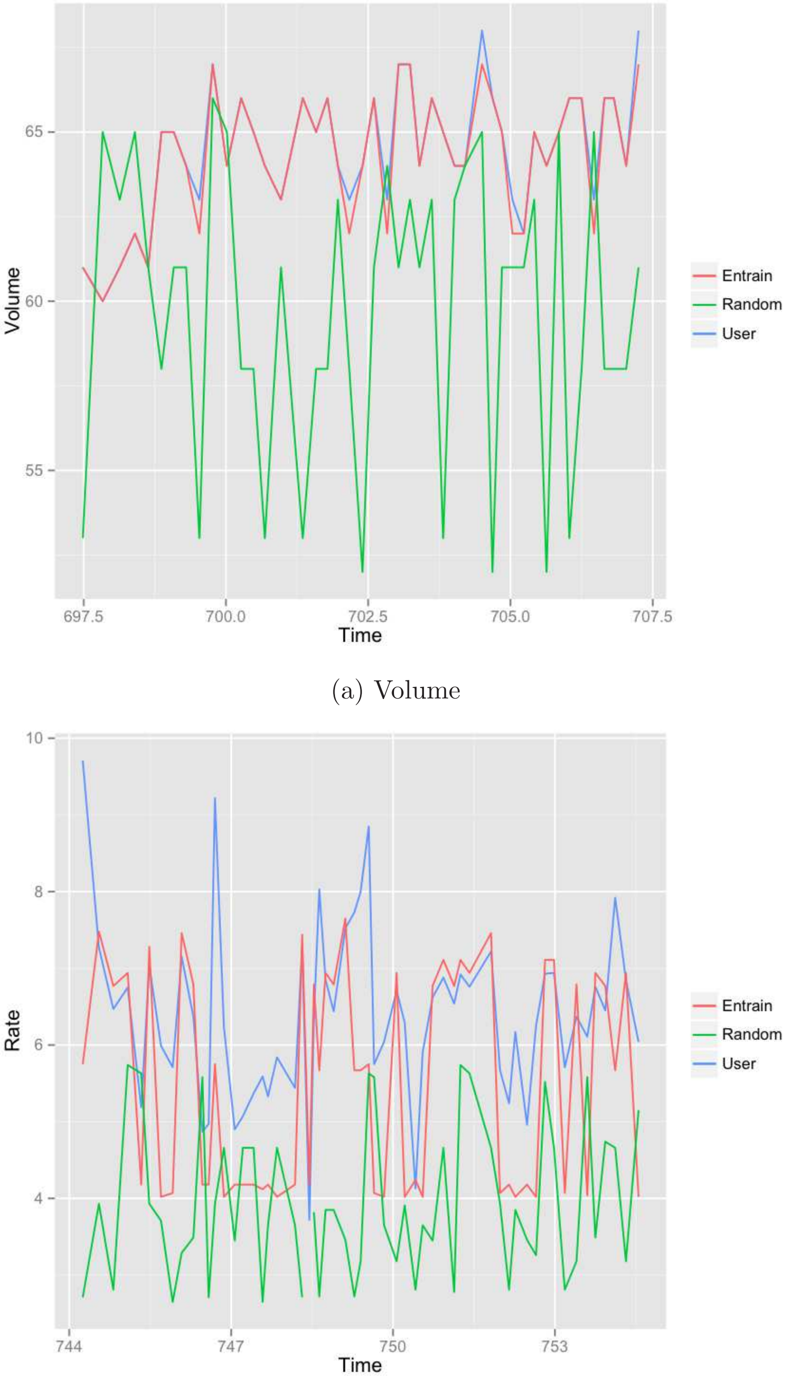
14.2.2 Error sources

If the method worked perfectly, the features of each **entrained** utterance would be identical to those of its corresponding user utterance. However, several factors can affect the accuracy at each stage of the entrainment process.

Accurate feature extraction is always difficult in a noisy environment. Signals from competing sounds may artifically inflate intensity values and distort pitch tracks. Even when recording conditions are ideal, as they were in our data collection process, with the subject wearing a head-mounted microphone in a soundproof booth, factors such as the subject's positioning the microphone too close or too far from his or her mouth may artificially affect the intensity of the input utterance. Speech rate can be difficult to compute accurately even when the signal is clean, since it involves finding stresses in the input utterance, a higher-level feature that is more complicated to detect; background noise can make this hopelessly inaccurate.

We minimize to some extent the potential impact of feature extraction errors by enforcing sanity checks on all output feature values. That is, we do not blindly inject extracted feature values into the SSML markup; instead, we enforce that each output parameter must lie within a predetermined range of acceptable values. We assume that values outside this range are errors, and abandon the goal of entrainment in favor of ensuring that the output sounds natural.

Sessions 2, 3 and 6 are examples of cases in which the user's input volume fell outside the normal range, most likely because the headset was worn incorrectly, but possibly because the user in fact has an unusually soft or loud tone. Instead of producing the output utterance inaudibly or with unacceptably loud volume, the system outputs each utterance at the range extreme closest to the extracted feature values. Since each utterance has the same intensity



(a) Volume

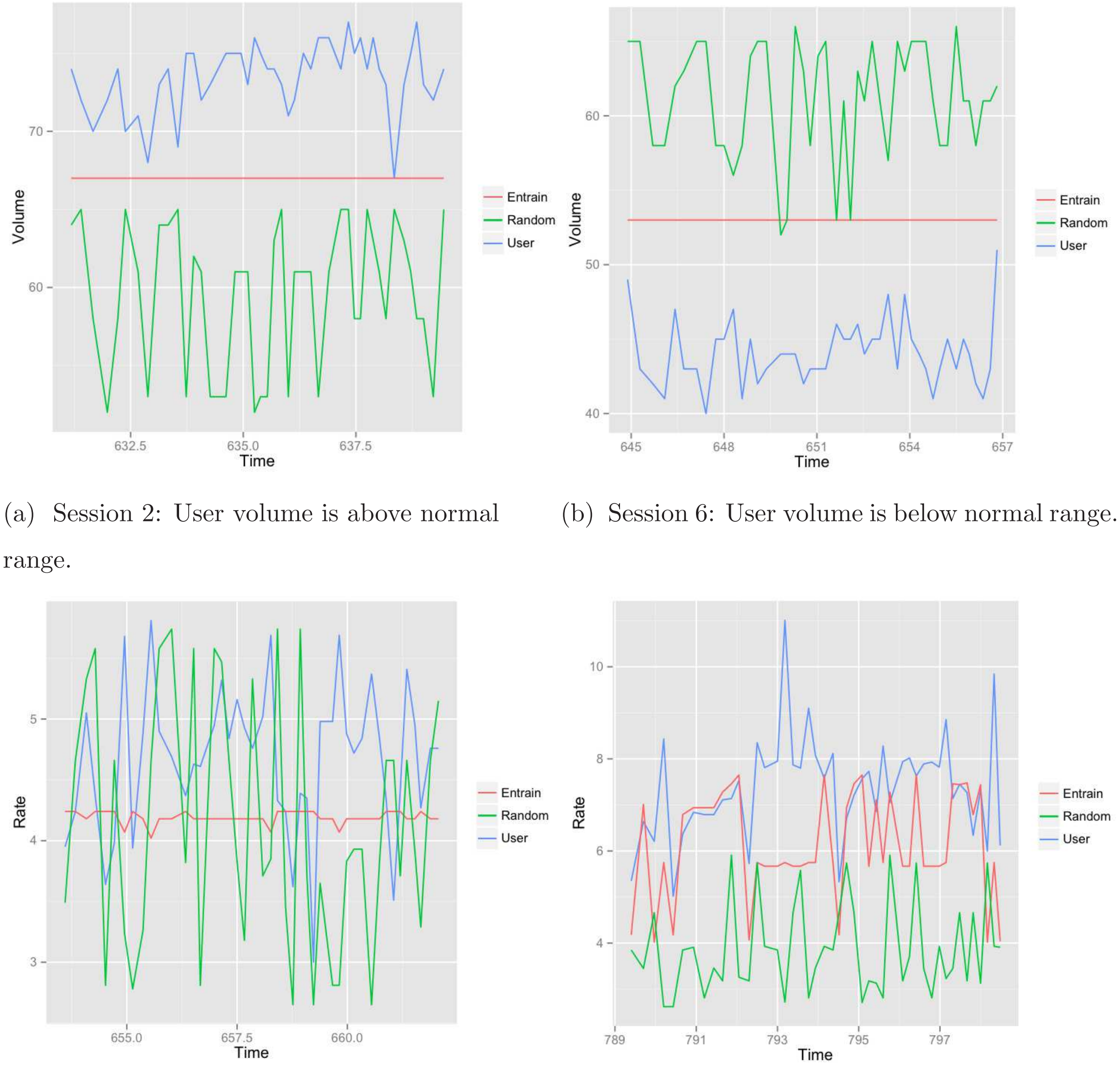
(b) Rate

Figure 14.2: Features of TTS output with respect to user input for a selected session. X-axis is turn index; y-axis is feature value.

value, correlation is impossible, but global entrainment is still present: the **entrained** mean is closer to the user's (perceived) mean than the **default** or **random** utterances. Even if the extracted feature values are wholly inaccurate, and the true means are dissimilar, the system has performed no worse than a non-entraining system would: it has effectively fallen back on the standard behavior for this feature. Figures 14.3a and 14.3b show the volume graphs for Session 2, where the user volume is too high, and Session 6, where the user volume is too low.

Another source of error manifests itself primarily in the output *speech rate* values, as shown in the rate graphs of two example sessions in Figures 14.3c and 14.3d. As stated before, the SSML markup is not intended to allow designers to enforce specific feature values; its purpose is rather to allow for relative, perceptual feature adjustments. Consequently, the TTS output does not always conform to the SSML specifications, possibly in cases where doing so would interfere with objectives that are given higher priority in the model. The degree of this "fuzziness" may vary across architectures. When using the Cepstral TTS, although the **entrained** utterances in every session were more similar in speech rate to their corresponding user utterances than the **default** utterances were, the correlations between **entrained** and user speech rates were not significant (or negatively correlated) for three sessions, and weaker than 0.5 for another eight. This is in contrast to the intensity values, for which over half the sessions have correlations stronger than 0.9; only one is weaker than 0.7. For speech rate, at least, the Cepstral TTS (this, too, may vary across architectures) does not exactly follow the prosodic directives embedded in the markup. This method can only succeed to the extent to which the TTS conforms to its specifications.

Another constraint comes from the quality of the TTS output. Many of the legal values for the `prosody` attributes yield output that sounds unnatural and even bizarre. As discussed above, we hedge against that possibility by enforcing that output feature values lie within a predetermined acceptable range. However, for some attributes and implementations, any manipulation at all results in distorted output. For the Cepstral TTS, while the `volume` and `rate` attributes can be manipulated to a certain extent and sound natural, any manipulation of the pitch attributes (`pitch`, `contour`, and `range`) causes the output utterance to sound distorted and unnatural. For that reason, we only entrain



(a) Session 2: User volume is above normal range.

(b) Session 6: User volume is below normal range.

(c) Session 14: Entrained rate does not track user rate.

(d) Session 19: Entrained rate does not track user rate.

Figure 14.3: Sessions with entrainment errors.

on intensity and speaking rate in this analysis. The ability to entrain on a given feature is constrained not only by the SSML specification but by the quality of a given TTS engine's implementation.

14.2.3 Entrainment accuracy

The rate and volume of user and agent turns in a sample dialogue are shown in Figure 14.2. The blue line represents the volume (in Figure 14.2a) and rate (in Figure 14.2b) of the user's input turns. The red line represents the feature values of the **entrained** utterances. For comparison, the green line shows the feature values of the **random** utterances.

The effect of entrainment is apparent from this pair of graphs. In Figure 14.2a, the user's intensity is tracked so exactly that the **entrained** feature values (in red) almost entirely overlay the user's feature values (in blue). In Figure 14.2b, the hills and valleys of the user utterances' rates are not as precisely tracked, but the overall shapes of the user and entraining agent's lines are clearly similar. In addition, *global* entrainment (Chapter 5) is more unambiguously present: the user and entraining agent have similar feature means for both volume and rate.

For each session of actual user utterances and simulated TTS utterances, we verify that the **entrained** utterances do in fact entrain globally and locally as in Chapter 5. For all except one session, the **entrained** utterances' session volume means are more similar to the user's than the **random** ones are; for all sessions, the **entrained** utterances' session rate means are more similar. The average difference in intensity between **entrained** and user utterances is 1.72 dB ($SD = 2.59$), while the average difference between **random** and user utterances is 6.12 dB ($SD = 4.70$); the corresponding average differences in rate (syllables per second) are 1.03 ($SD = 0.27$) and 2.26 ($SD = 0.74$). These differences are significant according to paired t-tests (volume: $t(18) = -7.56, p = 5.4e - 07$; rate: $t(18) = -8.51, p = 1.0e - 07$); we can therefore conclude that our method implements global entrainment as intended.

Locally, we compare the differences between each user turn and its corresponding **entrained** utterance with the differences between each user turn and its corresponding **random** utterance. Table 14.1 shows the results of paired t-tests for these comparisons. For every session, **entrained** utterances are more similar to their corresponding user utterances for both volume and rate than the **random** utterances are, which is the definition of local similarity. Finally, for each session, we calculate the Pearson's correlation between the user feature values and the corresponding **entrained** feature values (Table 14.2). The correlations for volume are nearly all extraordinarily strong (nearly half are over 0.95), indicating near-perfect synchrony. The correlations for rate are more moderate, though most are still significant; the strongest correlation (Session 14) is actually *negative*, while Sessions 12 and 19 do not show synchrony.

14.3 Discussion and future work

This work is the first to enable an interactive voice system to *entrain* to multiple aspects of its human interlocutor's prosody, as a human speaker does. This problem is entirely distinct from the problem of implementing *lexical* entrainment, since it involves continuous features. The method proposed here is lightweight, adding little latency to the system response time, and treats the rest of the system as a black box, modifying the TTS output using the existing SSML specification. It can therefore be easily integrated into an existing system whose TTS engine supports SSML, as illustrated in Figure 14.1, by sending the user utterance to the feature extraction script as it goes through the ASR (arc `a`) and then augmenting the planned TTS input with the `prosody` tags output by the entrainer (arc `b`). Aside for those modifications, the only other configuration step necessary is to test the range of each feature manipulation for quality verification, which can be semi-automated.

As the diagram shows, the entrainer works entirely in parallel with the dialogue system's core functions. However, state-of-the-art systems may have response latencies low enough to expose the latency added by the feature extraction algorithm. An additional weakness of the method is its sensitivity to recording conditions and feature extraction errors. Future work should make use of feature extraction algorithms that are faster and more robust.

The design of this method is informed by our research on entrainment in human-human conversations, but does not fully implement the range of entrainment behaviors observed in humans. Future work should implement global similarity without local similarity, local

Session	*t*	*df*	*p*	*Sig.*
1	-9.49	44	3.2e-12	*
2	-10.80	44	5.9e-14	*
3	-12.15	46	5.8e-16	*
4	-6.37	48	6.8e-08	*
5	-12.14	44	1.2e-15	*
6	-13.92	44	1.0e-17	*
7	-10.81	51	8.5e-15	*
8	-10.86	56	2.1e-15	*
9	-13.3	45	3.4e-17	*
10	-10.08	43	6.8e-13	*
11	-10.46	44	1.6e-13	*
12	-13.03	44	1e-16	*
13	-8.57	44	6.3e-11	*
14	-8.16	44	2.4e-10	*
15	-13.12	68	2.6e-20	*
16	-9.19	46	5.5e-12	*
17	-9.41	44	4.2e-12	*
18	-8.69	44	4.2e-11	*
19	-14.27	45	2.6e-18	*

(a) Volume

Session	*t*	*df*	*p*	*Sig.*
1	-5.5	44	1.8e-06	*
2	-6.29	44	1.3e-07	*
3	-4.96	46	9.9e-06	*
4	-9.07	48	5.7e-12	*
5	-3.74	44	0.00052	*
6	-3.42	44	0.0014	*
7	-7.31	51	1.8e-09	*
8	-7.37	56	8.3e-10	*
9	-5.46	45	2e-06	*
10	-4.78	43	2.1e-05	*
11	-9.18	44	8.8e-12	*
12	-10.47	44	1.6e-13	*
13	-4.7	44	2.6e-05	*
14	-4.21	44	0.00013	*
15	-6.79	68	3.4e-09	*
16	-8.37	46	8.4e-11	*
17	-4.15	44	0.00015	*
18	-6.42	44	8e-08	*
19	-9.96	45	5.9e-13	*

(b) Rate

Table 14.1: *T*-tests for differences between entrained and random similarities in utterance rate and volume.

Session	*t*	*df*	*p*	*Sig.*
1	0.91	43	0	*
2	NA	NA	NA	
3	NA	NA	NA	
4	0.91	47	0	*
5	0.71	43	6.4e-08	*
6	NA	NA	NA	
7	0.87	50	0	*
8	0.7	56	1.1e-09	*
9	0.97	44	0	*
10	0.98	42	0	*
11	0.57	43	4.7e-05	*
12	0.99	43	0	*
13	0.98	43	0	*
14	0.98	43	0	*
15	0.72	67	2.5e-12	*
16	0.79	46	2.2e-11	*
17	0.99	43	0	*
18	1	43	0	*
19	0.99	44	0	*

(a) Volume

Session	*t*	*df*	*p*	*Sig.*
1	0.74	43	5.8e-09	*
2	0.66	43	9.7e-07	*
3	0.73	45	7.4e-09	*
4	0.46	47	0.00089	*
5	0.63	43	3.1e-06	*
6	0.68	43	2.8e-07	*
7	0.29	50	0.036	*
8	0.59	56	1.2e-06	*
9	0.71	44	4.5e-08	*
10	0.75	42	5e-09	*
11	0.44	43	0.0026	*
12	0.05	43	0.77	
13	0.46	43	0.0013	*
14	-0.83	43	1e-12	*
15	0.5	67	1.5e-05	*
16	0.52	46	0.00016	*
17	0.48	43	0.00086	*
18	0.62	43	6.2e-06	*
19	0.2	44	0.19	

(b) Rate

Table 14.2: Pearson's correlation between user and entrained rates and volumes.

without global, and convergence at each level, individually and in combination; we broadly described how this might be done in the beginning of this chapter.

The set of which features can be entrained is multiply constrained by 1) which features are included in the SSML specification (for example, voice quality features are not included); 2) what parts of the SSML specification are implemented by the TTS engine; and 3) the quality of the TTS implementation (for example, manipulating the pitch of the Cepstral TTS results in distorted output). A potentially better alternative is to use a tool such as Praat to manipulate the audio of the TTS output after it has been generated by the core system, instead of working within the TTS engine. This approach has significant advantages: it is unconstrained by the capabilities of a specific technology, and it eliminates the need for a semi-automated configuration step, making integration considerably easier. However, it sacrifices the very important benefit of using the TTS's native prosodic specifications, and is likely to double the added latency; in addition, the output quality may suffer. Another interesting direction for future work is to explore the relative advantages of each of these approaches.

The implementation of the capability to prosodically entrain in an interactive voice system introduces new directions for research into how this behavior may affect the user-system dynamic. The next chapter describes a human-subjects study that addresses this question.

Chapter 15

User Interactions with an Entraining Avatar

To demonstrate the utility of implementing acoustic-prosodic entrainment in a spoken dialogue system, we created GoFishWithHelpers, a game in which subjects interact with an entraining avatar and a non-entraining one. GoFishWithHelpers is designed to implicitly test the hypothesis that speakers will trust an entraining avatar more than one that does not entrain. In addition, we explicitly ask the subjects for their impressions of each avatar, expecting that they will prefer the entraining avatar and its voice.

15.1 Data collection

Figure 15.1 shows screenshots of GoFishWithHelpers in progress. The player and the system are each initially dealt a hand of seven cards. The player's goal is to acquire cards from the system's hand, in order to earn points. In the canonical game of Go Fish, the player can ask her opponent for cards of any rank that she already has a representative of in her own deck. For example, in Figure 15.1a, she can ask for aces, kings, queens, sevens, threes, or twos. The opponent will then give her all the cards of that rank that he has in his hand. If he has no card of the rank she has requested, she has to "Go Fish," selecting a card from the top of the deck. In GoFishWithHelpers, the player receives ten points for each card she gets from her opponent, the system; 100 points for completing a "set" (a rank in all four suits); and loses 50 points for “Go Fish.” To motivate the participants, they receive a monetary bonus corresponding with the number of points they earn.

In GoFishWithHelpers, trust is modeled by requiring the player to ask the advice of one of two avatars at each turn. At the beginning of each turn, the player’s hand is disabled; she cannot ask the system for a rank directly. Instead, she presses a button marked “Ask for help,” and requests the advice of one of the avatars by name. The two “helpers” are called Bobby and Freddy (Figure 15.2); in every game, one entrains to the subject while the other randomly changes its prosody. In the instructions before the game, subjects were told:

> *“Bobby and Freddy can see the system’s hand and will tell you which rank to ask for. They will usually give you good advice, but sometimes they will give you bad advice. One will give you bad advice more often than the other. Your role in this game is to decide who to trust.”*

For every turn, there are a number of different outcomes depending on which rank is requested. For example, consider a turn in which the player has the hand in Figure 15.1a (AKQ7332) and the system has the hand AAAKK33. If the player asks for aces, she will receive three aces from the system and complete the set of aces for a total of 130 points. Asking for kings will earn her two cards and 20 points; asking for threes will earn her two cards and 120 points. Asking for queens or twos will lead to “Go Fish” and a loss of 50 points.

While the player can see the gain or loss a particular request yields, she is unaware of the possible alternative outcomes. A “helper” who tells her to ask for kings in the previous example may seem to have given good advice, but he is in fact an untrustworthy helper, because a request for aces would have earned 110 more points. Alternatively, in a case where the player has a hand of AAKJJ93 and the system has Q877732, a helper who tells her to ask for threes will earn her only 10 points but is giving her the best possible advice. To further obscure the quality of the advice received, the system is dealt a new hand after every turn, so the player cannot infer the contents of the hand based on responses to her previous requests.

Each helper persona is programmed to give advice according to an algorithm that keeps

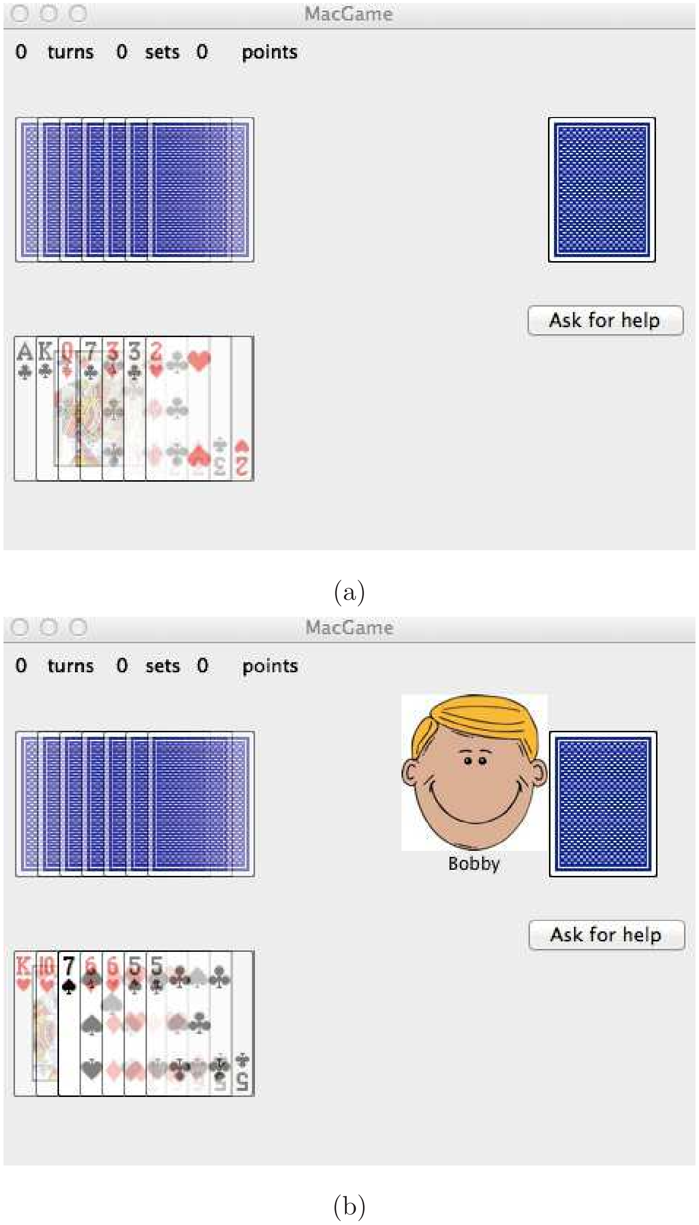


Figure 15.1: Screenshots of GoFishWithHelpers.

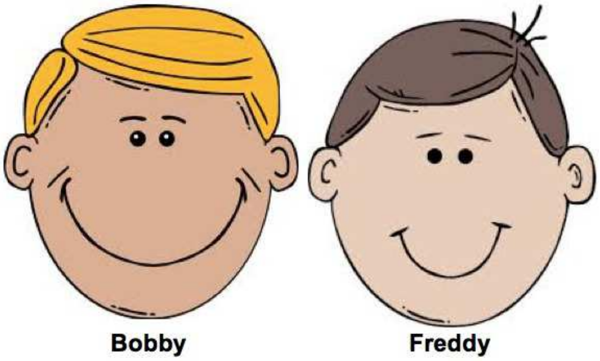


Figure 15.2: “Helper” avatars.

the persona’s global advice score—the overall perceived quality of the advice it has given so far—as close to zero as possible. Every rank that the player may request is assigned a score based on what the outcome of requesting that rank would be. If the rank would complete a set, its score is 5; if it would result in “Go Fish,” its score is −15; otherwise, its score is the number of cards the system has of that rank (1–3). These scores were empirically determined based on observing and discussing the gameplay with subjects in a pilot study, who expressed frustration at losing points to “Go Fish” that far outweighed their satisfaction at receiving cards from a successful request. At each turn, the helper that is asked for advice selects the rank whose score would bring the helper’s global advice score closest to zero.

The game design objectives of obscuring the *true* quality of an avatar’s advice and of enforcing that each avatar gives advice of the same *perceived* quality are intended to prevent the player from deciding whom to trust based on performance, and instead force him or her to rely on paralinguistic cues. The player cannot alternate between helpers for the first few turns and then rely on the one that performed best on those turns for the remainder of the game — a strategy some subjects used in the pilot study — because the *perceived* quality of each helper’s advice will be the same after several turns. However, the player is aware that the *perceived* quality does not necessarily reflect the *true* quality, which allows for the (false) illusion that one helper is in fact giving better advice.

The “Ask for help” button triggers recording. When the player has finished speaking, she toggles the same button (which now says “Done”); this finalizes the recording and sends it to the automatic speech recognizer and simultaneously to the feature extractor. Once these processes have completed, the program identifies the name of the requested helper from the user input, and the advice output is formulated based on the helper’s global advice score. If the helper is the entraining avatar, SSML markup is added to the TTS input to make it align to the intensity and speech rate of the user’s preceding turn, as described in Chapter 14. Otherwise, the prosodic parameters are randomly chosen from within ranges manually confirmed to sound acceptable. In the pilot study, the control (non-entraining) avatar maintained a static prosody, without intervention, as do most state of the art conversational avatars, but subjects in the pilot study reported noticing that “one helper sounds flat and boring.” The control avatar was therefore modified to vary its prosody within a natural range, as the entraining avatar does.

The selected helper’s face appears on the screen, the helper’s voice speaks the advice (e.g. “Ask for nines”), and the cards of the rank that the helper has chosen are enabled in the player’s hand (Figure 15.1b). The player completes the turn by clicking the enabled cards, which serves as the “request” for that rank from the system; the system either gives the player cards of that rank from its hand and awards the associated points or deals the player a card from the deck and deducts 50 points. At that point, the system is dealt a new hand from the deck and the turn is over. There are 15 turns in a game, and each subject plays three games.

15.1.1 Survey and analysis

After three games have been completed, the subject is directed to a short survey that asks the following questions:

- Which advisor did you like better?
- Whose voice did you like better?
- Who gave better advice?

- Please check the adjectives you would use to describe {Bobby's, Freddy's}[1] voice.
  - Loud
  - Soft
  - Pleasant
  - Annoying
  - Fast
  - Slow
  - Natural
  - Strange

The choice between Bobby and Freddy is forced for the first three questions.

In addition, subjects are asked to complete TIPI [Gosling *et al.*, 2003], a short personality test measuring the Big Five personality traits.

15.1.2 Subjects

Nineteen subjects participated in this study. They were recruited through advertisement in the Columbia University community via flyers and social media. Nine subjects were female and ten were male; their ages ranged from 20 to 35. All subjects were native speakers of Standard American English. One speaker did not submit the post-interaction survey correctly; the analysis of the survey therefore includes only 18 subjects. The subjects were paid for their time and given an additional monetary bonus that corresponded with the number of points they earned in each game.

15.1.3 Data

The final data consists of 19 sessions, 18 of which have associated survey responses. Each session contains approximately 45 user turns (some sessions have extra turns due to users repeating themselves) and an equal number of system turns, some from an entraining TTS and some from the control. On average, each session is 9 minutes long ($SD = 1.72$). Intensity, speech rate, and other acoustic-prosodic features have been extracted from all user and system turns using Praat. In addition, the `<prosody>` command associated with

[1] The order of presentation of the helper names in these questions and in the choices for the previous questions is balanced across experimental conditions.

each system turn has been recorded, as well as the advice given by each avatar at each turn, along with its perceived value.

15.2 Results

Our analysis explored how the entraining avatars differed from the control avatars in trust scores, liking, and perceived vocal characteristics.

15.2.1 Bobby vs. Freddy

In the interest of creating distinguishable personas for the two helpers, each was given a name, a face, and a distinct voice. Each of these variables has the potential to introduce bias, although they were chosen to be similar to each other. Instead of attempting to balance these variables across experimental conditions, each set of attributes remained consistent so we can control for bias based on persona, although we are unaware of whether the bias stems from the name, the face, or the voice. Bobby has the Cepstral[2] David voice, and Freddy has the William voice; both are 30 year old males speaking US English; neither has noticeable affect or personality.

In the course of our analysis, we did observe a bias in favor of Bobby. Eleven out of 18 subjects who completed the post-interaction survey liked Bobby better, and 12 thought he gave better advice, although these differences were not significant. This bias is also apparent in the implicit trust measures: On average, in each session Bobby was asked for advice more often than Freddy was ($t(18) = 2.76, p = 0.01$). However, when advice requests are weighted by their turn index in the game, the difference is not significant ($t(18) = 1.63, p = 0.12$), indicating that this preference is only initial, and that as the interaction progresses and trust develops, other factors — such as entrainment — come into play.

The results of the post-interaction survey suggest that the bias in Bobby's favor is not due to any qualitative difference in their voices: Subjects were no more likely to say that Freddy's voice was "annoying" or "strange", or any less likely to say it was "natural" or "pleasant." Instead, it may stem from any of the other differences between Bobby and

[2] http://www.cepstral.com/en/personal

Freddy, from their names to their hair colors.

15.2.2 Entrainment and trust

The game was designed to implicitly model trust by requiring the player to decide whom to ask for advice at each turn; this choice was framed in the instructions by telling the player that "Your role in this game is to decide who to trust." Two trust scores were calculated for each helper in a game. The **raw score** is the number of times the player asked that helper for help throughout the game. The **weighted score** is the sum of the indices of turns in which that helper was asked for advice, with the intuition that trust in later turns should be weighted more heavily as an indicator of developing trust. The subjects were also explicitly asked in the post-interaction survey to choose which helper they thought had given better advice.

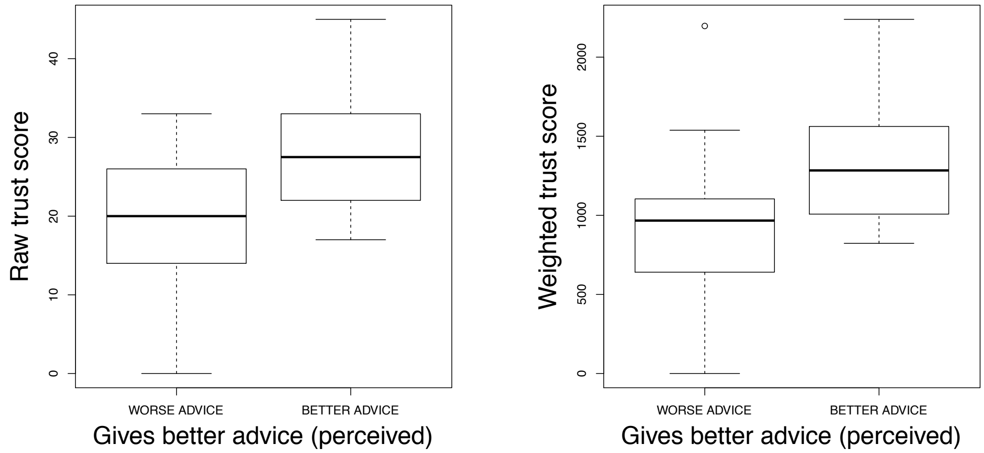
According to the post-interaction survey, subjects' perception of which avatar gave better advice was not related to whether an avatar entrained (Table 15.1). Instead, it was biased in Bobby's favor, although this difference was not significant ($X^2(1) = 2, p = 0.16$). The subjects asked for advice significantly more often from the helper whom they perceived as giving better advice, especially towards the end of the game (Figure 15.3) (raw: $t(15.37) = 2.81, p = 0.013$; weighted: $t(15.66) = 3.15, p = 0.0063$).

However, although there is no *explicit* relationship between entrainment and trust, the *implicit* measures of trust were affected by the entrainment condition. The entraining avatar's raw trust score was higher than the control's trust score for 12 out of 19 sessions, and the weighted score was higher for 14; on average, the raw trust score was approximately 6.2 points higher for the entraining avatar (standard deviation of the pairwise difference = 14.96), and the weighted trust score was approximately 253.42 points higher (standard deviation = 730.50). These differences approach significance for the raw score ($t(18) = 1.81, p = 0.087$) and fail for the weighted score ($t(18) = 1.51, p = 0.15$).

A linear model with trust score as the dependent variable and avatar condition (entrain vs. control) and persona (Bobby vs. Freddy) as the independent variabels further explicates the relationship between trust, entrainment, and the pro-Bobby bias. The "control" condition was a significant predictor of both a lower raw trust score ($\beta = -12.00, p = 0.00097$)

	Entraining avatar	
Better advice	*Bobby*	*Freddy*
Bobby	6	6
Freddy	3	3

Table 15.1: Which avatar was described as giving better advice.



(a) Raw trust score

(b) Weighted trust score

Figure 15.3: Trust scores related to subjects' perceptions of which avatar gives better advice.

and lower weighted trust score ($\beta = -488.50, p = 0.022$). The interaction of condition and avatar persona was significant in the raw score model ($\beta = 12.22, p = 0.016$), with the Bobby persona contributing to the raw score in the control condition, but not in the weighted score model ($\beta = 496.28, N.S.$). The model fits were $r^2 = 0.29$ and $r^2 = 0.18$ for the raw score and weighted score models, respectively.

The discrepancy between explicit and implicit measures suggests that entrainment can work as a *subconscious* force to influence the player to ask the entraining avatar's advice, even when she *consciously* trusts the other avatar more. Other unconscious biases, such as the pro-Bobby bias, can have an even stronger effect, although in our data the entrainment bias becomes stronger as the interaction develops.

	Entraining avatar	
Liked more	*Bobby*	*Freddy*
Bobby	8	3
Freddy	1	6

Table 15.2: Which avatar the subject preferred.

	Entraining avatar	
Liked voice better	*Bobby*	*Freddy*
Bobby	7	4
Freddy	2	5

Table 15.3: Which avatar's voice the subject preferred.

15.2.3 Entrainment and liking

While entrainment's influence on trust was subconscious, subjects explicitly liked the entraining avatar better. Table 15.2 shows that although some bias did exist in favor of Bobby here as well, the subjects usually chose the entraining avatar as the one they preferred; this difference approaches significance ($X^2(1) = 3.74, p = 0.053$). This result mirrors the well-documented finding that humans prefer other *humans* who entrain to them [Chartrand and Bargh, 1999; Ireland *et al.*, 2011; Manson *et al.*, 2013; Street, 1984] and supports the hypothesis that the benefits of entrainment observed in human-human conversations will apply in human-computer interactions as well.

15.2.4 Entrainment and voice descriptions

As stated before, bias existed with respect to the two avatars: when subjects were asked which speaker's voice they preferred, 11 out of 18 chose Bobby. However, of the 7 who preferred Freddy's voice, 5 had him as the entraining speaker (Table 15.3). This distinction is not statistically significant ($\chi^2(1) = 0.94, p = 0.33$), but it is consistent with our hypothesis that subjects would prefer the entraining voice.

In addition to the forced-choice question of preference, subjects were asked to indicate

	Entraining avatar	
Natural	*Entraining*	*Control*
TRUE	7	2
FALSE	12	17

(a)

	Entraining avatar	
Strange	*Entraining*	*Control*
TRUE	5	14
FALSE	14	5

(b)

Table 15.4: Perceived naturalness and strangeness for entraining and control voices.

traits that they felt were descriptive of Bobby's and Freddy's voices, respectively. We hypothesized that players would be more likely to view the entraining avatar's voice as natural (not strange) and pleasant (not annoying). We tabulated the counts for how often each kind of avatar was associated with a given descriptor. Each of the 18 subjects who completed the post-interaction survey evaluated an entraining voice and a control voice, for a total of 36 voice descriptions.

Naturalness. Only nine of the 36 voices were described as "natural." Of these, seven belonged to entraining avatars (Table 15.4a); this difference is consistent with our hypotheses but not significant ($\chi^2(1) = 2.33, p = 0.13$). However, 19 voices were described as "strange", and 14 of these belonged to control avatars (Table 15.4b). This difference is significant according to a chi-squared test ($\chi^2(1) = 6.74, p = 0.0094$).

Pleasantness. Only one control voice was considered "pleasant"; nine entraining voices were ($\chi^2(1) = 6.65, p = 0.099$). No entraining voice was considered "annoying"; nine control voices were ($\chi^2(1) = 9.32, p = 0.0023$) (Table 15.5). The descriptors of Freddy's voice were particularly interesting: Six of the nine subjects who had him as the control voice described his voice as "annoying", and none of the subjects to whom he entrained thought so.

Prosody. We also asked subjects to describe the prosodic characteristics of the voices, with the goal of determining which features had been most salient to them. The subjects' perceptions of rate were consistent with the voices' actual acoustic-prosodic characteristics: voices perceived as "fast" were faster ($t(142.57) = 3.67, p = 0.00034$), and voices perceived as "slow" were slower ($t(504.68) = -4.27, p = 2.3e - 05$). Volume, however, was perceived less accurately: both voices perceived as "loud" and those perceived as "soft" were in fact

	Entraining avatar	
Pleasant	*Entraining*	*Control*
TRUE	9	1
FALSE	10	18

(a)

	Entraining avatar	
Annoying	*Entraining*	*Control*
TRUE	0	9
FALSE	19	10

(b)

Table 15.5: Perceived pleasantness and lack thereof for entraining and control voices.

louder (loud: $t(80.42) = 1.82, p = 0.07$; soft: $t(238.17) = 2.14, p = 0.03$). The mean volume of "soft" voices (61.14 dB) is approximately equal to the mean volume of "loud" voices (61.36 dB).

There was no difference between entraining and control voices in whether the voice was perceived as "soft" or "fast" (soft: $\chi^2(1) = 0.70, p = 0.40$; fast: $\chi^2(1) = 1.78, p = 0.18$), but the control voices were slightly more likely to be described as "loud" ($\chi^2(1) = 3.68, p = 0.055$) and less likely to be described as slow ($\chi^2(1) = 3.96, p = 0.047$). Control voices were in fact 0.7 syllables per second faster on average than the entrained voices ($t(26.53) = -9.68, p = 3.4e - 10$), but the difference in volume means between entraining and control voices was not significant ($t(19.74) = -0.32, p = 0.75$). It is reasonable that subjects were less likely to perceive the entraining voice's volume, which mirrored their own, as unusual.

15.3 Discussion and future work

The game discussed in this chapter, GoFishWithHelpers, provides a method for data collection that implicitly measures a subject's trust in a given conversational avatar, and explicitly ask the subjects for their impressions of each avatar as well. We compare how subjects relate to an avatar whose voice entrains to the volume and speech rate of the user's preceding turn with how they relate to an avatar that varies its acoustic-prosodic features within a natural range.

We conclude that human users rely on an avatar more when it entrains to their acoustic-prosodic characteristics. This finding, in line with previous work that has shown that humanizing a robot promotes trust [Hancock *et al.*, 2011; Waytz *et al.*, 2014], has profound implications for the important task of facilitating interactions between humans and autonomous agents.

In addition, we show that human users like an avatar better when it entrains, prefer its voice, and are less likely to consider its voice "annoying" or "strange." These findings validate the central assumption of this study, which is that humans will relate to a computer that entrains in a manner comparable to how they relate to a human who entrains.

Global and local entrainment were coupled in the implementation of entrainment used in this study; in future work, we intend to explore how other entrainment strategies such as global and local convergence may interact with a human user's trust in the system and perceptions of it. In addition, we entrained on both volume and speaking rate in this study; future studies should look at entrainment on different features separately, as we have shown that in human-human conversations social features relate to entrainment on some features but not others.

In another direction for future work, we intend to look at the results of the personality test each subject completed in order to explore how a person's personality traits may affect his or her reaction to an entraining voice response system. For example, someone with a high score for "Openness" may have less trouble interacting in an unfamiliar way, while an "Anxious" person may benefit more from the availability of a human paradigm for the interaction.

The results of this study suggest that a system designer can incorporate acoustic-prosodic entrainment into a system's behavior in order to achieve a gain in user perception of the system's quality that is orthogonal to improvements in the ASR, TTS, dialogue management, other core components, or even the system's actual performance quality. The motivation to continue this line of research, especially the effort to improve and test the entrainer as described in the previous chapter, is therefore extremely strong.

Chapter 16

Conclusions and Future Work

Entrainment has been extensively studied in both human-human and human-computer conversations, and the link between entrainment and dialogue quality is well-documented. This link has recently motivated the development of several methods for incorporating entrainment behavior into natural language generation, resulting in systems that can entrain to the lexical and syntactic content of user utterances, as human interlocutors do. However, no similar methods have been proposed for implementing acoustic-prosodic entrainment. The problem is more difficult, because it does not involve discrete events whose probability can be manipulated, and because of the complications inherent in processing the audio signal.

This work fills a gap in the literature by proposing a method for implementing acoustic-prosodic entrainment. A system implementing this method can dynamically entrain to multiple acoustic-prosodic features of each user utterance, as humans do. Using this method, it can be simple to incorporate acoustic-prosodic entrainment into the behavior of an existing system.

Limitations of the method include its sensitivity to feature extraction errors. In addition, its greatest strength — the employment of SSML markup to manipulate the prosody of the output utterance in a manner native to the TTS, involving minimal additional latency — makes it dependent on architecture-specific implementations, which may result in output whose quality is less than optimal. Future work should explore the strengths and weaknesses of this method as opposed to transforming the output utterance using an audio processing tool such as Praat, which removes this dependence but may add unacceptable latency.

Based on the literature associating entrainment in human conversations with liking, we tested the hypothesis that human users would prefer an avatar that entrained to their prosody to one that randomly varied its prosody within a natural range. Additionally, since the literature suggests that more human-like behavior makes humans more likely to trust an autonomous agent, we designed the data collection method to implicitly model trust by requiring subjects to rely on the avatar of their choice to help them reach their game objectives. We found that subjects did in fact prefer the entraining avatar, used more positive adjectives to describe its voice, and were more likely to ask its advice. These findings strongly motivate the incorporation of acoustic-prosodic entrainment into the behavior of interactive voice systems.

The most important direction for future work is to test the utility of entrainment in a live spoken dialogue system with real users. Doing so will inform future development of the entrainment method to robustly handle input over a noisy channel, and investigate whether entrainment will play as large a role in the perception of dialogue quality when users have real objectives, in conversations likely to be significantly shorter.

The method as described here entrains to the user both globally and locally. Another interesting direction for future work is to implement other entrainment behaviors, such as global similarity, local similarity, synchrony, and global and local convergence, individually and in combination, as well as less aggressive or exact entrainment, and testing their utility in a study similar to the one described here. A system implementing this method can serve as a platform for testing hypotheses relating to entrainment, such as whether a user will prefer a system that becomes better at entraining later in the interaction or remains constantly similar throughout.

Part III

Conclusions

Chapter 17

Conclusions

Entrainment, one of the most pervasive and fascinating phenomena of human interaction, has been extensively studied but is not yet well understood. A more complete understanding of entrainment is necessary in order to accurately model human conversation, since it affects the expression of language at every level of communication. Furthermore, it is desirable to incorporate entrainment into a system's design, since a voice that does not entrain can never be perceived as truly natural.

In this thesis, we present a broad, multidimensional study of acoustic-prosodic entrainment that proposes and employs a consistent framework for studying entrainment and introduces novel dimensions to the body of research on entrainment. In addition, we present a method for implementing acoustic-prosodic entrainment in an interactive voice system, as well as a study that explores its utility.

The research presented in this thesis contains six major contributions:

- **A multi-dimensional study of entrainment.** The primary contribution of this thesis is its breadth, involving eight acoustic-prosodic features, six entrainment conceptualizations (global, local, exact, relative, static, convergent), three gender groups, and two languages. We make explicit the conceptualizations underlying various methods for measuring entrainment and look for evidence of entrainment according to each one, yielding a coherent, multidimensional analysis of acoustic-prosodic entrainment in task-oriented human-human conversation. Our approach builds a comprehensive understanding of the phenomenon and provides a framework for analysis that can influence the design and interpretation of future studies of entrainment.

- **Entrainment in Mandarin Chinese.** Our study of acoustic-prosodic entrainment in Mandarin Chinese — conducted jointly with Zhihua (Shirley) Xia — is, to our knowledge, the first to investigate entrainment in that language.

- **Cross-linguistic comparison of entrainment behaviors.** In addition, that study provides the first cross-linguistic analysis of acoustic-prosodic entrainment. We employ a consistent framework to explore entrainment in Mandarin Chinese in a way that allows for meaningful comparison with entrainment in Standard American English. Doing so exposes similarities and discrepancies in the two languages' respective patterns of entrainment behavior that have the potential to contribute empirical evidence to theories regarding entrainment. Similarly, we present the first cross-linguistic study of entrainment and gender, which provides a systematic comparison of the relationship between entrainment behavior and gender for each language.

- **Entrainment on backchannel-inviting cues.** Our study of entrainment on backchannel-inviting cues introduces the idea of entrainment on complex turn-taking cues. By showing that interlocutors entrain in this dimension, this work motivates further research into similar features, such as other turn-taking cues, questions, backchannel responses, and repairs. In addition, our finding that speakers are more likely to *respond* to the backchannel-inviting cues on which they entrain supports the theory that perception in this dimension of communication is related to production.

- **Entrainment on outliers.** Our study of entrainment on outliers contributes a novel perspective to the entrainment literature. We show that speakers entrain more to outlier values for some features, but not others. Assuming that outlier status can be used as a proxy for perceptual salience, our results suggest that increased perception is related to increased entrainment for intensity, voice quality, and speech rate, but not for pitch. This result supports theories that argue for a more nuanced relationship between perception and production. The knowledge that speakers are more likely to entrain on outliers may help resolve conflicting results in some studies that observe entrainment in certain cases — which may be outliers — but not others.

- **Implementation of acoustic-prosodic entrainment.** We present the first implementation of a system that can entrain to multiple acoustic-prosodic features of a human user's input, as a human interlocutor would. Our method is flexible, lightweight, and can be incorporated into an existing system with minimal configuration. It can be used to improve utterance generation in dialogue systems, or as a platform for testing hypotheses about entrainment in a controlled manner. In addition, we present a study showing that human users prefer an entraining avatar and rely on it more than one that does not entrain.

17.1 Significant observations

Several important observations emerge from this work:

- **Intensity.** Throughout the studies of human-human acoustic-prosodic entrainment described in this thesis, intensity (considering intensity mean and max together) is the feature for which evidence of entrainment is the most consistently observed. It is the only feature for which speakers are more similar to their interlocutors than they are to themselves in another conversation. Among speakers of Standard American English (SAE), it is the only feature on which speakers entrain both globally and locally; among speakers of Mandarin Chinese (MC), speakers entrain both globally and locally only on intensity and speech rate. For both SAE and MC, it is the feature on which interlocutors synchronize with each other most strongly. In SAE, it is the only feature on which female-female, male-male, and mixed-gender groups all entrain, and in MC, it is shows evidence of entrainment for both female-female and mixed-gender pairs.

  It is the only feature for which speakers entrain *more* to interlocutors for whom it is an outlier. Locally, it is one of two features for which speakers entrain more to *turns* for which it is an outlier. Finally, it is correlated with most coordination measures for every gender group.

Several hypotheses for intensity's prominence suggest themselves. It may be more perceptually prominent to the listener, or more accessible to manipulation at the motor level. It is possible that as in Natale's original model ([Natale, 1975]), it is more likely that a speaker's intensity level is the one he wishes his interlocutor to use, while pitch, for example, does not carry any such implication. Another possibility is that numerous intensity values are acceptable in a given context, while other features may be more constrained by pragmatics; this is unlikely, however, because intensity does play a strong pragmatic role. It is even possible to speculate that entrainment on intensity is so prevalent *because* it is associated with dialogue coordination. These hypotheses can be tested in perception and production experiments in which listeners are asked to describe and reproduce sounds or changes in sounds, as in [Kraljic *et al.*, 2008; Kreiman and Gerratt, 2005].

Since intensity shows evidence of entrainment in nearly every experiment, it is also useful to consider the cases in which it does not. It is not one of the features that show global convergence in SAE or local convergence in MC (no features converge globally in MC). Interlocutors do not entrain on their realization of intensity as a backchannel-inviting cue, and their entrainment on the use of intensity as a cue is not particularly strong. It is reasonable that the same factors contributing to the consistency of entrainment on intensity will also cause it to occur early enough in the interaction for convergence to have no effect. With respect to backchannels, it is interesting to note that these factors do not seem to apply when intensity is combined with other features to serve as a complex, higher-level feature. This may indicate that perception and production depend on the level of representation being activated.

Many of these observations also apply to speech rate, which is also prominent among our results.

- **Similarity between SAE and MC.** In our comparison of entrainment behaviors in SAE and MC, we observe striking similarities. Speakers of both languages entrain globally on intensity and speaking rate, entrain locally on intensity, entrain in synchrony for all features except speaking rate, and converge locally on pitch. For both languages, mixed-gender groups show evidence of entrainment on the greatest number of features, and male-male pairs the fewest. The prominence of intensity, and to a lesser extent speaking rate, is consistent across the two languages. These consistencies encourage us to hypothesize that these findings may extend to other languages as well.

We observe differences in entrainment behavior patterns between the two languages as well, most notably in the strength of synchrony. These findings suggest that for some features — such as intensity and speaking rate — entrainment may be more automatic, while for others — such as pitch — it might depend more on language-specific prosody and culture-specific social processes.

- **Differences in entrainment behavior between features.** The complement of the consistencies in our results is the set of differences. Intensity, for example, shows global and local similarity, moderately strong synchrony, no global convergence, and weak local convergence, while pitch shows no global or local similarity, and very weak synchrony, but it does show global convergence and slightly stronger local convergence, and speech rate shows global but not local similarity, no synchrony, and global but not local convergence. In addition, we observe differences in entrainment behavior based on the respective genders of the speaker and interlocutor. These differences show that all studies of entrainment must be interpreted in the context of the feature, method, and type of dialogue involved, since conclusions based on results from one experiment are likely to not apply to another. This observation motivates the necessity for a consistent framework for studying entrainment, as proposed by this thesis.

- **Automatic vs. social.** Several of our results are especially significant to the question of whether entrainment is automatic or pragmatic. We have shown that for some gender groups (but not others), entrainment on some features (but not others) is associated with some descriptors of pro-social behavior (but not others), indicating that entrainment is mediated by social goals — to some extent. With regard to perception, we have shown that for some features, people entrain more strongly to outlier values (which we assume to be more perceptually salient). We have also shown that speakers are more likely to respond to certain cues with a backchannel if they entrain on that cue, which shows that entrainment is associated with perception.

However, it is difficult to interpret any result as unambiguous support of either side of the debate, since in general, acoustic-prosodic events that are more perceptually salient carry stronger social signals, and interpersonal dynamics that motivate the reduction of social distance (such as liking or subordinate status) usually also entail greater attentiveness to the interlocutor's behavior. Future work may benefit from data collection methods that manipulate social goals in order to expose how they interact with entrainment behavior.

17.2 Future work

We have discussed possibilities for future work throughout the thesis. Here, we describe several directions that arise from the thesis as a whole.

- **Real users.** Although our data does approximate the kind of conversations in which we are most interested — task-oriented conversations between strangers — it was collected in a laboratory setting, with participants wearing head-mounted microphones in a soundproof booth. Future work should validate these findings in real-world situations, where social goals may operate differently and data is likely to be noisy.

- **Diverse subjects.** Since many of the participants in the Games Corpus were recruited from the Columbia University population, almost all were young and of relatively high socio-economic status. A very relevant direction for future work is to investigate how entrainment behavior may differ and how users' interactions with an entraining system may differ in other populations, such as people who are elderly, unfamiliar with computers, or on the autism spectrum. Similarly, we intend to investigate how different people's personalities affect how they perceive an entraining avatar.

- **Complex entrainment measures.** Throughout this work, we treat each entrainment measure as a separate phenomenon. We have shown that each can occur independently: local similarity without global similarity, global convergence without global similarity, synchrony on intensity without synchrony on speaking rate. However, the different aspects of an individual's entrainment behavior are likely to affect each other. Future work should explore interactions between these aspects, which may also serve to identify individual differences — are some people entrainers? Speech rate entrainers? Convergers? This work can yield a single metric to capture multiple aspects of entrainment behavior, so that the entrainment of a conversation can be characterized by a single measure (as in [Lee *et al.*, 2014]).

Furthermore, the measures of entrainment described here apply only to dyadic conversations. However, many interesting kinds of conversations involve multiple interlocutors, such as corporate meetings, online discussion forums, or Supreme Court proceedings. Extending these measures to capture *multiparty* entrainment is an interesting direction for future work.

17.3 Epilogue

In the first part of this thesis, we have investigated acoustic-prosodic entrainment in multiple dimensions and according to multiple conceptualizations and introduced several novel directions. In the second part, we have demonstrated that it is possible to incorporate this behavior into a spoken dialogue system, and that it is desirable to do so. The contributions of this work add considerably to our knowledge of human behavior and have the potential to improve spoken dialogue systems by incorporating that knowledge into their design, improving their naturalness and the user experience.

Part IV

Bibliography

Bibliography

[Babel, 2012] Molly Babel. Evidence for phonetic and social selectivity in phonetic accommodation. *Journal of Phonetics*, 40:177–189, 2012.

[Bateman, 2006] John A Bateman. A social-semiotic view of interactive alignment and its computational instantiation: A brief position statement and proposal. *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, page 157, 2006.

[Beattie, 1982] Geoffrey W Beattie. Turn-taking and interruption in political interviews: Margaret thatcher and jim callaghan compared and contrasted. *Semiotica*, 39(1-2):93–114, 1982.

[Bell *et al.*, 2000] Linda Bell, Johan Boye, Joakim Gustafson, and Mats Wirn. Modality convergence in a multimodal dialogue system. In *Proceedings of Gtalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue*, pages 29–34, 2000.

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

[Benus *et al.*, 2012] Stefan Benus, Rivka Levitan, and Julia Hirschberg. Entrainment in spontaneous speech: the case of filled pauses in supreme court hearings. In *3rd IEEE Conference on Cognitive Infocommunications, Kosice, Slovakia*, 2012.

[Beňuš *et al.*, 2014] Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia Hirschberg. Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, 2014.

[Bigi and Hirst, 2012] Brigitte Bigi and Daniel Hirst. SPeech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In *Speech Prosody*, pages 19–22. Tongji University Press, 2012.

[Bilous and Krauss, 1988] Frances R. Bilous and Robert M. Krauss. Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language & Communication*, 8(3/4):183–194, 1988.

[Blomgren *et al.*, 1998] Michael Blomgren, Yang Chen, Manwa L Ng, and Harvey R Gilbert. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *The Journal of the Acoustical Society of America*, 103(5):2649–2658, 1998.

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

[Bortfeld and Brennan, 1997] Heather Bortfeld and Susan E Brennan. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147, 1997.

[Bourhis and Giles, 1977] Richard Y Bourhis and Howard Giles. The language of intergroup distinctiveness. *Language, ethnicity and intergroup relations*, 13:119, 1977.

[Branigan *et al.*, 2000] Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25, 2000.

[Brennan and Clark, 1996] 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, 1996.

[Brennan and Hanna, 2009] Susan E. Brennan and Joy E. Hanna. Partner-specific adaptation in dialog. *Topics in Cognitive Science*, 1(2):274–291, 2009.

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

[Brennan, 1996] Susan E Brennan. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, pages 41–44, 1996.

[Broadbent *et al.*, 2013] Elizabeth Broadbent, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q Stafford, Bruce A MacDonald, and Daniel M Wegner. Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PloS one*, 8(8):e72589, 2013.

[Brockmann *et al.*, 2005] Carsten Brockmann, Amy Isard, Jon Oberlander, and Michael White. Modelling alignment for affective dialogue. In *Workshop on Adapting the Interaction Style to Affective Factors at the 10th International Conference on User Modeling (UM-05)*, 2005.

[Brown and Levinson, 1987] Penelope Brown and Stephen C. Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge University Press, 1987.

[Bruemmer *et al.*, 2004] David Bruemmer, Douglas Few, Michael Goodrich, Donald Norman, Nilanjan Sarkar, Jean Scholtz, Bill Smart, Mark L Swinson, and Holly Yanco. How to trust robots further than we can throw them. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, pages 1576–1577. ACM, 2004.

[Buschmeier *et al.*, 2009] Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. An alignment-capable microplanner for natural language generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 82–89. Association for Computational Linguistics, 2009.

[Carlson *et al.*, 2006] Rolf Carlson, Jens Edlund, Mattias Heldner, Anna Hjalmarsson, David House, Gabriel Skantze, and CSC CTT. Towards human-like behaviour in spoken dialog systems. In *Proceedings of Swedish Language Technology Conference*, 2006.

[Chan, 1998] Marjorie KM Chan. Gender differences in the Chinese language: A preliminary report. In *Proceedings of the Ninth North American Conference on Chinese Linguistics*, volume 2, pages 35–52, 1998.

[Chao, 1995] Fang-yi Chao. On the gender-marked pronoun 'renjia' in Chinese. Unpublished manuscript, OSU, 1995.

[Chartrand and Bargh, 1999] 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.

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

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

[Danescu-Niculescu-Mizil *et al.*, 2012] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: language effects and power differences in social interaction. In *Proceedings of WWW*, 2012.

[de Jong *et al.*, 2008] Markus de Jong, Mariët Theune, and Dennis Hofs. Politeness and alignment in dialogues with a virtual guide. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 207–214. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[Edlund *et al.*, 2009] Jens Edlund, Mattias Heldner, and Julia Hirschberg. Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech*, 2009.

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

[Farris, 1995] Catherine Farris. *A semeiotic analysis of sajiao as a gender marked communication style in Chinese*, pages 1–29. U. Chicago Center for East Asian Studies, 1995.

[Friedberg *et al.*, 2012] Heather Friedberg, Diane Litman, and Susannah BF Paletz. Lexical entrainment and success in student engineering groups. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 404–409. IEEE, 2012.

[Giles *et al.*, 1991] Howard Giles, Nikolas Coupland, and Justine Coupland. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1, 1991.

[Gnisci, 2005] Augusto Gnisci. Sequential strategies of accommodation: A new method in courtroom. *British Journal of Social Psychology*, 44(4):621–643, 2005.

[Goldinger, 1998] Stephen D Goldinger. Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105(2):251, 1998.

[Gosling *et al.*, 2003] S.D. Gosling, P.J. Rentfrow, and W.B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

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

[Gravano *et al.*, 2011] 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.

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

[Gregory *et al.*, 1993] Stanford Gregory, Stephen Webster, and Gang Huang. Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language & Communication*, 13(3):195–217, 1993.

[Gustafson *et al.*, 1997] Joakim Gustafson, Anette Larsson, Rolf Carlson, and K Hellman. How do system questions influence lexical choices in user answers? In *Eurospeech*. Citeseer, 1997.

[Hancock *et al.*, 2011] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(5):517–527, 2011.

[Heldner *et al.*, 2010] Mattias Heldner, Jens Edlund, and Julia Bell Hirschberg. Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech*, 2010.

[Hu *et al.*, 2014] Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Proceedings of 5th International Workshop on Spoken Dialog System*, 2014.

[Hu, 1991] Mingyang Hu. Feminine accent in the beijing vernacular: a sociolinguistic investigation. *Journal of the Chinese Language Teachers Association*, 26.1:49–54, 1991.

[Huggins-Daines *et al.*, 2006] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[Ireland *et al.*, 2011] Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44, 2011.

[Kraljic *et al.*, 2008] Tanya Kraljic, Susan E. Brennan, and Arthur G. Samuel. Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107(1):54–81, 2008.

[Kreiman and Gerratt, 2005] Jody Kreiman and Bruce R Gerratt. Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*, 117(4):2201–2211, 2005.

[LaFrance, 1992] Marianne LaFrance. Gender and interruptions: Individual infraction or violation of the social order? *Psychology of Women Quarterly*, 16:497–512, 1992.

[LaFrance, 2001] Marianne LaFrance. Gender and social interaction. In R.K. Unger, editor, *Handbook of the psychology of women and gender*, pages 245–255. Wiley, 2001.

[Laursen, 2013] Lucas Laursen. Robot to human: "trust me". *Spectrum, IEEE*, 50(3):18–18, 2013.

[Leaper and Robnett, 2011] Campbell Leaper and Rachael D Robnett. Women are more likely than men to use tentative language, arent they? a meta-analysis testing for gender differences and moderators. *Psychology of Women Quarterly*, 35(1):129–142, 2011.

[Lee *et al.*, 2010] 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.

[Lee *et al.*, 2014] Chi-Chun Lee, Athanasios Katsamanis, Matthew P. Black, Brian R. Baucom, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech & Language*, 28(2):518 – 539, 2014.

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

[Levitan *et al.*, 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.

[Levitan *et al.*, 2012] 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.

[Light, 1982] Timothy Light. On being deing: how women's language is perceived in Chinese. *Computational Analyses of Asian & African Languages = Ajia Afurikago no keisu kenkyu*, (19), Jan 1982.

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

[Lopes *et al.*, 2013] José Lopes, Maxine Eskenazi, and Isabel Trancoso. Automated two-way entrainment to improve spoken dialog system performance. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8372–8376. IEEE, 2013.

[Manson *et al.*, 2013] Joseph H Manson, Gregory A Bryant, Matthew M Gervais, and Michelle A Kline. Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, 34(6):419–426, 2013.

[Matarazzo and Wiens, 1967] Joseph D. Matarazzo and Arthur N. Wiens. Interviewer influence on durations of interviewee silence. *Journal of Experimental Research in Personality*, 1967.

[McLemore, 1991] Cynthia A McLemore. *The pragmatic interpretation of English intonation: Sorority speech. University of Texas at Austin Ph. D.* PhD thesis, dissertation, 1991.

[McMillan *et al.*, 1977] Julie R McMillan, A Kay Clifton, Diane McGrath, and Wanda S Gale. Women's language: Uncertainty or interpersonal sensitivity and emotionality? *Sex Roles*, 3(6):545–559, 1977.

[Mertens, 2004] Piet Mertens. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Speech Prosody*, 2004.

[Metzing and Brennan, 2003] Charles Metzing and Susan E. Brennan. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213, 2003.

[Michael and Otterbacher, 2014] Loizos Michael and Jahna Otterbacher. Write like i write: Herding in the language of online reviews. In *Proceedings of the Eigth International AAAI Conference on Weblogs and Social Media*, 2014.

[Namy *et al.*, 2002] Laura L. Namy, Lynne C. Nygaard, and Denise Sauerteig. Gender differences in vocal accommodation: the role pf perception. *Journal of Personality and Social Psychology*, 21(4):422–432, 2002.

[Nass and Lee, 2001] Clifford Nass and Kwan Min Lee. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001.

[Nass *et al.*, 1994] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78. ACM, 1994.

[Natale, 1975] 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.

[Nenkova *et al.*, 2008] Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 169–172. Association for Computational Linguistics, 2008.

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

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

[Pickering and Garrod, 2004] Martin J. Pickering and Simon Garrod. Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

[Ranganath *et al.*, 2009] Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. It's not you, it's me: detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 334–342. Association for Computational Linguistics, 2009.

[Reitter and Moore, 2007] 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.

[Reitter *et al.*, 2010] David Reitter, Johanna D. Moore, and Frank Keller. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. 2010.

[Scherer, 1989] Klaus R Scherer. Vocal correlates of emotional arousal and affective disturbance. *Handbook of social psychophysiology*, pages 165–197, 1989.

[Shen, 1995] Haibing Shen. An analysis of *niang niang qiang*. Unpublished manuscript, OSU, 1995.

[Silverman *et al.*, 1992] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: A standard for labeling English prosody. In *International Conf. on Spoken Language Processing*, volume 2, pages 867–870, 1992.

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

[Street, 1984] Richard L Street. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169, 1984.

[Taylor and Thomas, 2008] Paul J Taylor and Sally Thomas. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, 1(3):263–281, 2008.

[Thomason *et al.*, 2013] Jesse Thomason, Huy V Nguyen, and Diane Litman. Prosodic entrainment and tutoring dialogue success. In *Artificial Intelligence in Education*, pages 750–753. Springer, 2013.

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

[Turner and West, 2004] Lynn H Turner and Richard West. *Introducing communication theory: Analysis and application, 4th Edition*. Mountain View, CA: Mayfield, 2004.

[Vullinghs *et al.*, 2013] Anne Vullinghs, Martijn Goudbeek, and Emiel Krahmer. Crosslinguistic priming in interactive reference: Evidence for conceptual alignment in speech production. In *Proceedings of Interspeech*, 2013.

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

[Ward and Nakagawa, 2004] Nigel Ward and Satoshi Nakagawa. Automatic user-adaptive speaking rate selection. *International Journal of Speech Technology*, 7(4):259–268, 2004.

[Ward and Tsukahara, 2000] Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207, 2000.

[Ward, 1996] Nigel Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1728–1731. IEEE, 1996.

[Waytz *et al.*, 2014] Adam Waytz, Joy Heafner, and Nicholas Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52(0):113 – 117, 2014.

[Wolk *et al.*, 2012] Lesley Wolk, Nassima B Abdelli-Beruh, and Dianne Slavin. Habitual use of vocal fry in young adult female speakers. *Journal of Voice*, 26(3):e111–e116, 2012.

[Xia *et al.*, 2014] Zhihua Xia, Rivka Levitan, and Julia Hirschberg. Prosodic entrainment in Mandarin and English: A cross-linguistic comparison. In *Speech Prosody*, 2014.

[Yuasa, 2010] Ikuko Patricia Yuasa. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, 85(3):315–337, 2010.

[Zhang and Kramarae, 2012] Wei Zhang and Cheris Kramarae. Are Chinese women turning sharp-tongued? *Discourse & Society*, 23(6):749–770, 2012.