Toward Improving Dialogue Coordination in Spoken Dialogue Systems

Rebecca Lunsford
B.S. Computer Science, National University, 1991
M.S. Computer Science & Engineering, OHSU, 2005

Presented to the
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science & Engineering

March 2012

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. dissertation of

Rebecca Lunsford

has been approved.

Dr. Peter Heeman, Thesis Advisor
Research Associate Professor

Dr. Jan van Santen
Professor

Dr. John-Paul Hosom
Staff Research Technologist, Sensory, Inc.

Dr. Michael Johnston
Principal Technical Staff, AT&T Labs Research

Contents

List of Tables

List of Figures

Abstract

Toward Improving Dialogue Coordination in Spoken Dialogue Systems

Rebecca Lunsford, M.S.

Doctor of Philosophy
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine

March 2012
Thesis Advisor: Dr. Peter Heeman

When engaged in a conversation, speakers use both verbal and non-verbal mechanisms to help coordinate the dialogue, ensuring that, at each point, the other is engaged in the dialogue, and is capable of hearing, understanding and responding to the speaker. The problem is that current Spoken Dialog Systems (SDSs) do not take full advantage of dialogue coordination mechanisms, which can lead to interactions that are unnatural and inefficient. However, we posit that an SDS should anticipate, recognize and potentially emulate the full richness of dialogue coordination mechanisms. In this dissertation research, we aim to further understand dialogue coordination mechanisms, and to assess how they might be used to ease human-computer interaction. We start by investigating what cues a human speaker uses to differentiate computer-directed speech from self-directed speech, and from human-directed speech, finding that in both cases speech directed to the computer is much louder. We next conduct a perceptual study to determine what cues people attend to when determining whether a speaker is addressing a computer or nearby human. Here we found that people tended to rely on the direction of the speaker's gaze, although this led to systematic errors in their judgments of addressee. We next investigate whether

'um' and 'uh' result from the same, or different cognitive processes, using human-human interaction data collected while clinicians interacted with children with typical development, autism, or developmental language disorder. Here we found that 'um' appears to be listener-oriented, and 'uh' speaker-oriented. Next, again using the data from above, we investigated what factors impact the length of inter-turn gaps, and whether there is an interaction between gaps, disfluencies and social pressure to respond. Here we found that, after a question, speakers tend to respond more quickly, are more likely to start their speech with a disfluency, and that the likelihood of a disfluency increased with the length of the gap. Finally, we conduct a simulation study, using Reinforcement Learning, to demonstrate that dialogue policies can be created that take advantage of dialogue coordination mechanisms.

Chapter 1

Introduction

> "Natural communication is a social activity. Communicative behaviour therefore has to observe the norms and conventions of social activity in the culture to which the agents belong." *(Harry Bunt, 1997)*

The purpose of speaking is typically not the speech itself. People speak to accomplish some task or serve some goal. As such, speaking requires the speaker to manage two activities: the underlying task or goal, and the act of communicating itself. This second activity, that of managing the communication, is complex in itself. At each point in a conversation, speakers must ensure that their interlocutor is engaged, can hear and understand them, and is capable of, and willing to, respond. Listeners must attend to the speaker, interpreting the speaker's words, inferring the speaker's intent, signaling understanding (or lack of understanding) and planning a response. In addition, all of this is accomplished while the interlocutors coordinate speaking turns.

To coordinate a dialogue with others, participants employ a range of verbal, prosodic, and gestural mechanisms to help clarify their communicative intent. For example, a speaker can use rising pitch at the end of an utterance to both mark the preceding speech as a question and to signal a turn-release. Furthermore, a speaker could signal turn-assignment by looking at the intended responder or speaking their name. Likewise, a listener can nod and utter "uh-huh" to signal understanding or hold up a hand to signal the desire to interject. Dialogue mechanisms such as these allow participants to better infer their interlocutors intent and ease coordination.

As dialogue is a social activity, speakers conform to conventions when engaged in a

conversation. These conventions address both what to say so as to support the progress of the underlying goal or task, and how to coordinate the dialogue with others so as to ensure the smooth flow of dialogue and avoid problems such as misunderstandings, interruptions and un-intentional overlaps.

People follow these dialogue conventions with little effort. This is primarily because the mechanisms used to coordinate a dialogue are essentially automatic, both produced and interpreted without conscious planning. To illustrate, when speaking in the presence of loud background noise, speakers will engage in Lombard speech, even if they know their listener cannot hear the noise [45].

This automaticity, that speakers both produce and interpret dialogue coordination mechanisms without conscious effort, presents both a challenge and an opportunity for Spoken Dialogue Systems (SDSs). The challenge is that people cannot be expected to easily alter these behaviors, even if processing them proves challenging to an SDS. The opportunity is that these mechanisms should prove regular and readily interpretable, providing additional information about users that could lead to improved human-computer interaction.

However, consistent and regular use of a given mechanism does not, in and of itself, indicate that the mechanism is used by a speaker to coordinate the dialogue. Take the case of the fillers 'uh' and 'um', which can constitute as much as 9% of a person's speech [24]. Clark and Fox Tree [24] posited that 'uh' and 'um' are used by a speaker to signal a delay in his speech. If this position is correct, then fillers would clearly fall within the realm of dialogue coordination mechanisms. However, others contend that 'uh' and 'um' are the unintentional result of the speaker experiencing a speech production problem [51, 26]. If so, listeners might anticipate that a speaker would delay after a filler, but this anticipation would be based on inferring the speakers cognitive state, not because the speaker's filler was intended as a dialogue coordination mechanism. Thus, although fillers are both common and regular, there remains debate as to whether they are conventions.

1.1 Thesis Statement

Although human speakers readily produce and interpret dialogue coordination mechanisms, current Spoken Dialogue Systems (SDSs) tend to not use them. For example, SDSs typically require that the user explicitly engage the system prior to starting a dialogue, and assume that all subsequent speech is directed to the system. In contrast, human communication is rarely so rigid. Instead speakers employ one, or more, verbal, prosodic or gestural cues to imply their intent to engage in and continue a dialogue. The problem we address in this dissertation is that spoken human-computer interaction remains unnatural and inefficient, in part because of this disparity between humans and SDSs in their use, and understanding of, dialogue coordination mechanisms.

Because people produce, interpret, and respond to dialogue coordination mechanisms automatically and consistently, we hypothesize that many of these mechanisms could prove amenable to use by SDSs. The goal of this dissertation is to identify dialogue coordination mechanisms and assess how they can be used to improve human-computer interaction (HCI). From an HCI perspective, we anticipate three ways in which this work will impact SDSs.

- First, SDS designers can simply acknowledge that aspects of people's communication are realized through dialogue coordination mechanisms and design systems that react accordingly. For example, SDS designers could recognize that speakers typically identify to whom they are addressing a question or request, and build systems that respond only when being addressed.

- Second, SDSs can be designed to use dialogue coordination mechanisms that are appropriate for the current context. For example, an SDS could backchannel (e.g., "uh-huh") to indicate that it is able to understand the user's speech, but refrain when the message is unclear.

- Third, SDS designers could design systems that anticipate how the system's dialogue coordination mechanisms will impact a user's speech. For example, systems could be designed that anticipate longer inter-turn pauses when the user is asked a question,

or proactively adapt their actions so as to minimize a user's cognitive load.

1.2 Approach

In this dissertation, we use four different approaches. Each approach offers advantages (as described below) relevant to addressing the particular research goal.

First, to explore what mechanisms people use when addressing a computer systems, we analyzed data collected during a series of wizard-of-oz (WOZ) studies. In these studies people interact with simulated computer systems, in which a wizard (i.e., human researcher) interprets the user's speech and controls the system's feedback. By using a wizard, we can ensure that the user's behavior is not unduly influenced by speech recognition errors, and can instead proceed naturally [6]. By creating a WOZ system, which limits and directs what options the wizard has at any point, we can ensure the interaction remains computer-like.

Second, to explore how people interpret human cues of addressee, we conducted a perceptual study using recorded clips of multi-party human-computer interaction collected during a previous WOZ study. For this type of study, observers are asked to make judgments about the information presented, typically using forced-choice decisions. By using a perceptual study, we are able to determine what information leads human observers to correct versus incorrect judgments. For this work we coupled these perceptual judgments with a survey, thus allowing us to determine if participants are aware of which information they attend to when making these judgments.

Third, to explore whether certain aspects of dialogue are social in nature, we next analyzed data collected during human-human interaction. In these studies we contrast the dialogue behavior of participants without language impairments to participants with impairments in social language, and those with impairments in receptive and expressive language. By using data collected during human-human interaction, as opposed to human-computer, we can observe how both impaired and unimpaired speakers communicate naturally with a fully-skilled social agent.

Finally, to explore how SDSs might incorporate dialogue coordination mechanisms,

we trained dialogue policies using Reinforcement Learning (RL). By using RL to create dialogue policies using differing costs, we can compare how decisions about costs can affect the quality of interaction for users with differing abilities.

1.3 Dissertation Structure

Chapter 2 presents relevant background on dialogue coordination to provide a context for this research. Work relevant to specific chapters are included in those chapters.

In Chapters 3 and 4, we determine what dialogue coordination mechanisms an SDS might use to determine when it is being addressed. In Chapter 3, we show the high incidence of self-directed speech during human-computer interaction, and identify cues that do, and do not, differentiate self- from system-directed speech. In Chapter 4, we compare speech addressed to a computer to that addressed to a human cohort, and identify reliable cues which differentiate human- from system-directed speech.

Chapter 5 examines how people perceive other's cues of addressee. Using data collected in the previous study, this chapter explores whether the cues people use to identify a speaker's addressee are, in fact, those that speakers use when engaged in a multi-party human-human-computer interaction. By identifying any differences, SDS can be built that respond to the cues that are salient to human-computer, as compared to human-human, interaction.

In Chapters 6 and 7, we compare speech produced by children with Typical Development (TD), to those with Developmental Language Disorder (DLD), and Autism Spectrum Disorder (ASD). By contrasting the speech of children with TD to those with language processing impairments (i.e., ASD and DLD), and those with social impairments (i.e., ASD), we gain insights into what behaviors are common regardless of (dis)ability, or are impacted by processing or social impairments.

Chapter 6 explores two theories regarding *why* people produce the fillers 'uh' and 'um', and whether they are dialogue coordination mechanisms. One theory suggests that they are artifacts of a speaker's difficulties in processing, and the other suggests that they are signals intended to inform the listener of delay. In this chapter we compare the use of

fillers between children with TD, DLD and ASD to determine if filler usage is driven by cognitive processes responsible for language processing or social interaction.

In Chapter 7 we explore whether social pressure influences a speakers time to respond and likelihood of producing a filler or disfluency. To answer this question, we compare the inter-turn pauses of the three groups of children, specifically looking at how the timing of inter-turn pauses is impacted by the speaker's social (dis)abilities, processing (dis)abilities, and/or activity. In addition, we explore whether speakers are more likely to be disfluent when there is increased obligation to respond, such as when posed a question.

In Chapter 8, we explore whether Reinforcement Learning can be used to design dialogue policies that take advantage of dialogue coordination mechanisms to improve human-computer interaction.

Chapter 9 provides a summary of the work, draws conclusions and discusses the major contributions of this dissertation.

1.4 Contributions

1.4.1 Primary Contributions

In this dissertation, we identify dialogue coordination mechanisms that can be leveraged by SDSs to better meet user's needs and expectations. At a high level, we analyze dialogue coordination mechanisms in four ways, with each producing separate contributions. First, using *human-human* communication (including self-directed speech) as a baseline, we identify cues speakers use as dialogue coordination mechanisms during human-computer interaction, showing differences in speaker's dialogue coordination mechanisms when addressing a computer versus human. Second, we determine whether fillers are used by speakers as dialogue coordination mechanisms or are artifacts of language processing problems. Third, we investigate how interlocutors respond to dialogue coordination mechanisms, and whether the timing and fluency of the response is impacted by the speaker's social abilities. Fourth, we demonstrate that SDSs can be designed that take advantage of dialogue coordination mechanisms, adapting the system's behavior to better meet the needs of diverse users.

1.4.2 Secondary Contributions

Although the primary goal of this work is to identify dialogue coordination mechanisms and how they could be used in SDSs, this dissertation also contains secondary contributions relevant to other arenas of study. These include findings showing that:

- The fillers 'um' and 'uh' appear to arise from different cognitive processes, a finding relevant to socio-linguistics.
- Children with TD, ASD, and DLD differ in their inter-turn pauses, use of fillers, responsiveness to questions, and likelihood of turn-initial disfluencies, but that, for these measures, the children with DLD more closely resembled the children with TD. These findings are of interest to developmental psychology, as analyzing these speech aspects could aid in differential diagnosis.

In addition, this dissertation showcases a range of experimental approaches, illustrating how differing empirical methods can be used to explore the production, interpretation, and identification of dialogue coordination mechanisms.

Chapter 2

Background and Related Work

We start this review with background information that provides context for the studies included in this dissertation, specifically providing a more in-depth discussion of dialogue coordination. Related work relevant to the individual studies are included in the respective chapters.

2.1 Communicative Requirements

To engage in a dialogue, interlocutors must be both ready and able to communicate. Toward this end, Allwood describes four basic requirements of human dialogue: that the agents are capable of paying attention and willing to continue the dialogue; that the listener is capable of, and willing to, perceive what is being said; that the listener can understand what the speaker is attempting to convey, and that the listener is able and willing to respond [3]. The communicative functions used to address these requirements are referred to as contact, perception, understanding, and response (CPUR).

Allwood notes that:

> "Every language appears to have conventionalized means (verbal and prosodic means as well as body movements) for giving and eliciting information about the basic communicative functions." [3]

This quote makes two points of particular interest to this dissertation. First, that the mechanisms used to coordinate and manage dialogue are conventions. Thus, we can anticipate that they are driven by social norms and are likely to be both regular and

readily interpretable. Second, that to coordinate a dialogue, people produce, and look for, cues consisting of verbal, prosodic and bodily movements. From the perspective of SDSs, these two points suggest: 1) that these dialogue coordination mechanisms lend themselves to recognition, and that 2) the mechanisms may be realized through non-linguistic methods [23].

To ensure that CPUR is in place, participants must also attend to the biological and cognitive constraints on their interlocutor, as well as any demands related to the physical environment and communication channel [3, 21]. For example, Bunt discusses that during face-to-face communication, the need for "contact" is likely to be addressed using visual cues, such as eye-contact [20], but that when communicating on the telephone, participants may need to say something (e.g., "Hello?") to ensure continued contact. Bunt termed this *contact management*, which arises "because the perceptual and physical context is such that the speaker is in doubt as to whether he is currently in contact with his partner." [21].

The work reviewed in this section speaks to the most basic requirements of dialogue coordination, that of ensuring both parties are engaged. However, this work does not describe the mechanisms used except at a conceptual level. To design an SDS that can fully participate in dialogue coordination, more information is needed as to how people produce, and respond to, dialogue coordination mechanisms.

2.2 Communication Management

Having met the basic requirements for communication (i.e., having ensured CPUR), participants must also manage, and coordinate, the dialogue itself. Allwood describes *own communication management* (OCM) and *interaction communication management* (IACM). OCM are those mechanisms that allow a speaker to manage his own communication in regards to timing, processing, and change. IACM are those mechanisms used to manage the dialogue flow, specifically addressing sequencing, turn-taking and feedback (related to CPUR) [3]. However, from this description, it is unclear is as to exactly what mechanisms fall into each of these categories. For example, if a speaker utters "um" at the beginning

of their turn, are they engaging in OCM (i.e., buying themselves some time), or IACM (e.g., indicating uncertainty about the subsequent speech)?

Bunt uses the term "dialogue control" and a slightly different framework than that used by Allwood, to describe the management aspects of dialogue. In this framework, there are three major components of dialogue control; feedback, interaction management and social obligation management [21]. In this work Bunt incorporates Allwood's concept of OCM, and views it as a mechanism for providing information about the speakers' processing state. From Bunt's viewpoint, self-corrections indicate processing difficulties, and fillers a need for additional time. In Bunt's work, as in Allwood's, it is unclear whether mechanisms such as self-corrections and fillers are produced to inform the listener, or are merely artifacts of utterance production, used opportunistically by listeners to infer the speaker's processing state. From the perspective of an SDS, understanding the difference between these two will be important to correctly interpreting these cues when produced by a user and to producing these cues in such a way as to avoid user confusion.

2.3 Questions and Social Pressure

Bunt [21] also discussed the concepts of *reactive pressure* (RP) and *interactive pressure* (IP). RP describes the situation in which a speaker utterance places "pressure" on an interlocutor to respond with a certain type of utterance (e.g., "Thank you." – "You're welcome"). IP describes pressure on a participant to perform a certain action based on the context. Examples of IP include answering a phone with "Hello", producing back-channels to indicate understanding, or responding to a direct question. In earlier work, Bunt [20] suggests that, for educational interfaces, the interface must react appropriately to IP and RP (e.g., producing farewells), so that users can communicate in a manner they find natural. Yet, it seems that an SDS should also account for how IP conferred on the user by the system (e.g., by asking the user a question) might impact the user's response.

One area of particular interest is that of questions, and the IP associated with them. This pressure to respond to questions has been described in other works as an obligation conferred on, and adopted by, the listener [36, 2, 98] or as the second half of an *adjacency*

pair [82]. Regardless of the naming convention, there is general agreement that, having been queried, there exists pressure on the listener to respond to that query, and that, in the vast majority of situations, the listener does.

2.4 Turn Management

One area of particular interest, in terms of dialogue coordination, is that of turn-management. Sacks et al. [82] describes a model in which the a speaker, using both verbal and nonverbal cues, indicates a transition-relevant place (TRP) in the dialogue, at which a turn-transition can occur. In this model, there a number of rules that determine who will take the next turn. These rules are primarily contingent on the speaker who, by the construction of the turn-so-far, indicates whether another speaker is obligated to take the turn, or can choose to take the turn. Although this work was primarily concerned with how turn-transitions are managed, it is clear that speakers not only indicate when a turn transition is intended, but also indicate when no turn-transition is intended.

In addition to creating a model that describes how turn transitions are managed, Sacks et al. [82] also posit that, at turn transitions, speakers strive to avoid overlapping the preceding speaker, and to minimize silent gaps. Clark [23], agreed with Sacks et al., but suggested that the need to minimize the gap might be relaxed somewhat when it is likely that a responder needs processing time to understand the preceding speech, or to plan and organize a response. Smith and Clark [88] suggest that responders manage, and account for, gaps by producing cues signaling delay, such as the fillers ‘um’ and ‘uh’. From this work, it seems that there might be be an interaction between gaps and the content of the response.

2.5 Dialogue Coordination and SDS

To date, SDSs have generally used system-centric approaches to dialogue coordination, expecting that users will adapt to meet the system’s needs. In terms of contact, SDSs typically require the user to perform some explicit action to engage the system (e.g., place a phone call, press a button, or speak a keyword), and assume that all subsequent speech

is addressed to the system. Users are also expected to manage the communication channel (e.g., volume settings) to ensure they can hear the system. As user confusion or uncertainty can derail an interaction, system prompts are designed to ease user understanding, primarily with the intent of fostering user speech that the system can easily recognize and process [72, 49]. However, it seems likely that by taking advantage of the dialogue coordination mechanisms people use naturally, SDSs can be designed that are more user-centric and can adapt to user's needs.

Chapter 3

Self- versus System-directed Speech

In this chapter, we investigate self- versus system-directed speech. Self-directed speech is theorized as resulting from complex mental reasoning, and so when it does occur, it would be best for the computer to not interrupt. Self-directed speech is related to the notion of contact that Allwood [3] addressed, as during self-directed speech, it can be argued that contact is not being made between the user and the system. It is also related to turn-taking, as one could argue that in the model of Sacks et al. [82], the user does not intend the system to take the turn. From an HCI standpoint, we anticipate that there are reliable cues that a computer can use to determine whether it is being addressed, and should respond, or the user is talking to himself. In particular, we examine gaze and speech amplitude[1] as cues, as previous studies have found them to be useful.[2]

The primary focus of the present research is to establish better empirically-grounded models for distinguishing when users are addressing a computer. In particular, users' audio-visual activity patterns are examined when they are, and are not, addressing a system during human-computer interaction. In addition, this research explores self-directed speech as a user-generated source of noise in order to identify characteristics of self-directed speech that differentiate it from system-directed speech. The impact of age also is assessed on the presence of self-directed speech and the magnitude of users' amplitude separation during self- versus system-directed speech. Towards this end, corpora from two related

[1] What we refer to here as "amplitude" is commonly referred to as "intensity" in speech technology literature.

[2] Our expectations regarding amplitude are based on preliminary work published in the abstract *"Private Speech during Multimodal Human-Computer Interaction"* [56]

studies were analyzed in which participants interacted with a map-based interface using speech and pen input. The first study involved elderly users [107, 56], and the second compared younger and elderly adults. The specific goals of this research are the following:

- Compare rates of self-directed speech for younger and elderly adults, with the elderly users expected to engage in more self-directed speech due to declining working memory and hearing,
- Examine potential amplitude differences between users' self-directed speech and system-directed speech for a wide range of users representing different ages, with self-directed speech expected to be lower in amplitude,
- Explore the possibility of age-related differences in the magnitude of amplitude separation during self- and system-directed speech for younger versus older adults, and
- Compare gaze directed at the system during self- versus system-directed speech.

Finally, this research aims to evaluate the relative power of gaze and amplitude cues to reliably discriminate when users' speech is and is not addressed to the system during human-computer interaction.

The work in this chapter is based on an earlier work [61]: Audio-visual cues distinguishing self- from system-directed speech in younger and older adults, in *Proceedings of the 7th international conference on Multimodal interfaces* (ICMI '05). http://doi.acm.org/10.1145/1088463.1088494. [3]

3.1 Background and Related Work

3.1.1 Motivation

In pursuit of more natural computer interfaces for humans, researchers are experimenting with alternate interface mechanisms. The combination of multiple inputs modes such

[3]The number of elderly subjects in Study One who produced self-talk was previously reported in *"Modeling Multimodal Integration Patterns and Performance in Seniors: Toward Adaptive Processing of Individual Differences"* [107], in which I was the second author. My contributions to that work includes all analyses related to self-directed speech, referred to in that work as self-talk.

as speech with pen, facial gesture or other physical tools is a common theme within human-computer interface research. However, as interfaces become more human oriented, humans are, not surprisingly, behaving more naturally. One behavior recently observed within two multimodal interface studies was that of speech not directed to the system, but instead spoken solely for the speaker [107, 15]. Although defined slightly differently in each study, the underlying component of each is a well-known phenomenon, that of self-directed speech (typically referred to as private speech or self-regulatory speech within psychology literature).

Humans can resolve the difference between someone talking to them and someone talking to themselves, either by understanding verbal and non-verbal cues, or simply by asking the speaker. Current spoken dialogue systems do not make this differentiation. One can easily picture the following scenario in a home automation system:

> Human: (mutters) "Where the heck is that umbrella?"
>
> Computer: "I'm sorry, I didn't understand that."
>
> Human: "I wasn't talking to you. (muttering) I just can't find my stupid umbrella. I wonder if the kids took it?"
>
> Computer: "I'm sorry. I don't understand that command. The options are arm security system, lights off, lights dim, warm up car, ... "
>
> Human: "Stop it! I wasn't talking to you!"

In this scenario, the human enters into a dialogue with the system, whereas another nearby human would most likely recognize that the speech was self-directed and either ignore the muttering, or offer a possible location for the umbrella. Without the ability to recognize the speech as self-directed, the computer begins an inappropriate, and unhelpful, dialogue.

3.1.2 Open-microphone Engagement

For a dialogue to occur, both parties must be engaged in the dialogue. That is, at each point in the dialogue, each party is obligated to ensure that the other has "...the willingness

and ability to continue interaction..." [3]. For SDS, this is essentially an engagement problem, i.e., when should the system be listening to the user?

SDSs typically address the problem of engaging the system by requiring users to explicitly signal their intent to interact with the system. For example, a user would engage a telephone-based SDS by placing a call to the system. To interact with a device-based system, users would be expected to signal their intent to engage the system via "tap-to-talk", "push-while-speaking" or spoken keyword (e.g., "computer"). By requiring explicit user actions, SDSs can assume that users are willing and capable of interacting. However, there currently is interest in developing engagement techniques for speech interfaces that leave users' eyes, hands and attentional focus free for their primary task. This is an especially important consideration for mobile and pervasive interfaces due to safety concerns. In mobile tasks such as cell phone name dialing, interactions are brief and a substantial percentage of users' time (e.g., 30-40%) can be spent simply engaging and disengaging a system.

Current techniques typically assume that a user is speaking to the system if she is facing it with lips moving while speaking [44]. In fact, audio-visual processing of articulated speech that includes the user's head position and corresponding lip movements has shown an improvement in the rate of speech/silence classification compared with audio-only processing [68]. Another technique assumes that a user is speaking to the system if she is looking at the system and her spoken language appears to be a system-directed request [74]. However, to accommodate audiences who have a high rate of self-directed speech, such as children, seniors, or people completing difficult tasks [8, 17, 22], the system will need to be able to reason about whether it should be attempting to engage in conversation.

3.1.3 Self-directed Speech

Speech is both a communicative and cognitive tool. Self-directed speech is viewed as a self-regulatory behavior in which individuals verbalize poorly understood components of a task as they work on it [17]. Self-directed speech supports task performance, is indicative of planful and mature behavior, and has been associated with a reflective rather than impulsive cognitive style [17, 62, 66]. In fact, impulsive children taught to engage in

self-directed speech were able to slow down their response times and reduce performance errors [64]. Additionally, when individuals with Down's Syndrome were trained to use self-directed speech while performing a memorization task, their memory spans increased significantly [25]. Children and adults engage in self-directed speech, including during speech-based system interactions, with the highest rates of self-directed speech occurring during more difficult tasks [17, 29]. During childhood, self-directed speech initially is overt and fully audible, although it is inhibited and becomes progressively quieter as the child approaches adulthood [62, 105].

In the context of considering system processing of an acoustic scene, there is a sense in which a user's self-directed speech can be viewed as "background" speech (self-directed and secondary to the main task) relative to "foreground" speech (outward-directed, task-oriented) that is intended for the system [22]. This raises the question of whether and how speakers may mark their utterances in a reliable manner acoustically as either foreground or background during system interactions.

3.1.4 Elder Speech and SDS

Apart from the issue of self-directed speech, current speech recognition systems have difficulty processing elderly users' speech, with recognition error rates often double that of younger adults [103]. In addition, vulnerability to fatigue and declining working memory both make computer interaction potentially more difficult and error-prone for older adults compared with younger ones [28]. Hearing loss also can have an impact on elders' speech, since it decreases the ability to self-monitor their own speech, producing changes like higher overall amplitude [94].

3.2 Study One: Elderly Participants

In this study, elderly adults' naturally occurring self-directed speech and system-directed speech were compared during a range of realistic tasks varying in difficulty from low to high.

3.2.1 Methods

Participants

Study one included fifteen senior subjects aged 66 to 86 years, six male and nine female. All were native speakers of English and paid volunteers. None of the subjects were computer scientists, and they had varying degrees of computer experience from none to basic E-mail and office processing skills. All subjects were healthy, without any major cognitive deficits, physical limitations, or chronic diseases. All seniors also were living independently, and were physically active within the local community. The educational background of the subjects ranged from high school graduates to Bachelor's degrees. They also were from diverse professional backgrounds, such as nursing, property management, and real estate. All lived in the Portland, Oregon area.

3.2.2 Simulation Technique

The data collection process was based on a high-fidelity semi-automatic wizard-of-oz simulation technique similar to that used for previous studies involving adults [69] and children [106]. In the current simulation environment, the random error generator delivered a 5% task error rate.

Scenario

Subjects were presented with a scenario in which they were to act as non-specialists coordinating emergency resources during a major flood in Portland, Oregon. They were given a multimodal map-based interface on which they received textual instructions from headquarters. They then used this interface to deliver instructions to the map system using both speech and pen input. Individual tasks involved obtaining information (e.g., "Find out how many sandbags are at Couch School Warehouse"), placing items on the map (e.g., "Place a barge in the river southwest of OMSI"), creating routes (e.g., "Make a jeep route to evacuate tourists from Ross Island Bridge"), closing roads (e.g., "Close Highway 84") and controlling the map display (e.g., "Move north on the map").

Figure 3.1 shows a screen shot of the interface used in the experiment. In this example,

the message from headquarters was "Show the railroad along the east water front between Broadway Bridge and Fremont Bridge." Each task was designed for multimodal input. For example, a subject working with the task in Figure 3.1 might say "This is the railroad" and draw a line along the river on the map (see Figure 3.1, Area b).

Figure 3.1: Flood management interface.

The tasks included three levels of difficulty: low, moderate, and high. Low difficulty tasks required the subject to articulate just one piece of spatial-directional information (e.g., north, west), or one location (e.g., Cathedral School). Each additional direction or location translated into one level of difficulty higher. Therefore, moderate difficulty tasks contained two pieces of spatial-directional/location information, and high difficulty tasks contained three pieces. Table 3.1 shows sample tasks from each of these task difficulty levels.

Procedure

Instructional prompts that described tasks were delivered as text instructions on the lower part of the computer screen (see Figure 3.1, area a), which was displayed below a map showing the related area of Portland (area b). There was also a text area for system

Table 3.1: Examples of task difficulty levels, with spatial-directional/location lexical content in italics.

Task Difficulty	**Instruction from Headquarters**
Low	Situate a volunteer area near *Marquam Bridge*
Moderate	Send a barge from *Morrison Bridge barge area to Burnside Bridge dock*
High	Draw a sandbag wall along *east riverfront from OMSI to Morrison Bridge*

feedback (area c), where confirmation or error messages were displayed. The subjects were told to tap the computer screen to engage the microphone before communicating a task, to express themselves naturally using their own words, and to use both pen and speech to communicate each task to the map system. Subjects were told that they could integrate speech and pen input in any way they wished when delivering their multimodal commands to the system, as long as they used both modalities for each task.

The subjects were first given training until they were fully oriented and ready to work alone. Typically, the training took about 15 minutes. However, four subjects required two training sessions, which lengthened their training to 20-35 minutes. Senior adults frequently require longer training times and more help with computer tasks than younger adults [4]. During the training session, an experimenter was present to give instructions, answer questions, and offer feedback and help. Following training, the experimenter left the room and the subjects completed their session independently, which involved 80 tasks.

Upon completion, the subjects were interviewed about their interaction with the system, any errors they experienced, and were debriefed on the purpose of the study. Until that point, all subjects believed they were interacting with a fully-functional computer system. The entire experiment lasted about an hour per participant, although one subject required 1 hour and 40 minutes.

Research Design

The experimental design involved a within-subject comparison of users' interaction as a function of: (1) addressee (self- or system-directed speech), and (2) task difficulty (low, moderate, high).

Data Capture, Coding, and Analysis

For any task in which a participant engaged in self-directed speech, both self-directed speech and system-directed speech were identified and transcribed, which yielded an adjacent pair of self- versus system-directed utterances for that user and task. For amplitude measurements, these adjacent pairs were digitized and then analyzed using Praat [19] speech signal analysis software. All sessions were videotaped, and measures involving identification of self-directed speech and speech comprehensibility were coded using SVHS video-editing equipment.

The following is a description of the scoring conducted for each of the dependent measures.

- Self-directed Speech – audible speech verbalized by the user prior to and independent of addressing the system during a task. The presence or absence of self-directed speech for each task was coded, which then was converted to a percentage of tasks containing self-directed speech.

- Utterance Comprehensibility– utterances were classified by human coders as containing no comprehensible words, or else as containing partially or completely comprehensible lexical content. This data then was converted to a percentage of all self- versus system-directed speech that contained comprehensible lexical content.

- Amplitude – adjacent pairs that were scored as containing any comprehensible articulated speech and that did not involve reading aloud were analyzed for amplitude. Self-directed speech and system-directed speech regions were hand-labeled using Praat, and intra-sentential pauses over 0.33 seconds were excluded from analysis. Amplitude measurements also were normalized relative to consistent ambient noise in the recording room to correct for recording level variations between the studies and participants. To accomplish this, ambient room noise samples were labeled in each recording when there was no speech or extraneous noise.

  Amplitude in decibels (dB) is typically measured relative to a constant approximating the average auditory threshold for human hearing: dB = 10*log10(P/Pref),

where P is the power of the speech signal and Pref is the power of the referent auditory threshold. For our purposes, P is the power of the speech of interest (self- or system-directed) and Pref is the ambient noise within that recording.

Two separate dependent measures of amplitude were summarized: (1) amplitude across all spoken regions, and (2) amplitude across voiced regions. For amplitude across spoken regions, Praat was used to compute the power measurements of the self-directed speech, system-directed speech, and ambient-noise labeled regions within each adjacent utterance pair. Amplitude of speech in dBr (i.e., relative to ambient noise) then was computed for self- and system-directed speech. To check for convergence of results, amplitude across voiced regions also was measured using Praat to identify only the voiced regions within the self- and system-directed speech. This measure was more conservative, since it excised whispered speech which occurred more frequently during self-directed speech.

Reliability

Second scoring for presence of self-directed speech was completed for 20% of the data, with an exact match on 93%. Second scoring for comprehensibility was completed for 36% of adjacent utterances, with an exact match on 97%. Identifying the duration of speech regions was second scored for 29% of the data, with 80% having less than a 0.31 second departure. This degree of departure resulted in amplitude measurements reliable to within 0.18 dBr.

3.2.3 Results

In total, data were available for analysis on 404 tasks.

Presence and Comprehensibility of Self-directed Speech

Twelve of the fifteen participants (80%) engaged in self-directed speech at some point during their session, producing a total of 147 adjacent utterance pairs containing self-directed speech. Overall, 36.4% of the tasks contained user self-directed speech before addressing the system. However, as shown in Figure 3.2, there were large individual

differences in participants' rate of self-directed speech, ranging from 0 to 100% of their tasks.

Of the 147 adjacent utterance pairs that contained self-directed speech, 92 pairs produced by 11 participants were available for analysis of utterance comprehensibility after excision of clipped recordings and utterances that involved reading aloud. Self-directed speech utterances were partially or fully comprehensible 82% of the time, compared with 100% for system-directed speech, a significant difference by Wilcoxon signed rank test, $T+ = 15$, $N = 5$, $p < 0.031$, one-tailed.

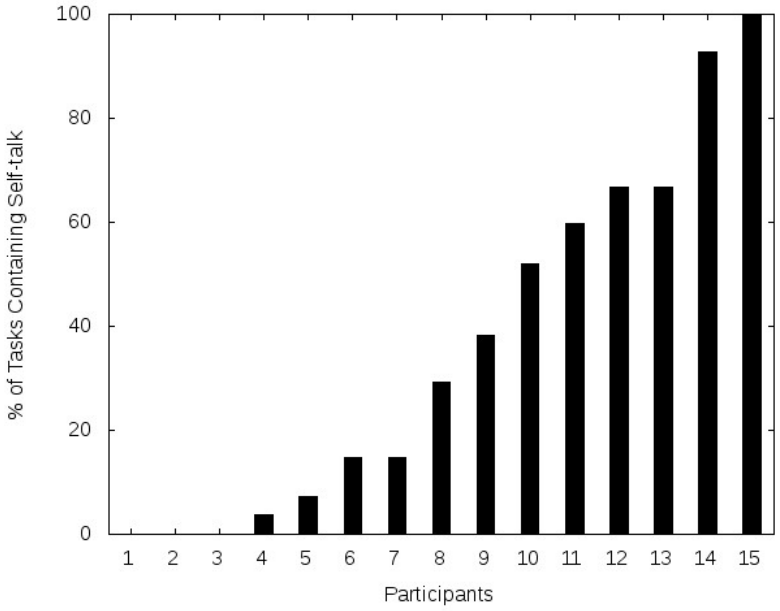


Figure 3.2: Individual differences in rate of self-directed speech for elder adults.

Amplitude

For amplitude analyses, only those pairs that included fully or partially comprehensible self-directed speech were included, resulting in 81 utterance pairs, produced by 10 participants, being available for comparison. Participants' average amplitude during self-directed speech was 13.95, 17.82, and 15.07 dBr for low, moderate and high difficulty tasks respectively, with no significant differences as a function of task difficulty by paired t-test, all t values < 1.54, N.S. Likewise, users' average system-directed amplitude was 30.68, 35.86,

and 37.66 dBr, again with no significant differences by paired t-test, all t values < 2.25, N.S. As a result, all further amplitude analyses were collapsed across task difficulty level.

As shown in Figure 3.3, participants' self-directed speech was lower in amplitude than their system-directed speech for 100% (81 of 81) of the utterance pairs. The average amplitude of self-directed speech was 15.15 dBr, significantly lower than the 33.96 dBr for system-directed speech, a priori paired t-test, $t = 17.58$ ($df = 80$), $p < 0.0001$, one-tailed. The average amplitude of voiced self-directed speech was 20.00 dBr, which also was significantly lower than 39.24 dBr for voiced system-directed speech, a priori paired t-test, $t = 15.21$ ($df = 76$), $p < 0.0001$, one-tailed.[4]

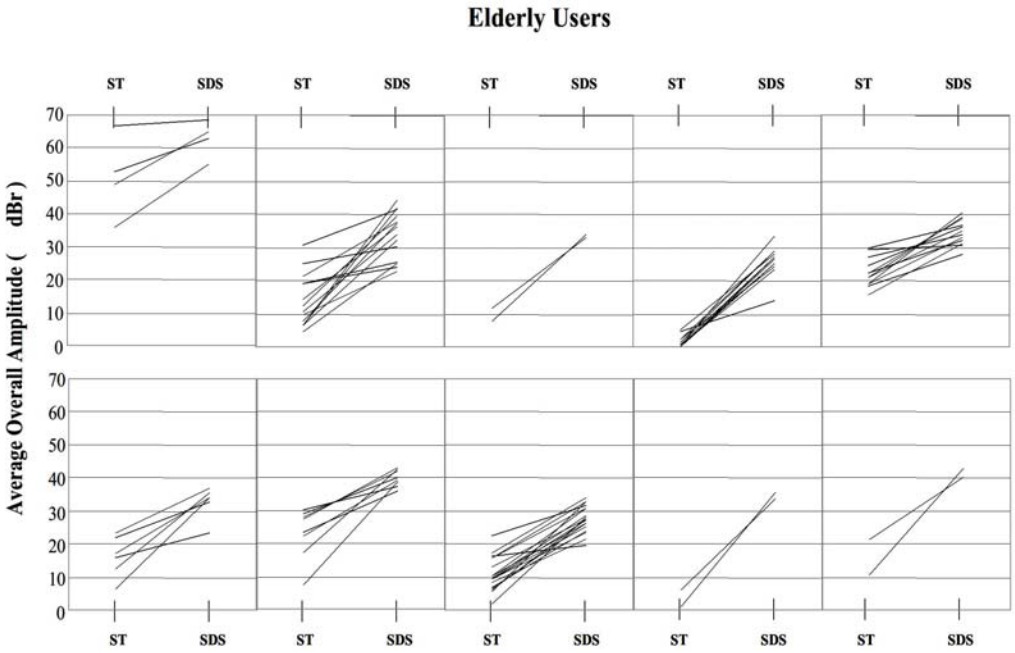


Figure 3.3: Individual differences in average amplitude of self-directed speech (ST) versus system directed speech (SDS) in adjacent utterance pairs for the ten elderly adults who produced comprehensible self-directed speech.

[4]Statistical comparisons using subject grand means were also significant for all amplitude measures.

3.3 Study Two: Younger and Elderly Adults

In this study, both younger and older adults' self-directed speech versus system-directed speech was compared, as was their gaze deployment at the system.

3.3.1 Methods

Study two included sixteen paid participants, ten adults who were 18-61 years of age, and six seniors 66-89 years of age. The younger adults included six females and four males, and the seniors included four females and two males. The tasks and procedure for study two were the same as study one, except an open-microphone simulation was employed, and the data collection used a wizard-of-oz simulation technique with a random error rate of 20% as described by Oviatt et al. [70]. As the users did not "tap" to engage, the wizard was instructed to interpret the user's command as they would human-human communication. The experimental design involved a within-subject comparison of users' interaction as a function of (1) addressee (self or system), and a between subject comparison of (2) age group (younger or elder adults).

Dependent Measures and Reliability

As in study one, both the presence and comprehensibility of self-directed speech were analyzed. In addition, amplitude measures for self-directed speech and system-directed speech in adjacent utterance pairs (i.e., for the same user and task) were analyzed. Finally, gaze during the adjacent utterance pairs was assessed to determine whether participants looked at the system or not when beginning their utterances (i.e., within the first 0.5 second).

Second scoring for presence of self-directed speech was completed for 13% of tasks and matched exactly for 92% of them. Comprehensibility of self-directed speech was second scored for 24% of utterance pairs containing self-directed speech, and matched exactly for 80%. Identifying the duration of speech regions was second scored for 26% of the data, with 80% having less than a 0.30 second departure. This degree of departure resulted in amplitude measurements reliable to within 0.18 dBr. Gaze was jointly scored by two

analysts for 75% of data, with 100% agreement.

3.3.2 Results

In total, data were analyzed on 766 tasks.

Presence and Comprehensibility of Self-directed Speech

Overall, 216 tasks (28.2%) contained self-directed speech. Twelve of the sixteen participants, seven younger and five older adults, engaged in self-directed speech at some point in the session. The average rates of self-directed speech for younger and older adults were not significantly different (25.9% and 32.1%, respectively) by Wilcoxon rank sum test, $z < 1$, N.S. As found in study one and shown in Figure 3.4, there were large individual differences in participants' rate of self-directed speech for both younger adults and elder adults.

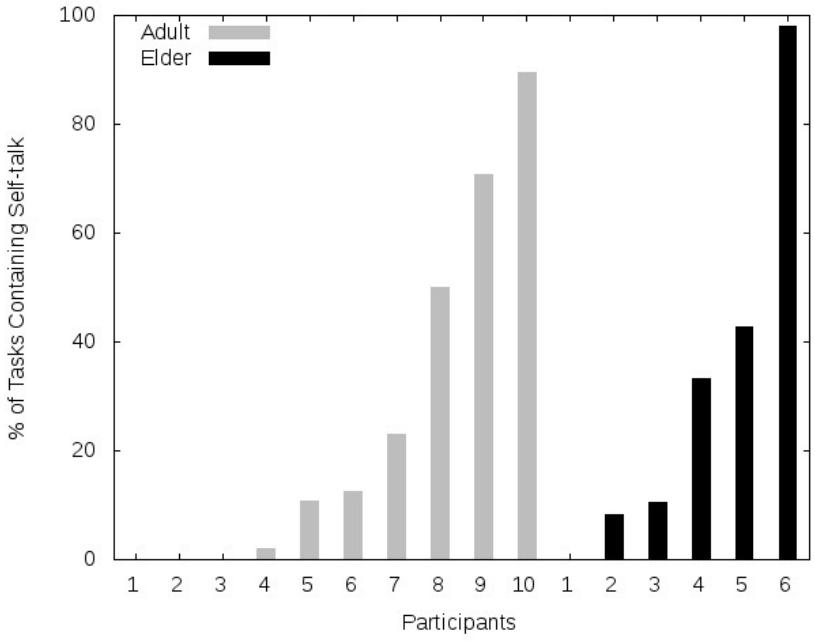


Figure 3.4: Individual differences in rate of self-directed speech for younger and elder adults.

Of the 216 adjacent utterance pairs containing self-directed speech, 152 were available for analysis of utterance comprehensibility after excision of clipped recordings and

cases involving reading aloud. Participants' average rate of producing self-directed speech that contained comprehensible lexical content was 63%, compared with 100% for system-directed speech. Seniors' rate of producing comprehensible self-directed speech averaged 81%, which was significantly higher than younger adults' rate of 50%, by Wilcoxon rank sum test, $z = 2.81$, $p < 0.003$, one-tailed.

Direction of Gaze at Start of Speech

The percentage of utterances for which participants' gaze was directed at the system at the start of speech was 99.5% for self-directed speech and 98.1% for system-directed speech.

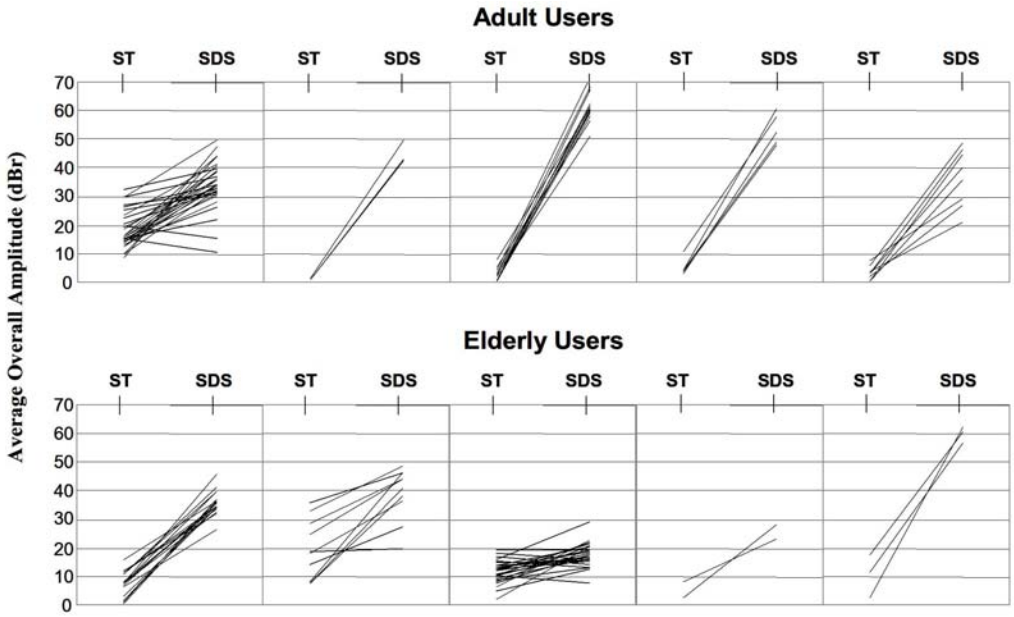


Figure 3.5: Individual differences for younger (top) and elder (bottom) adults in average amplitude of self-directed speech (ST) versus system-directed speech (SDS) in adjacent utterance pairs

Amplitude

For amplitude analyses, 119 utterance pairs were available for comparison, including 60 pairs from five of the younger adults and 59 from five of the elders. As shown in Figure 3.5, self-directed speech was lower in amplitude than system-directed speech for 94% of the adjacent utterance pairs. The amplitude of self-directed speech was 11.16 dBr, significantly

lower than that of system-directed speech at 36.14 dBr, a priori paired t-test, t = 14.58 (df = 118), p < 0.0001, one-tailed. The average separation between self-directed speech and system-directed speech for all users was 24.98 dBr.

As shown in Figure 3.6, the self-directed speech of seniors averaged 11.48 dBr, as compared to 29.35 dBr for their system-directed speech, a significant difference by a priori paired t-test, t = 9.82 (df = 58), p < 0.0001, one-tailed. Younger adults' self-directed speech averaged 10.84 dBr and their system-directed speech 42.82 dBr, also a significant difference by a priori paired t-test, t = 12.30 (df = 59), p < 0.0001, one-tailed.

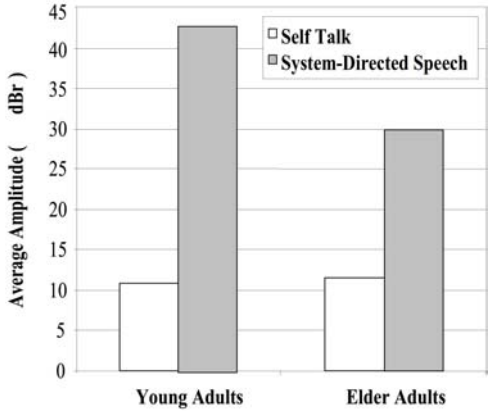


Figure 3.6: Average amplitude separation for self- versus system-directed speech in younger and elder adults.

Analysis of amplitude across voiced regions replicated these results. The average amplitude of self-directed speech was 14.96 dBr, which was significantly lower than that of system-directed speech at 39.99 dBr, a priori paired t-test, t = 15.55 (df = 113), p < 0.0001, one-tailed. The self-directed speech of seniors averaged 14.31 dBr, compared to 34.15 dBr for their system-directed speech, a significant difference by a priori paired t-test, t = 10.50 (df = 58), p < 0.0001, one-tailed. Young adults' self-directed speech averaged 15.66 dBr and their system-directed speech 46.25 dBr, also a significant difference by a priori paired t-test, t = 12.48 (df = 54), p < 0.0001, one-tailed.[4]

3.4 Combined Analyses on Amplitude Separation

Following up on the differences in amplitude separation between age groups shown in Figure 3.6, data were combined from study one and two. Analyses revealed a strong correlation between a user's age and average amplitude separation when they engaged in self- versus system-directed speech, by Pearson correlation r = 0.89, which was significant, F = 56.47 (df = 1,15) p < 0.0001, two-tailed. In fact, 79% of the variance among subjects in magnitude of amplitude separation could be predicted simply by knowing an individual's age, p2xy = 0.79, (N = 17). Figure 3.7 shows the best-fitting linear regression.

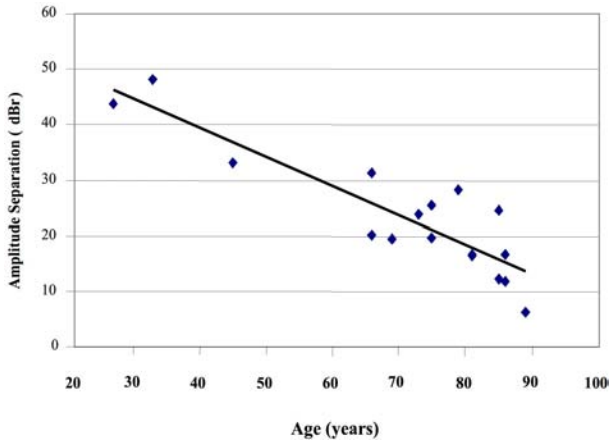


Figure 3.7: Linear regression showing amplitude separation between self- and system-directed speech as a function of age.

Follow-up comparisons evaluated whether elderly adults' diminished amplitude separation was due to higher amplitude self-directed speech, lower amplitude system-directed speech, or both. Across both studies, seniors' average amplitude during self-directed speech was 12.33 dBr, compared with 3.18 dBr for younger adults, a significant difference by independent t-test, t = 4.61 (df = 16), p < 0.0001, two-tailed. For system-directed speech, seniors' amplitude averaged 32.54 dBr, compared with 46.39 dBr for younger adults, also a marginal difference (i.e., 0.5<p<1.0) by independent t-test, t=2.74 (df = 5), p < 0.058, two-tailed.

To evaluate whether elderly adults' diminished amplitude separation could be attributable to fatigue over the session, elderly users' amplitude separation was compared

for the first versus second half of their session. A median-split on data from eight elderly adults during these two phases of a session revealed that amplitude separation between self-directed speech and system-directed speech averaged 18.38 dBr for the first half versus 17.38 dBr for the second, which was not a significant difference by paired t-test, $t < 1$, N.S.

To summarize, amplitude separation for self- versus system-directed speech decreased with age, with elderly adults suppressing the amplitude of their self-directed speech less than younger ones, and also marginally less likely to elevate amplitude when addressing the system. Elderly adults' diminished amplitude separation could not be attributable to increased fatigue during the second half of their one-hour session.

3.5 Discussion

The present findings indicate that people use amplitude to indicate their self- versus system-directed during human-computer interaction in a highly reliable way. Across both studies, all younger and older adults lowered their amplitude during self-directed speech and raised it during system-directed speech. This pattern also replicated for two different dependent measures of amplitude. In fact, only two participants out of twenty ever produced an instance of self-directed speech with higher amplitude than adjacent system-directed speech, accounting for only 3.5% of all data. On average, participants' magnitude of amplitude separation between self-directed speech and system-directed speech was over 26 dBr. As shown in Figures 3.3 and 3.5, participants also exhibited large individual differences in the magnitude of their amplitude separation, ranging from approximately 10 dBr to 60 dBr.

Furthermore, differences in amplitude separation between self-directed speech and system-directed speech were highly correlated with a participant's age. In fact, 79% of the variance in magnitude of amplitude separation was predictable simply by knowing a user's age. In particular, clear separation of amplitude to mark intended addressee diminished with increasing age, with the smaller amplitude separation in older adults mainly due to less suppression of their self-directed speech amplitude. This finding corresponds with

elder adults' lower rate of incomprehensible speech during self-directed speech.

Given previous research showing increased rates of self-directed speech with task difficulty, and seniors' diminished working memory capacity, it was anticipated that seniors would engage in more self-directed speech than younger adults. Interestingly, seniors' rate of self-directed speech was not higher than that of younger adults. However, they did produce louder self-directed speech. It is possible that seniors' increased self-directed speech amplitude could be accounted for by hearing loss and a related reduction of auditory feedback [94], which may have reduced their ability to self-regulate amplitude level. This loss of precision in seniors' ability to control their amplitude can be expected to degrade the performance of speech recognition systems with which they interact, including presenting greater challenges for reliable microphone engagement.

The use of amplitude to differentiate self-directed speech from system-directed speech is consistent with Buxton's [22] view that people separate their activities into "foreground" and "background" in a manner that corresponds with an ongoing task. In essence, people "mark" their utterances as either foreground (outward-directed, task-oriented) or background (self-directed, secondary to main task) by the forcefulness with which they assert an utterance acoustically. As a communicative tool, variations in amplitude can be used both to mark an intended addressee and to direct listeners' attention to task-oriented content.

3.5.1 Lessons for HCI

Self-directed speech was present during over 30% of users' tasks, with over 70% of participants engaging in self-directed speech at some time during their session. Furthermore, the incidence of self-directed speech was not significantly lower in younger adults than the elderly. This high rate of self-directed speech, which has not previously been acknowledged as an important source of "noise", will present a challenge for future engagement systems.

It is important to note that, although self-directed speech is not the users' primary task, it is an important support activity in organizing their behavior to complete tasks successfully. In developing user interfaces for realistic field tasks, it will be critical not only to support users' foreground tasks, but also to design in a manner that does not interfere

with background behaviors that involve self-regulation and performance enhancement.

In designing future interfaces that can successfully distinguish a user's intended addressee, developers will be able to take advantage of two predictable aspects of peoples' amplitude separation: 1) users reliably adapt their speech amplitude to differentiate self- from system-directed speech, and 2) their magnitude of amplitude separation, although very large, diminishes predictably with advancing age. In addition, the large individual differences revealed in the present data in magnitude of amplitude separation during self- versus system-directed speech indicate that user-adaptive system processing would be advantageous.

Chapter 4

Human- versus System-directed Speech

In this chapter, we investigate human- versus system-directed speech. From a dialogue coordination perspective, an SDS should strive to avoid responding when it has not been given the turn, and to respond when it has. To do so, an SDS will need to be able to differentiate speech that is addressed to it, and to which it must respond, and speech that is directed to other humans. The findings in Chapter 3 and that of other researchers (discussed in related work), has shown that people use increased amplitude to capture and direct an addressee's attention. Thus, in this chapter, we focus on amplitude as a potential cue of intended addressee during computer-assisted group interactions. Specifically we address:

- Whether the amplitude of a speaker's human- versus system-directed speech differs systematically and, if so, whether the magnitude of amplitude shifts are large ones. We hypothesize that speakers would use substantially higher amplitude when addressing a computer compared with human peers (i.e., comparing utterances matched on illocutionary force), essentially treating the computer as an inattentive or "at-risk" listener [72].

- Whether speakers dynamically and bi-directionally adapt their amplitude when switching between human versus computer addressee, as would be observable between adjacent utterances. We hypothesize that substantial, abrupt, and bi-directional amplitude shifts would occur across adjacent utterance boundaries representing such changes in addressee.

- Whether speakers use amplitude as an alternative strategy to explicit lexical markers (e.g., "computer") to identify their intended addressee. We hypothesize that amplitude differences would be largest when lexical markers were absent, and attenuated when present.

- Whether the dialogue style of a speaker's human- versus system-directed speech (i.e., matched on illocutionary force) differs during computer-assisted group interactions, reflecting a higher ratio of command-style utterances to a computer partner. We hypothesize large individual differences among speakers in the adoption of command-style speech, and also a higher ratio of such input directed to computers.

The work in this chapter is based on an earlier work [60]: Toward open-microphone engagement for multiparty interactions, in *Proceedings of the 8th international conference on Multimodal interfaces* (ICMI '06). http://doi.acm.org/10.1145/1180995.1181049.

4.1 Background and Related Work

4.1.1 Motivation

There is currently much interest in designing SDSs that can aid people engaged in multi-party and collaborative interactions in mobile or educational settings. To use a typical computer system during multi-party human interaction, a participant must disengage from the human interaction and attend to the process of using the computer. One potential advantage of a speech-based interface is the opportunity to provide computational assistance (e.g., information retrieval) without the need to stop the current task and manually engage and interact with the computer system.

However, to ensure that an SDS does not distract users from the task at hand, it will need to meet certain obligations, similar to those that would be expected of a human assistant. First, an SDS must recognize when it is being addressed and, per Sacks, when it "...is obliged to take the turn..." [82]. By doing so an SDS could avoid derailing the users' task, as a user would not need to interrupt their work to engage the system. Second, an

SDS must avoid responding to speech addressed to a human, since "The party so selected has the right and is obliged to take the next turn to speak; no others have such rights or obligations,..." [82]. By doing so, the SDS could avoid interrupting users' collaboration or train of thought.

To meet these obligations discussed above, an SDS needs to recognize the dialogue coordination mechanisms that differentiate speech addressed to the system from that addressed to humans. Towards this end, researchers have explored what mechanisms people use to differentiate addressee during multi-party human-human communication. For example, gaze has been found to be a particularly salient cue of addressee during multi-party human interaction [80]. However, given that current SDSs do not have the capacity to interact as humans do (e.g., lack of joint gaze), we anticipate that the cues speakers use to identify a computer addressee may differ from those used to identify a human addressee.

4.1.2 Multi-party Human-Human-Computer Interaction

Towards understanding what cues might best differentiate human- versus system-directed speech, recent research explored whether gaze direction would prove a reliable indicator. Research on human-human-computer and human-human-robot interaction in multi-person field settings indicates that gaze is not a reliable cue that a user is addressing the system [8, 99, 47]. In these studies, users looked at the system while addressing their human interlocutor 35-57% of the time. Researchers looking at human-human-computer interaction posited that this unexpected gaze behavior was due to the kiosk functioning as a "situational attractor" that encouraged gaze because the computer provides visual information. Thus, it is clear that gaze alone is not enough to differentiate human- versus system-directed speech.

Compared with gaze cues alone, Katzenmaier et al. [47] found that the accuracy level of automatically detecting a human versus computer interlocutor was 89% using a multilayer perceptron when gaze cues were combined with sentence length, number of imperatives, parseability, and the presence of a lexical marker for the robot. Other work also has incorporated dialogue and linguistic sources of information to attempt automatic

detection of intended interlocutor. In work by Turnhout et al., a Naive Bayes classifier that included gaze, utterance length, and dialogue events (i.e., button presses and system prompts) achieved a precision of 80%, but a recall of only 33% [99]. One caution in interpreting these results is that the utterances addressed to the computer versus human peer were not comparable sets with respect to illocutionary force and content. Furthermore, researchers typically assumed that speech to a computer would contain more command style language, although data on this issue were not systematically explored and reported.

Recent work by Reich et al. [78] examined whether using prosody in concert with automatic speech recognition decoding features could differentiate commands addressed to a SmartBoard from conversational speech addressed to a nearby human. An F-score of 0.830 was achieved on automatically segmented data using an unadapted acoustic model. However, the data was collected during "dry-runs" in which the system did not react to any of the speech, and it was assumed that speech addressed to the system would be fully parseable and contain no out-of-vocabulary terms. Thus, the question remains as to whether the speech is representative of actual real-time human-computer interaction.

4.1.3 Functions of Speech Amplitude

People actively use amplitude to attract and maintain the attention of a listener, to mark new lexical information [65] and discourse structure [41, 42, 33], to foster social cohesion [27, 102], and to identify speech addressed to a computer as shown in Chapter 3. Speakers' tendency to increase amplitude to capture the attention of others is supported by evidence in the cognitive neuroscience literature, which shows that changes in the intensity and frequency of speech trigger involuntary attentional shifts in the brain of listeners [31, 86]. Mothers addressing infants likewise increase their amplitude when teaching new lexical items [65] which serves to attract and maintain infants' attention during early socialization and language learning. From the viewpoint of social interaction management, the literature on Communication Accommodation Theory also has shown that speakers will adapt their amplitude to converge with that of an interlocutor during interpersonal interactions [102], which is believed to signal social status and to foster social cohesion.

From a linguistics and discourse processing viewpoint, speakers also are known to

increase their amplitude at the beginning of a discourse segment and decrease it at the end during both elicited monologues and read-aloud speech [42]. Speakers also decrease their amplitude to mark information or discourse segments that are more extraneous or less critical to a listener's understanding. For example, in monologues read aloud by professional newscasters, parenthetical statements were lower in amplitude than preceding speech [41], which effectively separated them acoustically.

In human-computer contexts, people interacting with a spoken language system typically decrease their amplitude 1 to 2 dBs prior to a topic change, and increase it when starting a new topic [33](G. Levow, personal communication, 2006). In studies of young children interacting with animated computer characters, it was shown that they will alter their spoken amplitude to converge with that of text-to-speech (TTS) output from computer characters, dynamically increasing or decreasing their amplitude by up to 1.1 dB in such contexts [27]. In Chapter 3, speakers were observed to reduce their amplitude a substantial 26dB during self-directed speech, compared with speech addressed to a computer. This latter finding suggests that large amplitude shifts may play a particularly important role in marking intended addressees, or when attracting listeners' attention during communication.

Speaking at an appropriate amplitude is critical to maintaining an effective communication channel, as inappropriate amplitudes affect listeners' understanding and performance. However, the appropriate amplitude depends on the listener and their environment. Baldwin found that audible, but lowered, amplitude can negatively affect both younger and older subjects' reaction time and ability to respond correctly while multitasking [9]. Elderly listeners are likely to need higher amplitudes than younger cohorts to maintain similar performance [9]. Hard of hearing individuals often have a diminished amplitude range, resulting in difficulty understanding both speech that is too loud and speech that is too soft. Just as low amplitude can present communication issues, high amplitude can be annoying or cause pain.

4.2 Methods

4.2.1 Participants

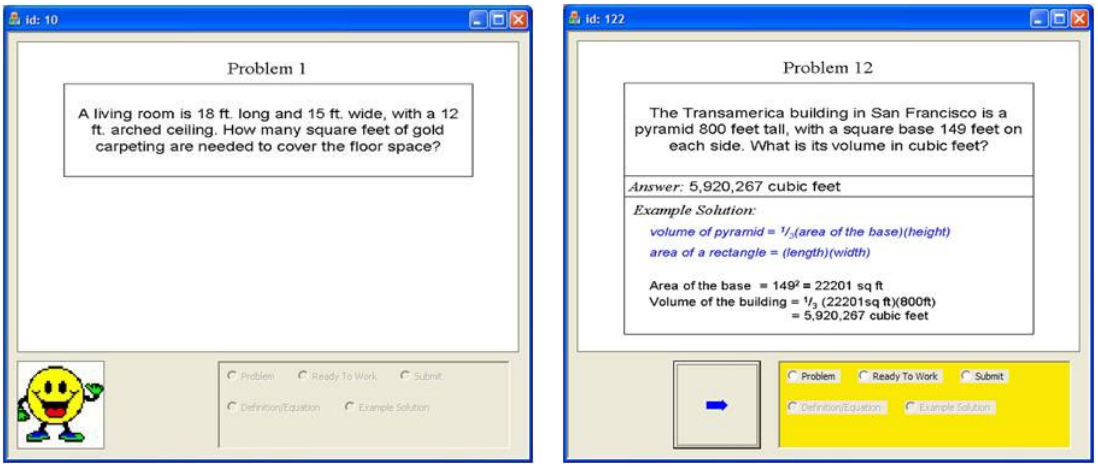
Participants in this study included 9 female and 9 male high school students who ranged in age from 15 to 17 years old. All had recently completed Geometry 1 at a local high school and represented a range of geometry skills from average to high performers. Performance rankings were based on teacher feedback and student self-reports. Participants were paid volunteers and all were native English speakers.

4.2.2 Tasks

Each session consisted of 16 basic algebra and geometry problems presented as word problems with each set of 4 including a low, moderate, high and very-high difficulty problem. Problem difficulty was controlled by creating math problems varied along dimensions known to make them more challenging. These dimensions included the number of math terms, the number of equations needed to solve the problem, whether or not the equation could be applied directly, whether units required translating (i.e., inches to feet), and the number of steps required. Figure 4.1(a) shows an example of a low-difficulty problem which included one math term (e.g., square feet) that could be solved by directly applying the equation for area of a rectangle. Figure 4.1(b) shows an example of a very-high-difficulty problem, which contained four math terms (e.g., pyramid, square, volume, cubic feet) and required equations for the area of a rectangle and the volume of a pyramid. The difficulty levels of the problems were validated using: 1) teacher records of percentage correct on similar problems for high school students in introductory geometry, 2) pre-experiment piloting, and 3) students' percentage of correct solutions in the current study.

4.2.3 Procedure

The six groups of students participating in the study worked collaboratively in groups of three to solve basic geometry and algebra problems, and each triad participated in two separate sessions. Groups were seated at a table with one randomly selected student designated as group "leader" next to the computer monitor and his peers in the remaining



(a) Interface displaying an easy math problem.

(b) Interface displaying a very hard math problem, answer and example solution.

Figure 4.1: Screen shots illustrating the two different interface modes, problem difficulty levels, and computer features.

two chairs, as shown in Figure 4.2(a). In addition, a TI-83 calculator, digital graph paper, Nokia digital pens, and a mouse were provided for students' use, as shown in Figure 4.2(b). Since a different leader was selected from each group for their second session, a total of 12 separate sessions involving 12 different leaders was conducted.

Each group was instructed to exchange information and expertise as they worked on solving the problems, so everyone understood the solution and could explain it if asked. To ensure that all students participated fully, they were told that during the session each participant would be randomly asked to explain one or more of the group's math solutions. The leader and was told that he or she was responsible for both coordinating the group's activity and getting information from the computer.

Participants were told that the leader would interact with the computer in one of two ways, either directly or mediated through a peer, depending on the computer's current mode. When the computer was in direct-communication mode, the leader could instruct it to show the next problem, clear any screen distractions while they worked, submit their answers, show terms and equations, and show problem solutions. They could communicate with the computer as they would with another participant, using speech, gestures, and other input. The system was in direct mode when the yellow smiley persona was present

(a) Leader view with computer monitor to the right and two student peers to the left.

(b) Room view showing papers, digital pens, calculator and mouse.

Figure 4.2: Room setup for study on computer-assisted peer tutoring.

in the lower left corner of the screen, as shown in Figure 4.1(a). When the computer was not available for direct communication, buttons showing the available computer functions were activated on the bright yellow panel shown in Figure 4.1(b) on the lower right side. In this mode, the leader instead addressed any communications to a peer to complete the same functions. Their peer then would use a mouse to select the appropriate action on a labeled radio button. The leader's peers were cautioned by the experimenter not to initiate an action unless the leader instructed them to do so. By specifying that the leader initiate and guide all interaction with the computer we were able to orchestrate the collection of matched human- and system-directed instructions spoken by the same participant (i.e., the leader).

Following orientation and instructions, each group was given three sample problems to familiarize themselves with the computer's features and the different modes of interaction. After completing the practice problems, the experimenter left the room and the group began the main session which lasted about an hour. During the session, the distance between the leader and the peer or computer addressee was controlled, with two distances, 2 and 4 feet. To accomplish this, the 16 math problems were presented in blocked sets of four, after which the experimenter re-entered the room and moved the monitor or had the participants change seats as needed (see Table 4.1), so that all data collected on peer versus system-directed communication would be completely counterbalanced and matched

on addressees' distance from the leader.

Table 4.1: Counterbalancing of addressee and distance from the leader during each session.

Problem	**Target Addressee**	**Distance to Leader**
1-2	Computer	Distal
3-4	Peer 1	Distal
Pause	*Peers exchange seats*	
5-6	Peer 2	Distal
7-8	Computer	Distal
Pause	*Computer moved to proximal location*	
9-10	Computer	Proximal
11-12	Peer 1	Proximal
Pause	*Peers exchange seats*	
13-14	Peer 2	Proximal
15-16	Computer	Proximal

At the end of each of the two sessions, every participant completed a written questionnaire. At the end of the first one, people were asked questions about the interface quality and ease of use (e.g., computer voice, visual display, and digital pens). After the second session, participants were asked questions about interactions with the computer, including how the leader indicated whether he or she was addressing the computer or a peer.

4.2.4 Computer-assisted Instruction and TTS

The computer provided basic functions that included: 1) displaying the math problem (Fig. 4.1(a)), 2) turning off the display to avoid distraction during problem solving, 3) accepting the groups' submitted answer, and 4) displaying the correct answer (Fig. 4.1(b)). Upon request, the computer also provided: 5) definitions of math terms or equations via pre-recorded TTS, and 6) example solutions (Fig. 4.1(b)). As a reminder, a summary of available computer functions was taped to the bottom of the computer monitor. In concert with display changes during the above requests, the computer also responded with TTS phrases such as "problem one" when displaying a new problem, "okay, take your time" when turning off the display so students could focus on work, and "here's one solution" when displaying an example solution.

The computers' TTS voice was gender matched with the group, and its amplitude was

matched to that of the leader's human peers. In matching amplitude, care was taken to ensure that the TTS amplitude was the same or slightly lower than the human peers, since past research has shown that two speakers will converge on one another's amplitude [71]. This provided the required context for investigating the hypothesized amplitude effects without any contamination from amplitude changes due to social convergence per se.

4.2.5 Simulated Computer Interface

The computer's interface was implemented using a Wizard-of-Oz simulation, which was presented to participants as a fully functional system. Simulation infrastructure was developed for this study that permitted the wizard to view multiple video feeds of the group's interaction as data was collected, while simultaneously listening to the leaders' speech in one ear and a room microphone of the group's discussion in the other. Using audio and video cues, the wizard could differentiate verbal instructions spoken by the leader from those spoken by other group members. During the session, the wizard controlled the system's responses to participants' requests and instructions as detailed in [6]. The simulation was designed to provide speedy and accurate wizard responding, and it generated 5% random errors for realism and credibility.

4.2.6 Research Design

The experimental design involved a within-subject comparisons as a function of: (1) Addressee (human versus computer), and (2) Distance between leader and addressee (proximal (2 feet) versus distal (4 feet)).

Each student group was matched to have the same gender and geometry skill level (low, moderate/high, and high) to facilitate more reciprocal and collaborative interactions. See Table 4.1 for the counterbalancing protocol.

4.2.7 Data Capture, Coding and Analysis

For each session, a high-resolution digital video close-up of each participant was captured using Point Grey digital firewire cameras, as well as a wide angle room view and a view of the tabletop with paper, pens, and other artifacts. Digital audio recordings also were

collected of each participant's speech using Countryman hyper-cardioid microphones connected to Shure wireless transmitter/receivers. In addition, a digital audio recording of the room was collected using a studio-quality omni-directional microphone hung above the leader's head. Finally, each participant's writing was collected using Nokia digital pens and Anoto paper. The audio-visual media streams then were synchronized for analysis purposes (see [6]).

Areas of interest in the leaders' speech were identified and transcribed using a modified version of the MockBrow annotation tool [6]. Using this tool, annotators could hear and view the five synchronized audio/video data streams during the group meeting, and could mark selected segments for amplitude measurement using the Praat speech signal analysis tool [19]. Leaders' spoken utterances were pitched-tracked, and mean intensity measurements were taken over all pitch-tracked voiced regions. By using only the voiced regions, we eliminated any potential contamination of average amplitude due to silent regions.

Instructions

Speech regions in which the leader was attempting to activate a computer function (see Section 4.2.4), whether addressed directly to the computer or a human peer, were hand-labeled and transcribed. This resulted in system-directed and human-directed instructions, matched on illocutionary force (i.e., ordering), for each of the 12 leaders. Utterances contaminated with laughter or other extraneous noise were excluded, as were misdirected instructions (e.g., instructions directed to the computer when it was not available for direct communication).

Lexical Marking

All instructions were also coded as containing a lexical marking of intended addressee (e.g., "Computer", "Susan") or not. Amplitude analysis of these instructions compared all instructions, only those with lexical markings, and only those without lexical markings.

Adjacent Utterances

Immediately Adjacent Pairs: Utterances immediately preceding the instructions were also hand-labeled and transcribed, so that adjacent utterance pairs could be analyzed (i.e., ones separated by less than 2 seconds with no intervening speech). This resulted in adjacent pairs of human-human and human-computer utterances for each speaker. For these adjacent utterances, the first utterance varied from queries to replies to clarifications (e.g., "Are we ready to work the next problem?", "yeah, 50 pi"). Utterances that were laughter, expletives, or reactions to system actions were excluded. Adjacent utterance pairs were coded as including a lexical marker of intended addressee if either utterance in the pair contained such a name. Amplitude separation then was compared for all pairs, pairs with lexical markers, and pairs without lexical markers.

Across Phases Shifts: In addition, since speakers addressed their instructions to either the computer or a human for blocks of 2-4 problems, as shown in Table 4.1, amplitude was compared for instructions immediately preceding and following these phase shifts. For this adjacency measure, the amplitude of three instructions immediately preceding and following a change of target addressee were averaged and compared.

Dialogue Style

The instructions were also coded as either a command or not. Instructions coded as commands either incorporated an imperative verb (e.g., "get problem"), or elided the verb, leaving only a noun phrase (e.g., "next problem"). When coding instructions, any disfluencies, lexical discourse markers, politeness terms, and lexical markers of intended addressee were disregarded. For example, the instruction "Okay, uh next problem please Susan." was coded as a command.

Reliability

Second scoring for identification of leaders' spoken instructions was completed for 19% of the data, with an 87% match on the speaker utterances identified. Second scoring for identification of scorable adjacent preceding speech was completed for 12.5% of the

data, with a 92% match on the speaker utterances identified. Second scoring for coding instructions on dialogue style was completed for 19% of the data and matched for 98%.

4.3 Results

In total, ten hours of meeting data were collected, from which data were available for analysis on 658 instructions, and 334 adjacent utterance pairs.

4.3.1 Instructions

With respect to physical distance, speakers' average amplitude for system-directed instructions was 66.6dB when the computer was at the proximal distance and 66.4dB when the computer was distal, not a significant difference by paired t-test, $t < 1$, NS. Likewise, for human-directed instructions, speakers' amplitude was 63.9dB when the addressee was proximal and 64.2dB when distal, again with no significant difference by paired t-test, $t < 1$, NS. As a result, all further amplitude analyses were collapsed across distance.

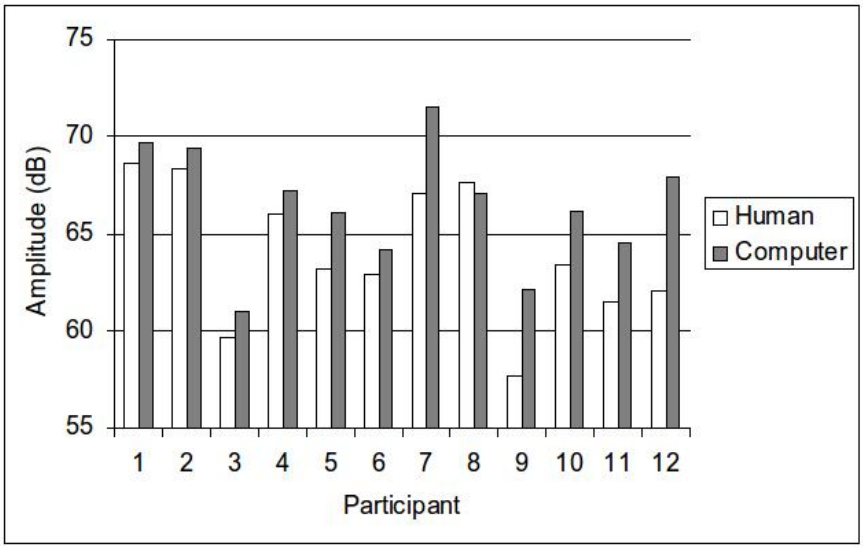


Figure 4.3: Amplitude of human-addressed versus computer-addressed instructions for all 12 leaders, matched on illocutionary force.

As shown in Figure 4.3, eleven of the twelve leaders (92%) used higher average amplitude when addressing instructions to the computer compared with a human peer. As predicted, leaders' amplitude for all human-directed instructions averaged 64.0dB, significantly lower than the 66.4dB amplitude for system-directed instructions, a priori paired t-test, $t = 4.58$ ($df = 11$), $p < 0.0005$, one-tailed.

Eight of the twelve leaders used a lexical marker to identify their intended addressee at some point in their session, although two only used lexical markers when addressing a peer, so were not included in paired analyses of computer- versus human-directed instructions. When only instructions involving a lexical marker were analyzed for amplitude, 182 instructions were available for analysis, and leaders' amplitude for human-directed instructions averaged 66.23dB, significantly lower than 67.95dB for system-directed instructions, a priori paired t-test, $t = 2.35$ ($df = 5$), $p < 0.033$, one-tailed. When only matched instructions not involving lexical marking were compared, 387 instructions were available for analysis, and amplitude to humans averaged 62.96dB, compared with 66.21dB to the computer, a priori paired t-test, $t = 5.57$ ($df = 7$), $p < 0.0005$, one-tailed. As shown in Figure 4.4, the amplitude difference between computer- and human-directed instructions was 3.25dB for instructions with no lexical markers, compared with 1.72dB when a lexical marker was present, or 89% greater.

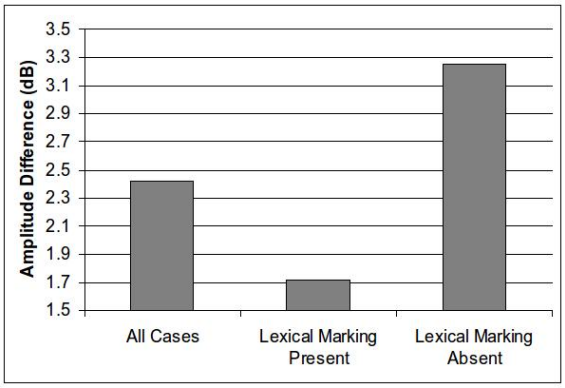


Figure 4.4: Amplitude difference between human- and system-directed instructions for all instructions, and those with lexical marking present versus absent.

4.3.2 Adjacent Utterances

An example of an adjacent utterance pair is shown in Figure 4.5. Leaders increased their amplitude 1.97dB on human-human adjacent utterance pairs, significantly less than the 4.63dB on human-computer adjacent pairs, a priori paired t-test, $t = 3.59$ ($df = 11$), $p < 0.002$, one-tailed. Their amplitude increase was 135% greater when the subsequent instruction was addressed to a computer compared with a human, as shown in Figure 4.6. These amplitude differences were replicated using only those pairs with no lexical marking (1.83dB amplitude increase in human-human pairs, 4.41dB difference in human-computer ones), a priori paired t-test, $t = 1.86$ ($df = 9$), $p < 0.05$, one-tailed. Results also were replicated using only those pairs in which a lexical marker was present, for which the amplitude increase was 2.53dB for human-human utterance pairs versus 5.09dB for human-computer, a priori paired t-test, $t = 2.53$ ($df = 5$), $p < 0.03$, one-tailed.

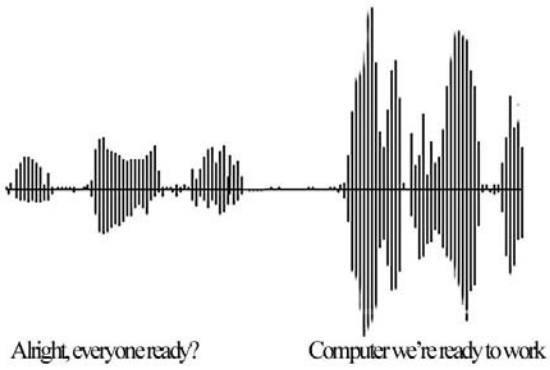


Figure 4.5: Lexical content and wave form of an adjacent utterance pair, with initial utterance addressed to human peers and second one to computer with lexical marking.

As shown in Table 4.2, speakers also altered their amplitude when their target addressee changed from human to computer and vice versa across phase shifts. Speakers' average amplitude dropped 2.5dB between computer- and human-directed instructions, a significant decrease by a priori paired t-test, $t = 4.50$ ($df = 11$), $p < 0.001$, one-tailed. In contrast, amplitude increased 3.1dB between human- and system-directed phases, a priori paired t-test, $t = 4.89$ ($df = 11$), $p < 0.0001$, one-tailed. These analyses confirm the

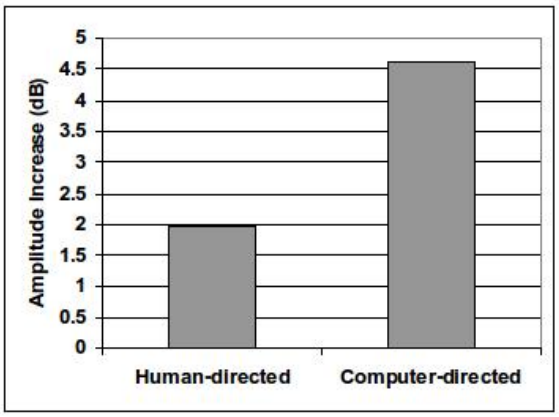


Figure 4.6: Average amplitude increase on adjacent utterance pairs when an instruction was addressed to a human versus computer

bidirectional nature of speakers' use of amplitude shifts to mark their intended addressee.

Table 4.2: Differences in average amplitude for instructions preceding and following target addressee phase shifts.

	Computer followed by Human			**Human followed by Computer**		
Subj	**C**	**H**	**Magnitude Drop**	**H**	**C**	**Magnitude Gain**
1	70.5	68.8	-1.7	68.1	70.8	2.8
2	69.0	70.7	1.6	66.7	70.9	4.3
3	63.1	59.0	-4.0	59.9	60.5	0.6
4	64.8	64.0	-0.8	66.7	70.2	3.5
5	65.8	64.6	-1.1	61.9	65.4	3.5
6	64.0	63.9	-0.2	63.2	65.0	1.8
7	71.3	67.2	-4.1	64.4	73.2	8.8
8	67.4	65.2	-2.3	64.6	65.6	1.0
9	61.7	57.3	-4.4	59.3	62.0	2.7
10	66.7	63.5	-3.1	64.6	64.9	0.3
11	66.1	61.3	-4.8	60.3	62.6	2.3
12	64.6	59.2	-5.4	64.1	70.3	6.3
Ave	66.3	63.7	-2.5	63.6	66.8	3.1

4.3.3 Dialogue Style

Participants used commands when addressing the computer 37.4% of the time versus 35.5% when speaking to humans, not significantly different by paired t-test, $t<1$, NS. As shown in Figure 4.7, a linear regression revealed that participants' ratio of command use when addressing their peers and the computer was highly and significantly correlated

by Pearson correlation, r=0.898, F=41.6 (df=1,10) p<.0001, two-tailed. In fact, 81% of the variance in a participants' likelihood of addressing commands to the computer was predictable by knowing their ratio of commands when addressing humans.

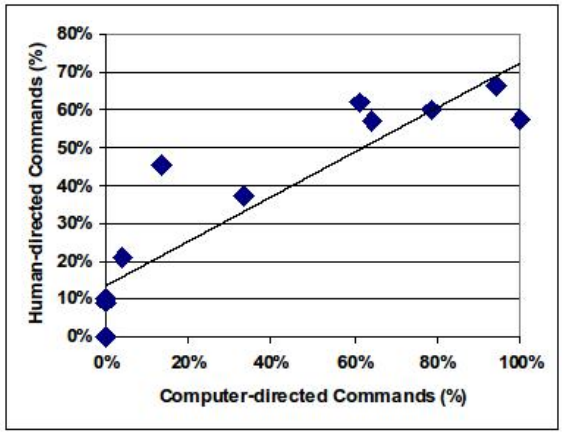


Figure 4.7: Linear regression of participants' ratio of commands when addressing computer versus human.

4.4 Discussion

People use substantially higher amplitude to differentiate computer- from human-directed instructions during computer-assisted group interactions. In fact, in this study they increased their average amplitude when addressing the computer, relative to human-directed instructions, by a substantial 2.4dB overall and by 3.25dB when not also using a lexical marker naming their intended addressee explicitly. When they did use a lexical marker ("computer", "Susan") this difference in amplitude diminished to 1.72dB, indicating that amplitude and lexical marking are alternative strategies used by speakers to clarify their intended addressee. Apart from these differences for instructions carefully matched on illocutionary force, amplitude also increased significantly when speakers shifted between addressing a computer versus human during immediately adjacent utterances and adjacent session phases. Furthermore, speakers shifted their amplitude bi-directionally when

changing between a computer and human addressee, increasing it 3.1dB when shifting between a human and computer addressee, and correspondingly decreasing it 2.5 dB when shifting between a computer and human.

From a functional evolutionary viewpoint, the present results confirm that people use relatively large amplitude changes during communication in order to attract the attention of an intended addressee. Apart from this dynamic change during interpersonal communication, the present results reveal that strong amplitude marking also occurs during more complex computer-assisted group communication as a means of distinguishing an intended addressee, and that during these exchanges the computer is effectively treated as an "at-risk" listener [72]. The large magnitude amplitude changes observed are consistent with the 26 dB difference observed between speech directed to a computer and self-talk during human-computer interactions shown in Chapter 3. In fact, follow-on research exploring the use of normalized amplitudes and amplitude shifts as cues for system engagement during real-time speaker-adaptive processing found amplitude to be a highly reliable indicator of addressee, achieving an 86% correct identification rate [73].

Although amplitude was a strong distinguisher of when a computer was being addressed, the present study revealed that speakers did not direct significantly more command-style instructions to a computer than to human peers, contrary to previous assumptions in the literature. In the present set of instructions matched on illocutionary force, participants issued 37.4% of their instructions as commands to the computer compared with 35.5% to peers. In fact, the frequency of command language clearly was more a function of the individual speaker than their intended addressee, since large speaker differences were observed and there was a high correlation between speakers' frequency of command language to the computer and to their human peers. It may be that the social constraints of an interpersonal meeting may dissuade speakers from abruptly adopting a curt command style when addressing the computer assistant within a group setting.

4.4.1 Lessons for HCI

From an HCI perspective, this study underscores the importance of empirically-grounded models and results in guiding the design of future SDS. By leveraging the dialogue coordination mechanisms that people naturally use to differentiate among computer, human peers, or self as intended addressees, such as the substantial amplitude shifts revealed in the present findings, new engagement techniques can be developed that minimize users' effort in complex mobile and multi-party field environments.

Chapter 5

Human Perception of Human versus Computer Addressee

In spite of interest toward developing computer systems that can identify human- versus computer-addressed speech, there currently is little empirical work exploring what cues humans actually attend to in making such judgments, or how accurate they are in determining who the intended interlocutor is during human-computer group interactions. Most of the current research on detection of intended addressee assumes that the most valuable cues are the same as those used during interpersonal interactions. However, the results shown in Chapter 4 suggest that people do not speak to computers as they do humans.

The general goals of the current study are to gain insights into what audio-visual cues people use to mark their communication as directed to a human versus computer addressee, as judged by a human observer. In particular, this study examines what information people in different age groups (i.e., teenage and adult) use to determine whether videotaped interactions involving utterances matched on illocutionary force were addressed to a computer or a human interlocutor. It also assesses human accuracy levels in making such judgments.

To accomplish these goals, previously collected data was used in which students interacted during a group study session with a computer tutoring assistant and two human peers (as described in Chapter 4). This data was optimal in that speakers were required to produce equivalent instructions to both the computer assistant and to their peers, permitting comparison of human- versus computer-addressed utterances matched on the actions requested.

Using these data, this study explores what audio and visual information people most strongly attend to and how accurately they can identify a speakers' intended addressee using this information. Specifically, we explore:

- People's accuracy and speed in identifying a speaker's intended addressee during different presentation formats, including lexical transcriptions only, audio information with transcriptions, visual information with transcriptions, and full audio-visual playback with lexical transcriptions. We anticipate that participants' judgments will improve as more information sources are made available, but that visual information would support lower accuracy in identifying human-directed utterances in mixed human-computer groups since speakers tend to look at the computer even when addressing humans, as shown in Chapter 4 and in previous work by Katzenmaier et al. [47] and van Turnhout et al. [99].

- Whether additional contextual information about the group's interaction would be beneficial in determining intended addressee. Participants' accuracy was compared between visual presentations showing only a close-up of the speaker versus the speaker and their peers.

- What information people thought was most valuable in deciding whether the speaker was addressing a computer or human. Participants were asked to note during their session what information they used in making their decisions. In addition, post-session questionnaires elicited their ranking of how important different information sources were in judging the intended interlocutor during different presentation formats.

The work in this chapter is based on an earlier work [59]: Human perception of intended addressee during computer-assisted meetings, in *Proceedings of the 8th international conference on Multimodal interfaces* (ICMI '06). ©ACM, 2006. http://doi.acm.org/10.1145/1180995.1181002.

5.1 Methods

5.1.1 Participants

Participants in this study included 8 teens 14 to 17 years of age, and 8 adults 31 to 55 years old. They were paid volunteers, and all were native English speakers.

5.1.2 Tasks

People were told that they were participating in a study to help improve understanding of what information people use to decide who a person is addressing when they speak in a meeting. To accomplish this, they viewed short digital recordings of students interacting during a group study session with two human peers and a computer tutoring assistant. For each clip, participants were asked to identify whether the speaker was addressing a human or the computer assistant. They were also given the option of selecting "other" if they believed the speaker was addressing neither the computer nor a peer (e.g., talking to herself).

Each session was divided into four sections. In all sections, a text transcript of the speaker's utterance was visible for the participant to read. In the first section, the utterance transcript was the only information available to the participant. During the second and third sections the participant could either see the video along with a transcript but not hear the audio, or hear the audio with lexical content but not see the video. In the fourth section, the participant could both see and hear the speaker as she delivered a particular utterance.

Figure 5.1 illustrates the interface used by participants for making these judgments. Video segments were displayed in the middle of the screen. A panel at the bottom of the screen contained (a) video control buttons for playing/pausing and stopping the video, (b) the lexical content of the speakers' utterance, (c) radio buttons for providing answers judging the speaker's intended addressee, and (d) a submit button to input the participant's selection and move on to the next clip.

When video was presented, two different views were used. In one view, only the speaker was visible, as illustrated in Figure 5.1. In the other view, the two peers were also

displayed in relative locations to the speaker, as illustrated in Figure 5.2. Within each section, half of the video segments were presented in the speaker-only view, and the other half in group view, and were not intermixed.

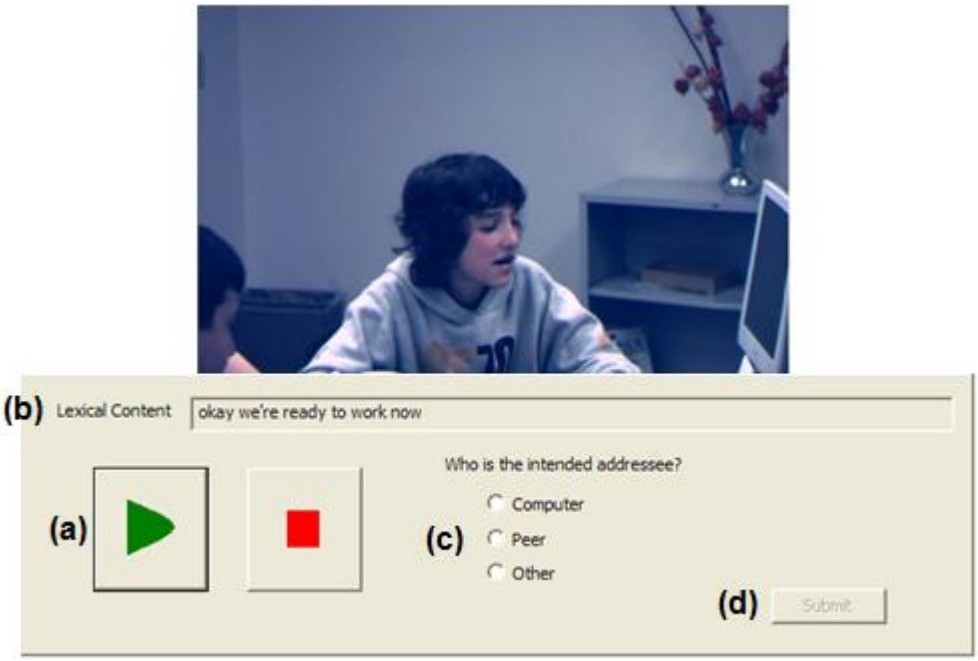


Figure 5.1: Interface showing speaker-only video view.

Figure 5.2: Alternate group view, showing the speaker (right) and two peers (left and center)

5.1.3 Video Data

The video segments presented to the participants were captured during the study described in Chapter 4. Ninety-six unique video segments were created and balanced to contain equal quantities of human- and computer-addressed speech from nine different speakers. These were divided into four groups containing 24 video segments each, with each group containing an equal number of segments from any given speaker. Within each group, the segments were ordered so that no video segments from the same speaker were presented sequentially.

5.1.4 Procedure

Prior to training on the system interface, participants were given an illustration of how the meeting room was set up with respect to participants' locations in the video segments. It was pointed out that in all the videos the computer was to the speaker's left and the human peers to the speaker's right, as illustrated in Figure 5.2. In addition, the experimenter emphasized that the speakers they would be viewing requested the same types of actions from both the computer and peers. Specifically, the experimenter said, "Sometimes the speaker talked directly to the computer and it would handle their requests, and other times they asked one of their peers to complete the request. This means there's no difference in what the speaker is asking for, just in who they're asking to do it, either computer or human peer."

During training, participants were given eight tasks to familiarize themselves with the system and the different presentation formats, two audio-visual, two visual only, two audio only and, two text only. After the participant completed training, the experimenter left the room and the participant completed interlocutor judgments. At the end of each presentation format, a screen was displayed which instructed the participant to contact the experimenter to continue the session. At this point, participants were instructed to indicate on a form what information they had used to make their judgments. The experimenter then returned and reviewed these notes, clarifying participants' feedback as needed before initiating the next session phase.

At the end of the study session, participants were given a questionnaire that listed potential lexical (text), audio, and visual cues of intended addressee, as shown in Figure 5.3. The list of cues was derived from comprehensive participant feedback gathered during early pilots. Participants completed the questionnaire by ranking pre-defined cues in order of how important the cue was in making his or her decisions during each presentation format. The questionnaire had columns for each presentation format, and rows for each cue. Cells were grayed out when not applicable to a given presentation format.

Instructions: Below are some cues that others have said they used in determining whether a speaker was addressing the computer or a human. If you believe you used any of these, please rank them in order of importance for that part of the session, with "1" being most important.

		Text Only	Audio	Video	Audio Video
Content	Fluency of the speech, (no ums, uhs or corrections)				
	Politeness terms (please, thank you)				
	Number of words				
	Conversational versus command style				
Visual	Speaker gaze: Looking at computer or glancing at peers				
	Peers' gaze: Looking at computer or glancing at speaker				
	Peers' reaction: Expression or movement				
Audio	Loudness of the speech				
	Careful pronunciation				
	Tone of voice: question or command				

Figure 5.3: Post-session questionnaire.

5.1.5 Research Design

The experimental design involved a within-subject comparison of speaker's judgment accuracy and the number of times they played the segments when making judgments as a function of: (1) Presentation format (lexical transcript only, audio/lexical, visual/lexical, and audio-visual/lexical); (2) Speaker addressee (computer or peer), and (3) Video view (close-up of speaker only; group view of speaker and peers). Questionnaire feedback was also compared as a function of presentation format. An additional between-subject factor was: (4) Age group (teen, adult).

Order of presentation was counterbalanced between the four video segment groups, the audio/lexical and visual/lexical presentation formats, and the speaker-only and group-view video presentation formats.

5.1.6 Data Capture, Coding and Analysis

Data for each session were logged automatically for the number of times each video segment was played and the participants' judgment of addressee (computer, peer, or other).

Correct Judgments

Each participant's addressee judgments were coded as either correct or incorrect in relation to the speaker's actual known addressee, and a percentage correct was calculated for each presentation format.

Speed of Judgments

The speed of a participant's judgments was estimated from the average number of times he or she played the video segments in each presentation format (excluding text only), when making judgments about interlocutor.

Questionnaire Rankings

Composite rankings were calculated by averaging participants' rankings of each information source within each presentation format. Prior to calculation, information sources not ranked by a participant were assigned a value of one over maximum rank score for that

presentation format. For example, if the participant only ranked 2 of the potential information sources for the "Text Only" presentation format, the remaining 2 would would be assigned a score of 5 (i.e., 4 possible+1). To ease comparison between the scores for the different presentation formats, the scores for each presentation format were scaled to fit a 10-point range, and inverted so that 10 represented the most important cue and 1 the least important.

5.2 Results

In total, data were available for analysis on 1536 judgments of intended addressee, and 28 questionnaire rankings for each of the 16 participants.

5.2.1 Correct Judgments

Teen judgments of intended addressee were correct for 45% of the segments during lexical-only presentation, 59% during audio/lexical, 56% during visual/lexical and 59% during audio-visual/lexical. Adults were correct for 47% of the segments during lexical-only presentation, 57% during audio/lexical, 58% during visual/lexical, and 67% during audio-visual/lexical. A comparison of the accuracy of teens versus adults revealed no significant differences for any presentation format, independent t-test, all t's < 1.3, NS. For this reason, all further accuracy analyses are collapsed over age group. Additionally, speakers averaged 59% correct when the video view was speaker only, not significantly different than the 61% correct when the video view included the group with speaker and peers, paired t-test, $t < 1$, NS. Therefore, further accuracy analyses also were collapsed over view. Participants selected "other" for 3% of both the human- and computer-directed segments. For the present analyses, judgments of "other" were treated as incorrect judgments.

As shown in Figure 5.4, overall participants were correct in their identification of the speaker's intended addressee for 46% of the utterances when presented with lexical only, 58% for audio/lexical, 57% when video/lexical, and 63% for audio-visual/lexical. The two presentations involving a unimodal signal (i.e., audio only and video only) were not significantly different by paired t-test, $t < 1$, NS, so were collapsed in further follow-up

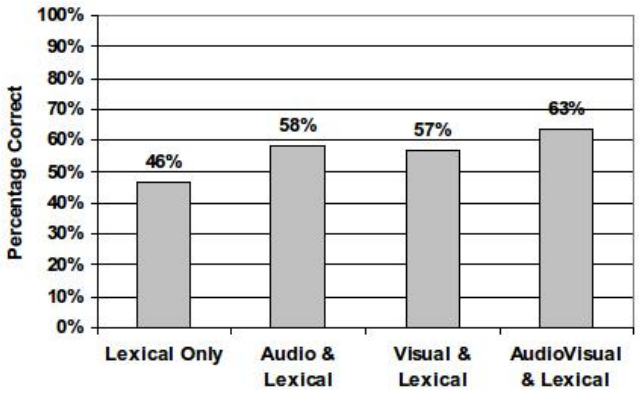


Figure 5.4: Participants' overall percentage of correct judgments in each condition.

analyses on judgment accuracy.

Overall, participants were correct significantly more often when a unimodal audio or visual signal was available (58%) than when only lexical content was available (46%), a-priori paired t-test, t=3.66 (df=15), $p < 0.001$, one-tailed. However, participants were correct only marginally more often when an audio-visual signal was available (63%) than when a unimodal signal was available, paired t-test, t=1.53 (df=15) , $p < 0.08$, one-tailed.

Figure 5.5 shows participants' percentage correct when the speakers' addressee was a human peer. For these cases, participants actually were correct most often when the presentation format was audio/lexical (46%), which was significantly better than lexical only, a-priori paired t-test (35%), t=1.87 (df=15), $p < 0.05$, one-tailed, and visual/lexical (34%), a-priori paired t-test, t=2.12 (df=15), $p < 0.03$, one-tailed, but not significantly better than during the audio-visual/lexical presentation format (40%), $t < 1$, NS. In addition, participants' judgments were not significantly better during the audio-visual/lexical format than during either visual/lexical or lexical only, all t's < 1.4 NS.

Figure 5.6 shows participants' percentage of correct judgments when the speaker's addressee was the computer. For these cases, participants were correct significantly more often during audio/lexical presentation (70%,) than during lexical only (58%), a-priori paired t-test, t=2.66 (df=15), $p < 0.01$, one-tailed. In addition, participants were correct significantly more often during visual/lexical (82%) than during audio/lexical, paired t-test, t=1.82 (df=15), $p < 0.05$, one-tailed, or lexical only, a priori paired t-test t=4.07

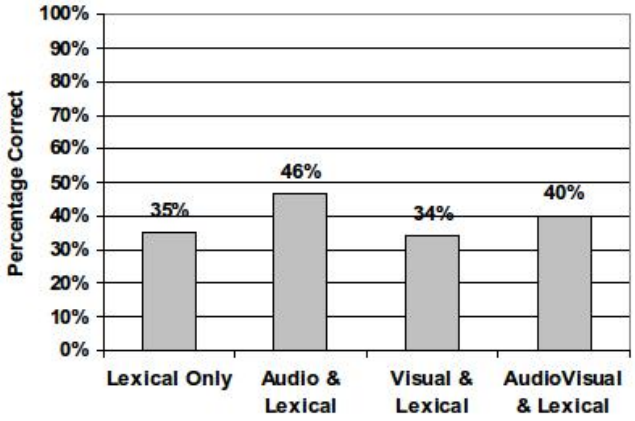


Figure 5.5: Participants' percentage of correct judgments when the speaker's addressee was a human peer.

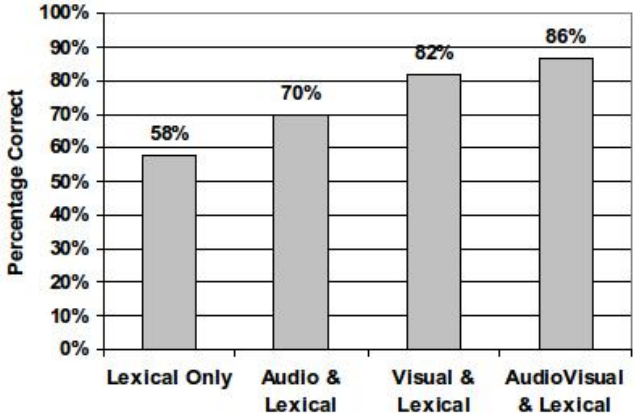


Figure 5.6: Participants' percentage of correct judgments when the speaker's addressee was the computer.

(df=15), $p < 0.001$, one-tailed. Participants' percentage of correct judgments was not significantly different between the visual/lexical and audio-visual/lexical (86%) presentation formats, $t < 1.2$, NS. Participants also were correct significantly more often during audio-visual/lexical presentations than during audio/lexical, paired t-test, t=3.23 (df=15), $p < 0.005$, one-tailed, or lexical only, a priori paired t-test t=6.66 (df=15), $p < 0.0001$, one-tailed.

Finally, overall participants were correct in their judgments for those segments in which

the addressee was human for 39% of the segments, significantly less often than their 71% rate of correct judgments for those segments in which the addressee was the computer, paired t-test, t=7.39 (df=15), $p < 0.0001$, two-tailed.

5.2.2 Speed of Judgments

Adults and teens were not significantly different in the number of times they played the segments, with teens averaging 2.06 plays and adults 2.37 plays, independent t-test, $t < 1$, NS. Further analyses on number of plays were collapsed over age group. As shown in Figure 5.7, participants played the segments significantly more often when the video view was the group of speaker and peers (2.63 times) compared with speaker only (1.80 times), a-priori t-test, t=5.73 (df=15), $p < 0.0001$, one-tailed. Participants also played the segments on average 1.44 times during audio-only presentation, significantly less than during the speaker-only visual (1.80), paired t-test, t=3.01, $p < 0.005$, one-tailed, or the group visual (2.63), paired t-test, t=6.47, $p < 0.0001$, one-tailed.

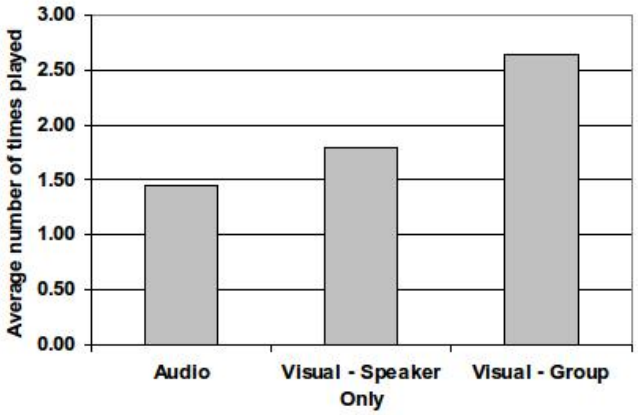


Figure 5.7: Participants' number of replays when making interlocutor judgments during audio and visual conditions (speaker versus group).

5.2.3 Questionnaire Ranking

As shown in Figure 5.8 for the lexical-only presentation format, participants' ranking of the importance of fluency of speech (i.e., no ums, uhs or corrections) was 5.85, and it was 5.68

for dialogue style. Participants' average ranking for these two cues was significantly higher than the 3.35 mean ranking for each of the other two lexical cues cited by participants, Wilcoxon signed rank test, $z = 2.6$, $p < 0.01$, two-tailed.

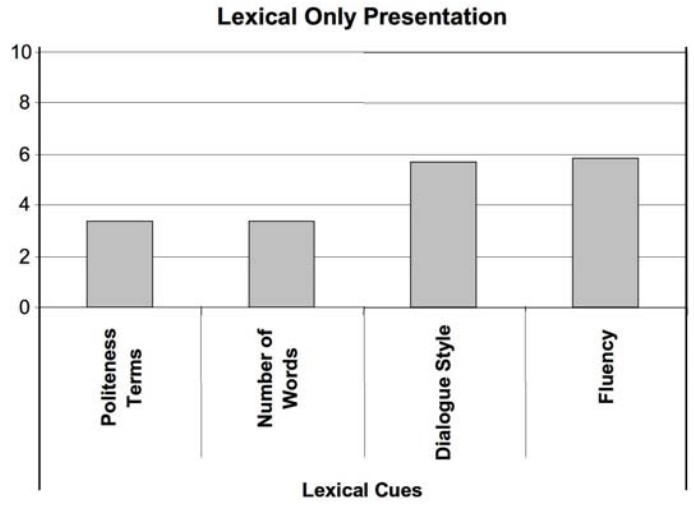


Figure 5.8: Participants' ranking of cue importance during lexical-only presentation.

As shown in Figure 5.9 for the audio/lexical presentation format, participants ranked "tone of voice" (intonation) 7.96, significantly higher than the mean of the two top ranked lexical cues, which were fluency (5.56) and dialogue style (5.81), Wilcoxon signed rank test, $z = 2.42$, $p < 0.02$, two-tailed. That is, when audio information was added, this audio information source was considered more valuable than the available lexical information sources. In addition, intonation was ranked higher than the mean of the two other audio cues, speech loudness (3.94) and careful pronunciation (4.66), Wilcoxon signed rank test, $z = 2.80$, $p < 0.005$, two-tailed.

As shown in Figure 5.10 for the visual/lexical presentation format, participants ranking was 8.59 for speakers' gaze, significantly higher than the mean of the other visual cues, peers' gaze (6.36) and peers' movement (5.20), Wilcoxon signed rank test, $z = 3.21$, $p < 0.001$, two-tailed. Speakers' gaze also ranked significantly higher than the mean of the highest ranked lexical cues of fluency (3.64) and dialogue style (4.17), Wilcoxon signed rank test, $z = 2.59$, $p < 0.01$, two-tailed. However, the mean of peers' gaze and peers' movement was not significantly higher than an average of the top lexical cues of fluency

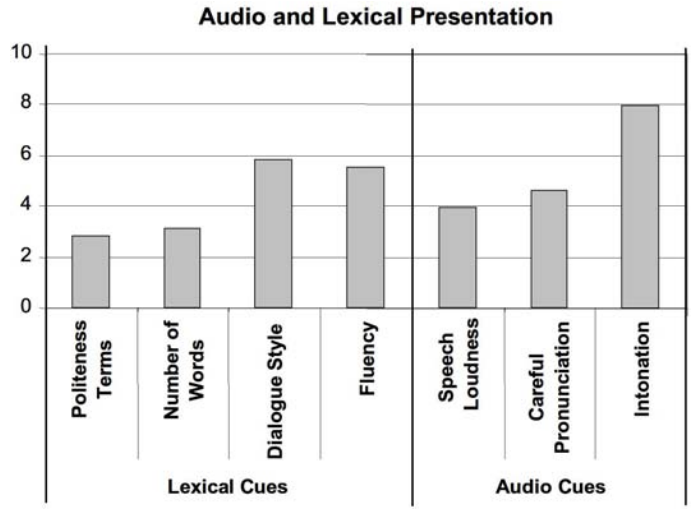


Figure 5.9: Participants' ranking of cue importance during combined audio/lexical presentation.

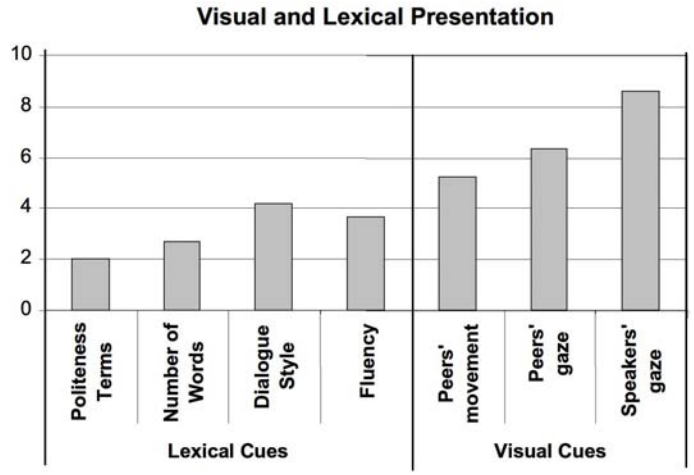


Figure 5.10: Participants' ranking of cue importance during combined visual/lexical presentation.

and dialogue style, Wilcoxon signed rank test, $z < 1.6$, NS.

Figure 5.11 shows participants' ranking of cues for the audio-visual/lexical presentation format. Of all available cues, speakers' gaze was ranked most important at 8.44, although it was only marginally higher than the second-highest ranking of 6.56 for intonation, Wilcoxon signed rank, $z = 1.82$, $p < 0.07$, two-tailed. In addition, speakers' gaze was ranked significantly higher than the third-highest ranked peers' gaze at 6.12, Wilcoxon

signed rank test, z = 3.37, p < 0.001, two-tailed. Also, the mean of the two highest lexical cues, which were fluency (4.00) and dialogue style (5.81), was significantly lower that the mean of speakers' gaze and intonation, Wilcoxon signed rank test, z = 2.62, p < 0.009, two-tailed.

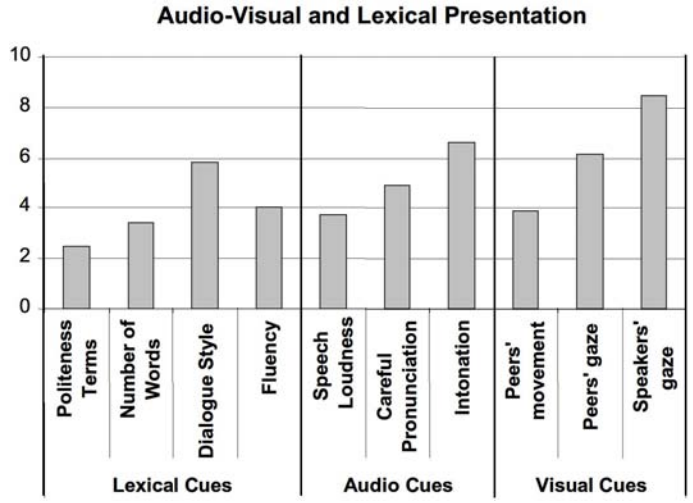


Figure 5.11: Participants' ranking of cue importance during audio-visual/lexical presentation.

5.3 Discussion

As anticipated, human observers' overall ability to correctly judge an intended interlocutor during computer-assisted group interactions improved as more information sources became available, from text-only, to unimodal (audio or visual), to combined audio-visual. Perhaps surprisingly, people's accuracy in determining human versus computer interlocutors did not exceed chance levels with lexical-only content (46%). In comparison, accuracy improved significantly with both audio (58%) and visual (57%) unimodal signal information added, and further with combined audio-visual information (63%). As we showed in Chapter 4, speakers do not direct more command-style language to a computer than to human peers during mixed group interactions, which may account for the chance-level accuracy level observed in the lexical-only condition.

When the accuracy of judging human versus computer interlocutors was analyzed separately, it was apparent that people's accuracy levels in detecting human interlocutors was significantly worse than that for computer interlocutors during these mixed group interactions. As shown in Figure 5.5, people systematically failed to correctly identify human interlocutors whenever visual information was present because speakers often looked at the computer when addressing their peers. In particular, performance levels for judging human interlocutors were significantly better in the audio condition than the visual one. These results reveal that people rely on cues appropriate for interpersonal communications when judging computer versus human interlocutors in mixed group contexts, even though this default model degrades their accuracy in judging human interlocutors.

In contrast, accuracy in judging computer interlocutors actually was significantly better whenever visual information was present than with audio alone, and it yielded the highest accuracy levels observed (86%). However, this apparently higher level of accuracy in judging the computer as an intended interlocutor to a large extent may have been due to the higher overall base-rate of gazing at the computer, which provided the group with visual feedback. That is, as documented in other work [99], gaze can be a misleading indicator of an intended addressee whenever its deployment also is required to complete basic aspects of a task. During any mixed human-computer group interaction in which the computer delivers visual feedback, a built-in gaze asymmetry typically exists that should not be misconstrued as a higher level of correctly detecting the computer as the intended addressee. In this sense, the apparently higher accuracy levels of detecting the computer as an intended addressee were spurious.

Apart from judgment accuracy, participants' questionnaire data revealed that they believed speakers' gaze, peers' gaze, and tone of voice were the most valuable information sources for determining an intended interlocutor. Secondarily, among lexical cues people believed that dialogue style (i.e., command versus conversational) and fluency also were valuable in making such judgments, especially if audio or visual signal cues were unavailable. These self-reports confirm that people attempt to base their judgments on cues relevant to interpersonal group interactions, even when observing human-computer

interaction. Of course, both speakers' and peers' gaze would be misleading during judgments involving human-computer group interactions, for the reasons outlined previously. In contrast, the highly ranked intonation cue clearly assisted people in making relatively accurate interlocutor judgments in the audio-only condition, during which time misleading gaze cues were absent. Given that our earlier research (Chapters 3 and 4) revealed that amplitude is a powerful cue of intended interlocutor, it is perhaps surprising that people did not cite it as an important information source. This may have been due in part to people's lack of explicit awareness of changes in this paralinguistic cue. People also may have been less aware of amplitude changes indicating an intended interlocutor because audio playback during this study involved individual utterances without surrounding dialogue context.

The work in this chapter has had an impact on the field. Terken et al. [97], building on our work, explored how well a human observer could identify the addressee in multi-party conversation with all humans. In this case, the multi-person interaction consisted of two 'friends' planning a trip together a human travel agent. This work found that gaze, in concert with the lexical content of the speech, resulted in the most accurate judgments of addressee. This contrary result is probably due to the computer, in our work, not producing gaze cues itself and also being the source of task-critical information.

5.3.1 Lessons for HCI

From an HCI perspective, this chapter serves as a precaution to designers of multimodal SDSs. The results suggest that people have a predilection to base judgments of an intended addressee on their experience with interpersonal communication cues. However, this predilection resulted in systematic errors when asked to differentiate human and computer addressee during human-computer group interactions. These errors were likely due to mis-interpreting gaze at the computer to imply that the speaker is addressing the computer. Thus, it seems clear that SDS designers will need to be cautious to avoid assumptions as to what cues people use to differentiate addressee during human-computer group interaction, if those assumptions are based on human-human communication.

However, as shown in Chapter 4, people do produce reliable cues of human versus

computer addressee. That is, the speakers are producing cues, just not those that might be expected based on human-human communication. Thus, it's likely that future systems, that process actual rather than expected communication patterns, could be designed that perform as well as, if not better than, human observers.

Chapter 6

Distinctions Between 'um' and 'uh'

In this chapter, we focus on the fillers 'um' and 'uh'. Previous researchers have suggested that fillers could be used by a speaker when managing his own communication, or used to help coordinate the dialogue with another, as we discussed in Chapter 2. In earlier work, in which I was a second author [38], we used data from children with either Typical Development (TD) or Autism Spectrum Disorder (ASD), finding that children with ASD seem to use 'uh's in a manner similar to those with TD, but not so with 'um'. This led us to posit that 'um' and 'uh' might arise from different cognitive processes. We included data from children with ASD because they have social impairments, and so might also have impairments in their dialogue coordination mechanisms.

The goal of this chapter is to further understand the role that 'um' and 'uh' each play in dialogue coordination, specifically whether they are used as dialogue coordination mechanisms, or are speaker-oriented and produced without intent to inform the listener. To this end, we continue investigating the differences between how 'um' and 'uh' are used by children with TD and ASD. We also include children with Developmental Language Disorder (DLD), as they have language impairments, as children with ASD have, but do not have social impairments. The specific aims of this chapter are to determine:

- Whether diagnosis (i.e., TD, DLD, and ASD) has a significant effect on a speaker's ratio or 'um's and 'uh's, and whether the filler usage of children with DLD will more closely resemble that of children with TD, or those with ASD.
- Whether the ratios of 'um's and 'uh's differ based on location within the C-unit (i.e., beginning of turn, beginning of utterance, or utterance-medial).

- Whether the differences between 'um' and 'uh' can be explained by factors (e.g., age, sex, mean utterance length) other than those previously found relevant, and if these factors interact with diagnosis.

- Whether the pauses following 'um's are more prevalent, and longer, than the pauses following 'uh's, as predicted by Clark and Fox Tree [23].

The research in this chapter can serve two distinct purposes. First, by comparing the spoken language of children with language and social impairments, we will gain insights into the underlying cognitive processes that control these aspects of language. In doing so, we can inform SDS design as to when and how fillers should, and should not, be used, and how they can be interpreted when present in a user's speech. Second, although not part of the main goal of this dissertation, it can assist in the diagnostic process for ASD and DLD. The tools currently used in diagnosis of ASD and DLD are primarily subjective. Easily administered quantitative analyses would be a helpful adjunct to the diagnostic process.

The work in this chapter subsumes that published in *"Autism and the Use of Fillers: Differences Between 'um' and 'uh' "* [58]. That paper subsumed the work on fillers included in *"Autism and Interactional Aspects of Dialogue"* [38], on which I was the second author.[1]

6.1 Background and Related Work

6.1.1 The Role of Fillers in Dialogue

Many researchers have explored the role of 'um' and 'uh' in dialogue. This research typically uses one of three approaches: 1) examining the occurrence of fillers in a dialogue,

[1] My contribution to those works, in terms of filler analyses, was to use subject means (rather than raw, unbalanced data) in the group-wise analyses, and to perform within-subject comparisons. In addition, I replaced t-tests with non-parametric analyses, and examined the length of pauses after fillers. We extend those works here by including children with DLD, and performing more sophisticated statistical analyses. Specifically we use linear mixed-effects regression models to account for individual differences and for potential confounds such as sex, MLU, and age.

with the goal of identifying regularities in the use of fillers; 2) examining how fillers effect the listener's behavior; and 3) evaluating how fillers impact overhearers impressions of the speaker. We now review research using each of these approaches.

By examining the occurrence of fillers in a dialogue, researchers provide insight into the ways in which fillers may be used to manage dialogue coordination. Smith and Clark [88] found that fillers are more likely to precede an answer that the speaker is not confident about. Swerts [95] found that fillers appear to carry information about topical units in a dialogue, with stronger breaks in the discourse more likely to co-occur with fillers than weaker breaks and that fillers at strong breaks are more likely to be preceded and followed by pauses. Goldman-Eisler [34] suggested that fillers signal a speaker's word-searching problems and Stenström's [89] work suggests that fillers can function as turn-holders. Barr [13] found that speakers were more likely to precede a new referent with an utterance-initial 'um' and an old referent with an initial 'uh'.

Examining the effects of fillers on listener behavior provides insight into how fillers are interpreted. Barr [13] shows that listeners will anticipate new information if the speaker says 'um' prior to a referential description, but 'uh' does not produce the same effect. Arnold [5] found that listeners were more likely to look at a new (rather than previously mentioned) object when the object name was preceded by "... thee, uh,". Kidd et al. [48] found that children as young as two years would look toward a previously un-mentioned object if the speaker preceded an object name with "... thee, uh,", but that younger children did not. Corley [26] substantiated the concept that disfluency impacts a listener's comprehension, and interestingly, that words preceded by "... thee, uh" were more readily recognized later.

Examining the impressions of overhearers provides insight into whether, and if so how, fillers impact other's perceptions about the speaker. Fox Tree [32] examined multiple theories of how the use of fillers and pauses impacted an overhearer's impression of a speaker's speech production difficulty, honesty, and comfort with the given topic. This work found that preceding a statement with 'um' generally led to less positive impressions of the speaker.

It is important to note that research described above does not identify whether speakers

produce fillers with the intent to provide additional information to listeners (listener-oriented). It is instead possible that listeners take opportunistic advantage of (speaker-oriented) fillers, making inferences based on evidence that the speaker is having difficulty. Thus, the question remains as to *why* speakers produce fillers.

6.1.2 Theories of Filler Production

There are two general theories as to why speakers produce fillers. First, fillers have historically been viewed as speech errors [51] or as symptoms of speech production difficulty, with 'um' related to deeper planning problems than 'uh' [34]. From a dialogue coordination perspective, this viewpoint casts fillers as part of the speaker's own communication management [3]; not as a mechanism used to intentionally inform a listener of the speaker's processing state.

Second, in contrast, more recent work by Clark and Fox Tree presented the "filler-as-word" hypothesis [24]. This hypothesis states that fillers are *words* that are used to announce the initiation of what is expected to be a delay in speaking, with 'uh' signaling a minor delay and 'um' signaling a major delay. Clark and Fox Tree suggest that, when using fillers, speakers' fulfill a specific social obligation; that of informing the listener of upcoming delays in the speaker's speech. To support this hypothesis, Clark and Fox Tree compared speakers use of 'um' and 'uh', in a study primarily using the London-Lund corpus of face-to-face conversations [24]. They found that (a) speakers use fillers most often near utterance boundaries, primarily 'um's, (b) 'um's are more likely to be followed by a pause than 'uh's, and (c) pauses following 'um's are longer than pauses following 'uh's.

Although differing in how they explain the source of hesitations, these theories (and most of the previous research) assume that the two fillers have a common source, with a speaker's choice of 'um' versus 'uh' being driven by personal preference and level of difficulty or anticipated hesitation.

However, our recent research suggests that 'um' and 'uh' may result from different sources [38]. In that work we found that the rate of 'uh's was similar between the two groups, but that the children with ASD had a significantly lower rate of 'um's. From these

results, it appears likely that 'um' and 'uh' arise from different cognitive processes, and that the process responsible for 'um' is effected by ASD, but the process responsible for 'uh' is not. However, the question remains whether these differences are due to impairments in the ASD groups' ability to learn social cues, or in their ability to formulate their speech.

6.1.3 ASD, DLD and Spoken Language

To help illuminate whether differences in filler usage are related to social or processing impairments, in this chapter we include in our analyses children with Developmental Language Disorder (DLD). ASD and DLD are both conditions that impact a child's ability to engage in spoken communication. Children with ASD are characterized by impaired reciprocal social interaction and communication, repetitive behaviors, and restricted interests according to the APA's DSM-IV-TR [4]. In terms of communication, even high-functioning children with ASD, who have semantically and syntactically correct spoken language, will have faulty pragmatics due to difficulties in understanding and using social cues during conversations. Children with DLD are characterized by an inability to communicate in a manner appropriate to the child's age, in which the inability is not attributable to physical or intellectual impairments, or ASD [4]. These inabilities can include shortcomings in expressive language, resulting in difficulties organizing and formulating an utterance, or in receptive language, leading to difficulty in comprehending others' speech.

In terms of diagnosis, researchers have noted "... the lack of a clear dividing line between language disorders and autism." [18], and that "... a clearer understanding of the factors that are markers of ASD and those that differentiate groups of children with DLD are needed" [67]. To illustrate, a child with DLD may appear to have restricted interests (a marker of ASD), but instead be limited by a restricted vocabulary and syntax, outside of which it is difficult for them to meet the social demands of spoken interaction (e.g., timely responses). To further complicate diagnosis, the dividing line between DLD and ASD has changed as diagnostic criteria have evolved. Bishop et al. [18] found that in a group of 38 adults who had received a diagnosis of DLD as children, 12 would meet current criteria for ASD. In addition, the research of Mouridsen and Hauschild [67] suggests that individuals with DLD may have increased vulnerability to ASD.

Difficulty in the differential diagnosis of ASD versus DLD is to be expected as the language skills, and impairments, of children with ASD and DLD are quite varied and will often exhibit the same, or similar, linguistic features. However, despite these overlaps, there is a substantial difference; that of impaired pragmatics. That is, we can expect that children with ASD will often display language impairments similar to those seen in children with DLD, but children with DLD will not display the same high degree of faulty pragmatics seen in children with ASD [79].

Essentially, children with DLD have impaired language processing skills, but do not have the social impairments seen in ASD. Thus, if their use of fillers more closely resembles that of children with TD, this provides evidence that the difference in usage of 'um' between children with TD and ASD is attributable to social impairments. Conversely, if their filler usage more closely resembles that of children with ASD, this provides evidence that the differences are attributable to processing impairments.

Analyses of the pragmatic impairments of children with ASD have generally focused on higher-level features of language such as the appropriateness of a response or whether the child is overly talkative [75]. In this chapter, we instead look at pragmatics from the viewpoint of dialogue coordination mechanisms, anticipating that impairments in social interaction will also manifest in the content of the child's utterance and his speech timing. Specifically, we compare the use of fillers and post-filler pauses between children with TD, ASD, and DLD.

6.1.4 ASD and Fillers

As children with ASD have, by definition, social impairments, the difference found in the usage of 'um' between children with TD and ASD in our previous work suggests that 'um' is listener-oriented, whereas 'uh' is not. However, children with ASD also suffer deficits in executive functioning and processing [77]. Thus, although our previous work suggests that 'um' and 'uh' are not produced by the same cognitive process, the question remains whether the process responsible is related to social or processing impairments.

The hypothesis that fillers are listener-oriented (i.e., social in origin) is supported by work showing decreased use of fillers in speakers with ASD, when compared to speakers

without social impairments [50]. However, this work did not separate 'uh' and 'um', and included only responses to questions. Thus it is unclear whether the participants used 'um' and 'uh' differently.

6.1.5 ASD and Private Speech

Children with ASD are known to have deficits in both executive functioning and the use of social speech. In contrast, recent work has suggested that children with ASD have relatively unimpaired private (i.e., self-directed) speech [104]. High-functioning children with ASD were found to use private speech at the same rate as children with typical development. In addition, in contrast to children with TD, children with ASD were more likely to get items correct when talking versus when silent [104]. This works suggest that children with ASD use private speech to bolster their executive functioning and improve task performance. Thus, if a filler were to arise from a cognitive process responsible for producing private speech, we could expect children with ASD to use fillers in a manner similar to children with TD. In contrast, if fillers arise from a cognitive process responsible for the use of social speech, we would expect children with ASD to use fillers less often and less effectively than children with TD.

6.2 Methods

6.2.1 Participants

One of the goals in collecting this data is to determine what markers delineate children with TD, DLD, and ASD. Because of this, the protocol used to determine participant inclusion in this research is particularly stringent. First, the children are diagnosed. The children with ASD are diagnosed using tests administered and scored by trained clinicians, in concert with an Autism-specific parent questionnaires. In addition, children participate in an extensive protocol that gathers neurocognitive and developmental information via standardized tests and parent questionnaires. Results of these tests are used in a consensus process whereby a group of clinicians and therapists trained in various disciplines meet to discuss and agree on each child's diagnosis. Second, to be included in the study, children

must have an IQ above 70,[2] be "verbally fluent", and must have no neurological conditions or gross motor impairments. In addition, a child is excluded from the TD group if any immediate family member has been diagnosed with DLD or ASD, or the child has a history of psychiatric disturbance (e.g., ADHD). Data for 91 children (27 with TD, 21 with DLD and 43 with ASD) ranging in age from 3.9 to 8.9 were available for analyses.

6.2.2 Activities

The data in this study was collected while children were engaged in the Autism Diagnostic Observation Schedule (ADOS) [55] with a trained clinician. The ADOS is designed to engage the child in interpersonal interaction during different activities, allowing the clinician to observe and assess the child's communication skills. Sessions typically lasted one to two hours.

The ADOS activities used in this chapter consist of having a conversation, describing a wordless picture or book, and playing with toys. Each of these activities is designed to allow the clinician to observe different communication skills. The *conversation* activity consists of interactions in which the clinician asks the child about their personal experiences with different emotions and the physical manifestations of emotion. This activity allows the clinician to observe how aware the child is of his own emotions. During the *description* activity, the clinician and child peruse and discuss a wordless picture/book. The clinician then tells a story from her past, relating it to the picture/book. In this activity, the clinician observes whether the child engages the clinician, interacting with the clinician to gain a better understanding the story. During the *play* activity, toys are made available for the child and clinician to engage in joint play. This activity allows the clinician to observe how well the child engages in interactive play and maintains a joint activity.

[2]For the children with ASD, an IQ above 70 classifies them as "high-functioning".

6.2.3 Research Design

The experimental design involved a within-subject comparison of the ratio of a speaker's 'um's versus 'uh's to total words overall and as a function of: (1)Activity (conversation, description, or play); and (2) Position (turn-initial, utterance-initial or utterance-medial). An additional between-group factor was: (3) Diagnosis (TD, DLD, or ASD).

To account for both potential confounds and the factors of interest, an additional experimental design was correlational, predicting the ratio of 'um's, or 'uh's, as a function of (1) Activity, (2) Age, (3) Sex, and (4) Mean length of utterance (MLU).

The length of pauses after 'um's versus 'uh's were compared both within-subject and between-groups.

6.2.4 Data Capture, Coding, and Analysis

The speech recordings used in this study were captured using a single microphone placed near the participants.

C-units

The child's and clinician's speech was transcribed and using Praat [19]. The speech was segmented into communication-units (C-units) [83] and annotated with a start and end time. A C-unit includes one main clauses and all the subordinate clauses attached to it. C-units were transcribed with a '.', '!', '?' to mark semantically and syntactically complete sentences, '>' to mark incomplete ones, and ':' to mark the start of an intra-unit pause.

'um's and 'uh'

'Um's and 'uh's were included in the transcriptions of the children's speech. The fillers were counted, as were the number of words, and a ratio of 'um's and 'uh's to words computed for each child.

Activity

Sections of the session in which the clinician and child were engaged in each of the activities were annotated on a separate tier using Praat [19]. This information was aligned with

the C-units, and the activity (i.e., conversation, description, or play) identified for each C-unit.

Position

All 'um's and 'uh's were annotated as to their position within the dialogue: turn-initial if the filler immediately followed a C-unit produced by the clinician, and utterance-initial if the filler followed a C-unit produced by the child. All other fillers were annotated as utterance-medial.

Pauses after 'um's and 'uh'

For fillers that were annotated as preceding an intra-unit pause (i.e., those immediately followed by a ':'), pauses were calculated as the start time of the following utterance minus the end time marked after the filler. Pauses that were not annotated using a ':' (i.e., those that were not identified as being followed by a perceptible pause) were counted as 0-length. For this measure, data were excluded for fillers which were followed by the clinician's speech, as the length of the pause was not under the child's control.

MLU

Mean length of utterance (MLU) was computed for each child as the average length of their C-units in words. When calculating utterance length, words before and after an intra-unit pause were counted as part of the same C-unit.

6.2.5 Analysis

For this research, we are interested in how different factors may influence a child's production of fillers. Thus, in addition to tests of central tendency (e.g., Wilcoxon), we also used linear mixed-effects regression models (a.k.a. mixed-models). Specifically we used the *lmer* function in the R package lme4 [14], to estimate models of the children's ratio of 'um' or 'uh'.

Mixed-models, such as *lmer*, are particularly well suited to these analyses as they; 1) are robust for unbalanced data, 2) include the ability to model individual differences as

random effects, and 3) allows both the factors of interest (e.g., diagnosis and activity) and potential confounds (e.g., sex and age) to be modeled as fixed effects.

As we do not expect the estimated models to be the same for the three different groups of children, especially between the children with TD versus those with ASD, we first create separate models for each group of children. In these models, we include activity, the child's MLU, age and sex as factors (i.e., fixed-effects). To avoid potential covariance effects between continuous factors, both age and MLU are first centered. To account for individual differences, Subject is modeled as a random-effects term. Because the values we wish to predict are ratios, the models are based on a binomial distribution. As such, the resulting coefficients are log values.

Referring ahead to Table 6.2, we now look at an example prediction based on the model for the children with TD. The first factor 'Intercept' represents the predicted value using the reference levels (i.e., activity=conversation, sex=M, mean MLU and mean age). The remainder of the Factors specify how to alter the prediction for non-reference values. At the bottom of the table is the standard deviation (SD) of the random effects (i.e., the Intercept offset that accounts for each subject's individual differences). For example, to determine the predicted ratio of 'um's for a male child with TD, of average age, who is engaged in *conversation* and has an estimated Intercept offset of 0.5, we would calculate exp(-4.43 + 0.5) = 0.02. For the same child engaged in *play*, the expected ratio of 'um's is exp(-4.43 + 0.5 - 0.39) = 0.013.

For this research, we are interested in determining both; 1) how the factors impact the ratios for that group of children, and 2) how the three groups of children differ. Thus we follow the creation of by-group models (e.g., models using only the children with TD) with a model that includes the data from all three groups of children and adds diagnosis as a factor. In doing so, we can ascertain to what degree diagnosis, and other factors interacting with diagnosis, impact the ratios.

To assess the quality of the models, we compute correlations between the actual ratios, and those predicted by the models. The r values for these correlations are reported along with the model results. It is important to note that our goal here is not to create highly predictive models. Instead, our goal is to determine if, and to what degree, the factors

included contribute to the estimated models. Thus, the reported r values are used to compare the models, and more specifically, to assess the ability of the included factors to produce a predictive model.

As an adjunct to the correlations, we also create a base model using only the random effect of subject. This model is essentially a combined model that includes no fixed-effect factors. By correlating the predictions of this base model to the actual values, we can assess how much of the variability in the data is accounted for solely by taking into account the subject's individual differences. Also, by comparing the base model's r value to that of the combined model, we can determine how much additional variance is accounted for in the combined model.

6.3 Results

There were 40,016 utterances consisting of 187,858 words available for analysis. Children with TD averaged 471 utterances, children with DLD averaged 442, and children with ASD averaged 432 utterances. All children produced both fillers, with the exception of two children (one TD and one DLD) who produced only 'uh's (no 'um's) and one TD child who produced only 'um's. In total, the children produced 3148 fillers, 2192 'um's and 956 'uh's.

6.3.1 Ratio of Fillers

Each group's median ratio of 'um's and 'uh's, both overall and by position, are shown in Table 6.1.[3] We look at these contexts individually as fillers have been posited to serve different roles, such as turn-taking, stalling for time or as part of a disfluency, and these roles are related to their position in a turn. Also shown are the number of children who produced an 'um' or 'uh' in that position. Looking at the "overall" row in the table, we see that a total of 89 of the 91 children produced 'um's and 89 out of 91 produced 'uh's.

Comparing within-subject using Wilcoxon signed rank tests, we find that the overall

[3]We excluded analyses of utterance-initial fillers, as fillers were quite sparse in this position, with only 46 out of 91 children producing either 'um's or 'uh's.

ratio of 'um's is significantly higher than the ratio of 'uh's for children with TD (V=348, n=27, p<.0001, one-tailed) and DLD (V=162, n=19, p<.01, one-tailed) and that these results are robust across position. For the children with ASD, the ratios of 'um' and 'uh' are not significantly different in any position, all p's>.05, NS.

	um						uh					
	TD		DLD		ASD		TD		DLD		ASD	
	ratio	n/N	ratio	n/N	ratio	n/N	ratio	n/N	ratio	n/N	ratio	n/N
overall	1.2%	26/27	1.0%	20/21	0.3%	43/43	0.3%	26/27	0.3%	20/21	0.6%	43/43
turn-initial	3.8%	25/27	2.2%	20/21	0.9%	37/43	0.8%	20/27	0.7%	19/21	1.2%	39/43
utt-medial	0.7%	25/27	0.4%	20/21	0.1%	37/43	0.2%	22/27	0.1%	18/21	0.4%	39/43

Table 6.1: Medians of the children's ratio of fillers in each position and the number of children with fillers in that category.

6.3.2 Ratio of 'um's

We next compare the group's ratio of 'um' to total words, as shown in Table 6.1. Wilcoxon rank sum tests reveal significant differences in the ratio of 'um's between the TD and ASD groups, (W=1251, nTD=27 nASD=43, p<.001), which is in line with our previous work. The DLD group's median falls between the TD and ASD groups, and there are no significant differences between the TD and DLD groups or the DLD and ASD group, all p's>.5, NS.

By-group models: As a first step, before creating the by-group mixed-models, we perform visual analyses, looking at how subjects' ratios were impacted by different predictors. Figure 6.1 shows a scatter-plot of the children's ratio of 'um's to total words (y-axis), as a function of age (x-axis), plotting separately by diagnosis and activity. The line in each chart is a loess curve (i.e., a regression using data in the 'local neighborhood', rather than all the data, to compute the line). From the chart, it appears that the children's activity impacts the ratios of 'um's with both *description* and *play* having lowered ratios. Visual analysis also suggests that age may effect the ratio of 'um's for children with TD and ASD, with the effect reversed for the two groups. A similar plot, using MLU instead of age, produced no visual evidence of an effect of MLU.

We now look at the by-group models. Table 6.2 shows the coefficients for each factor

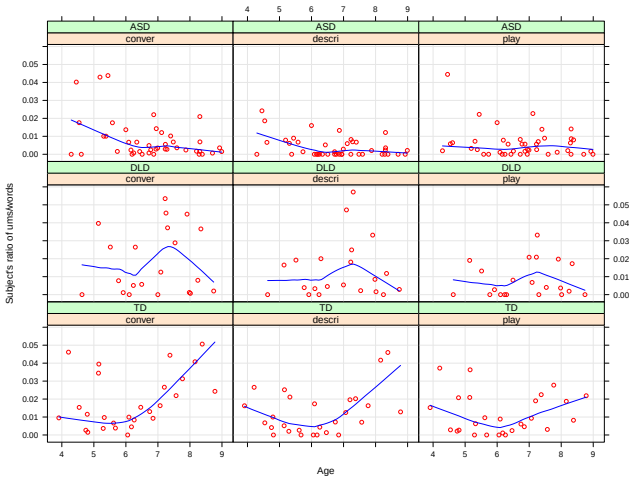


Figure 6.1: Children's ratio of 'um's to words plotted across age. The plotted lines show a loess curve fitted to each scatter-plot.

and their significance. These models support the impression that there is an effect of activity, with all models showing a significant negative effect for *description* (TD=-0.44, DLD=-0.52, and ASD=-0.70) and *play* (TD=-0.39, DLD=-0.82, and ASD=-0.82). There was a significant negative effect of age (-0.40) only for the ASD model. Although sex has been previously shown to be correlated to filler usage in adults [87], it was not found to be a significant contributor in any by-group model using this data. The children's MLU also showed no significant effect in any model.

Factors	TD		DLD		ASD	
Intercept	-4.43	***	-4.53	***	-5.42	***
age (centered)	0.26		0.19		-0.40	***
sex=F	0.24		-0.20		-0.01	
MLU (centered)	0.00		-0.31		0.17	
activity=*desc*	-0.44	***	-0.52	***	-0.70	***
activity=*play*	-0.39	***	-0.82	***	-0.82	***
SD (subj)	0.94		1.30		0.98	
r	0.91		0.85		0.89	

'***'<0.001, '**'<0.01, '*'<0.05, '.'<0.1

Table 6.2: Coefficients(log) for the 'um' estimated models. 'Intercept' represents the predicted value using the reference levels for each factor (i.e., activity= conversation, sex=M, mean MLU, and mean age).

Combined model: Next, to compare the effects of diagnosis, we create a combined model for 'um's that includes all three groups of children, as shown in the 'um' labeled column in Table 6.3. This model includes only those factors found significant in the by-group models. Additionally, looking at the by-group models in Table 6.2, we see that the coefficients for activity and age differed between the three groups of children, thus we include interactions between diagnosis (dx) and activity, and dx and age, in the combined model. Because the by-group models, and preliminary combined models, did not show significant effects of MLU or sex, these two factors were not included in the final combined model.

Looking at the combined model in Table 6.3, we see a significant effect of diagnosis for the children with ASD (dx=ASD), but not for DLD (dx=DLD). This means that the children with ASD differ significantly from those with TD, but the children with DLD do not. In addition, in the combined model, age has a marginal positive effect for the children with TD, and a significant negative effect for children with ASD, and no interaction for the children with DLD. This suggests that for children of this age, the ratio of 'um's in their speech increases as they age, but that for children with ASD, the ratio of 'um's declines. We also see that the children with DLD have a significantly lower ratio of 'um's when engaged in play.

Analysis of model fit: Next, we assess how well the models predict the ratio of 'um's, by examining the r values. In essence, these values address whether, and how well, the factors included in the model are capable of accounting for the data. All three by-group models, as shown in Table 6.2, had r's > 0.85, with the DLD model being the least precise. Thus the models account for between 73% and 83% of the variance in the data. However, for mixed-effect models, much of the variance is often accounted for by individual differences, as reflected in the r of 0.83 for the base model, as shown in Table 6.3. Thus, the factors included in the combined model account for 13% of the variance accounted for by the combined model.

Factors	um			uh		
	Combined		Base	Combined		Base
Intercept	-4.25	***	-5.13	-5.54	***	-5.56
age (centered)	0.08	.				
sex=F						
MLU (centered)				-0.35	***	
activity=*desc*	-0.44	***		-0.43	*	
activity=*play*	-0.39	***		0.08		
dx=DLD	-0.33			-0.27		
DLD:age	-0.15					
DLD:desc	-0.08			-0.09		
DLD:play	0.43	**		0.08		
dx=ASD	-1.11	***		0.44	.	
ASD:age	-0.64	**				
ASD:desc	-0.26			-0.27		
ASD:play	0.11			-0.57	**	
SD (subj)	1.07		1.25	0.78		0.88
r	0.89		0.83	0.79		0.71

'***'<0.001, '**'<0.01, '*'<0.05, '.'<0.1

Table 6.3: Coefficients(log) for the 'um' and 'uh' estimated models. 'Intercept' represents the predicted value using the reference levels for each factor (i.e., activity=conversation, sex=M, mean MLU, mean age, and dx=TD).

6.3.3 Ratio of 'uh's

We now compare the group's ratio of 'uh' to total words, referring back to Table 6.1. Here that the children with ASD have significantly higher ratios than the other two groups by Wilcoxon rank sum tests; TD versus ASD, (W=1695, nTD=27 nASD=43, p<.03), and DLD versus ASD (W=1534, nDLD=21 nASD=43, p<.03). No significant difference was found between the TD and DLD groups, p>.5, NS.

By-group models: Figure 6.2 shows the ratio of 'uh's as a function of the children's MLUs. Visual inspection suggests that there may be some effect of MLU and activity, and that the ratios are generally low. A similar plot, using age instead of MLU, produced no visual evidence of an effect of age.

For 'uh's we used a procedure similar to that used for 'um's. First, we create by-group models, as shown in Table 6.4. For these models, the activity *description* is the only factor to show a significant effect in all three models, and was the only significant factor

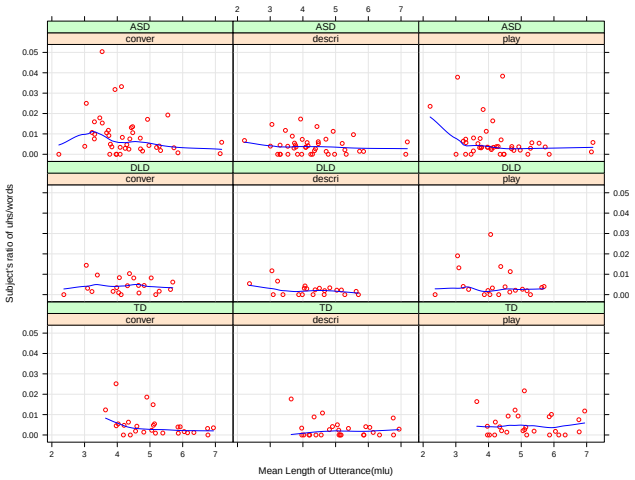


Figure 6.2: Children's ratio of 'uh's to words plotted across MLU. The plotted lines show a lowess curve fitted to each scatter-plot.

for children with DLD. For the children with TD and ASD, MLU was also a significant factor. The activity *play* was a significant contributor to the models only for the children with ASD.

Factors	TD		DLD		ASD	
Intercept	-5.81	***	-5.55	***	-4.97	***
age (centered)	0.20		-0.01		0.05	
sex=F	0.11		-0.33		-0.35	
MLU (centered)	-0.47	*	-0.26		-0.44	**
activity=*desc*	-0.43	*	-0.52	*	-0.70	***
activity=*play*	0.88		0.17		-0.49	***
SD (subj)	0.79		0.78		0.76	
r	0.85		0.62		0.78	

'***'<0.001, '**'<0.01, '*'<0.05, '.'<0.1

Table 6.4: Coefficients(log) for each of the 'uh' estimated models. 'Intercept' represents the predicted value using the reference levels for each factor (i.e., activity=conversation, sex=M, mean MLU, and mean age.)

Combined model: We next create a combined model, including all three groups of children. For 'uh's, neither sex nor age showed a significant effect in the by-groups models, thus are excluded from the combined model. An interaction between diagnosis

and activity is included, but interactions between MLU and diagnosis were not found to be significant in preliminary combined models, thus are also excluded. The resulting model is shown in the 'uh' labeled column in Table 6.3.

Looking at the combined model, we see that the *description* activity and MLU have a significant effect. Interestingly, the coefficient for ASD is positive, suggesting that children with ASD may have a *higher* ratio of 'uh's than children with TD or DLD, although this factor was not significant. However, when engaged in *play*, the children with ASD have a significantly lower ratio of 'uh's than children with TD.

Analysis of model fit: It is important to note that for 'uh's, all models are less predictive than those for 'um's with r's ranging from 0.62 for DLD, to 0.85 for TD, as shown in Table 6.4. Thus, these models, or more specifically the factors used, are less able to predict the ratios of 'uh's. The same is true for the 'uh' combined model, with an r of 0.79, versus 0.89 for the 'um' combined model, as shown in Table 6.3.

6.3.4 Filler Ratio by Position

We next examine the ratio of turn-initial and utterance-medial fillers. For these analyses, we recreate the process used above for 'um's and 'uh's, but here present only the combined models in Table 6.5.

Um, by position: Looking first at the 'um's in Table 6.5, we see that for both turn-initial and utterance-medial positions, the same factors are significant as those found in the previous overall combined model, and that the models both have substantial r's. This finding suggests that the factors the predict how often a child will use 'um's are robust, regardless of position in the utterance.

Uh, by position: Looking next at the 'uh's in Table 6.5, we see that the models are less consistent. In the turn-initial position, only *description* has a significant effect, with MLU and a diagnosis of ASD having marginal effects. In utterance-medial, we find that the only factors with a significant effect are the activities *play* and *description*, with MLU and diagnosis having no significant effect. However, this model should be given less weight, as it showed little ability to predict the actual values with an r of only 0.66.

	um				uh			
Factors	turn-initial		utterance medial		turn-initial		utterance medial*	
Intercept	-3.14	***	4-.73	***	-4.39	***	-6.12	***
age (centered)	0.34	.	0.30	.				
MLU (centered)					0.30	.		
activity=*desc*	-0.67	***	-0.32	**	-0.72	*	-0.48	**
activity=*play*	-0.53	***	-0.32	**	0.19		-0.42	**
dx=ASD	-1.10	**	-1.42	***	0.29			
dx=DLD	-0.29		-0.63	.	-0.41			
ASD:age	-0.11	**	-0.56	*				
DLD:age	-0.23		-0.16					
ASD:desc	-0.11		-0.51	.	0.00			
DLD:desc	0.25		-0.31		0.04			
ASD:play	0.24		-0.07		-0.51	.		
DLD:play	-0.59	*	-0.24		0.38			
SD (subj)	1.19		1.14		0.94		0.89	
r	0.86		0.87		0.81		0.66	

'***'<0.001, '**'<0.01, '*'<0.05, '.'<0.1

Table 6.5: Coefficients(log) for each of the 'um' and 'uh' by-position combined models. 'Intercept' represents the predicted value using the reference levels for each factor (i.e., dx=TD, activity=conversation, sex=M, mean MLU, and mean age). **The utterance-medial model for 'uh's contains only activity, as more complex models using age and MLU were not justified via likelihood ratio test* [7].

6.3.5 Pauses after Fillers

Next we examined the ratio of fillers that were followed by pauses. For these analyses a total of 2936 fillers were available, of which 881 were followed by non-zero length pauses.

Table 6.6 shows the ratio of fillers followed by pauses for each group, as well as the number of children in that group who produced any pauses after that filler. Comparing within-subject, we find that the children with TD have a significantly higher ratio of pauses after 'um' than after 'uh', a significant different by Wilcoxon signed rank test, V=266, N=25, p<.01, one-tailed. In contrast, neither the DLD or ASD groups significantly differed in their ratios for 'um's versus 'uh's, both z's<0.5, NS.

Comparing the incidence of pauses between the groups, we see that only 28 out of 42 (67%) children with ASD ever paused after 'um', whereas all but one of the children with TD and one of the children with DLD did, a significant difference by Fisher exact

	TD		DLD		ASD	
	ratio	(n/N)	ratio	(n/N)	ratio	(n/N)
um	48.35%	(24/26)	42.45%	(19/20)	34.11%	(28/42)
uh	28.43%	(17/26)	40.73%	(16/21)	37.46%	(36/42)

Table 6.6: Ratio of fillers followed by pauses and the proportion of children (n/N) who produced pauses.

test, p<.01, one-tailed. However, for 'uh's no significant difference was found for the three groups of children, p>.1, NS.

We next compare the length of pauses following fillers, including 0-length pauses, as shown in Table 6.7. All three groups averaged longer pauses after 'um' versus 'uh', but this difference was significant only for the children with TD by Wilcoxon signed rank test, V=265, N=25, p<.01, one-tailed, not for the children with DLD or ASD, both p's>0.3, NS.

	including 0-length			excluding 0-length		
	TD	DLD	ASD	TD	DLD	ASD
(N=)	(25)	(20)	(40)	(16)	(15)	(24)
um	0.63	0.43	0.46	1.43	0.93	1.34
uh	0.37	0.39	0.40	1.09	1.08	0.95

Table 6.7: Mean length of pauses (in seconds) following fillers, both including and excluding 0-length pauses.

We also looked at only those fillers that were followed by a non-zero length pause, as shown in Table 6.7, columns 3 and 4. Data is included only for those children who produced pauses after both 'um' and 'uh'. Comparing within-subject we find that the TD and ASD groups have significantly longer pauses after 'um' than after 'uh' by Wilcoxon signed rank test, one-tailed: TD V=101, N=25, p<.05; ASD V=215 N=24, p<.04. The DLD group showed no significant difference, z<1, NS. This is particularly interesting in that the DLD group had the highest percentage of children who produced pauses after both 'um' and 'uh' (70%), as compared to 59% for the TD group and 56% for the ASD group.

6.4 Discussion

The focus of this chapter is to better understand how fillers are used and whether their use can be traced to social or processing issues. To do this, we compared the use of fillers between children with TD, DLD and ASD. As determining the source of a speech phenomena can be challenging, we include the latter two groups as they each have impairments that impact their speech, but only the children with ASD have, by definition, impaired pragmatics.

Comparing the use of fillers by the children with DLD to that of the other two groups of children, the results were as expected. The DLD group had a significantly higher ratio of 'um's than 'uh's, like that of the TD group, and unlike the ASD group, whose ratio of 'um's and 'uh's did not differ significantly. These results suggest, as posited in our previous work, that 'um's and 'uh's result from different cognitive processes: 'uh' from an internally focused process, perhaps similar to self-directed speech, and 'um' from an externally focused process, in which the speaker intentionally uses 'um' to inform, or assist, the listener.

By creating mixed-effects models of the children's ratio of 'um's and 'uh's, we find further evidence to support this hypothesis. When modeling ratios of 'um's we find that, regardless of diagnosis, the children's activity is a significant predictor. In addition, as expected, ratios for the children with TD and DLD do not significantly differ, except during the *play* activity, in which the children with DLD had higher ratios of 'um's. Of particular interest is that, for the children with ASD, age is also a significant negative factor, with older children being less likely to produce 'um's, whereas, for the children with TD and DLD in this age group, the ratio did not change with age.[4] One possible explanation is that the children with TD and DLD may already have learned to use 'um's and 'uh's much as they will as adults.

The factors found to be significant for children's overall ratios of 'um' were robust for both turn-initial and utterance-medial positions. This result suggests that not only are

[4]In fact, older children with TD and DLD may be more likely to produce 'um's, although this result was only marginal, and appeared only in the combined model, thus must be suspect.

the factors modeled predictive of children's ratios of 'um's, but that the production of 'um's is strongly related to a child's activity and diagnosis, and potentially less related to more localized aspects of the child's utterance, such as position in the utterance.

Models for the children's ratios of 'uh's were dissimilar to those for 'um's. For these models, only the activity *description* was found to be a significant predictor in all by-group models, and was the only significant predictor for the children with DLD. However, MLU was a significant predictor for both the TD and ASD groups, and the activity *play* was a significant predictor for the children with ASD. In the combined model, we found the diagnosis was *not* a significant predictor, but both MLU and *description* are significant predictors. These results suggest that 'uh's are not listener-oriented, but are more related to other factors, such as the nature of the activity, and the child's propensity to produce longer utterances.

The results for ratios of 'uh's were somewhat replicated in the turn-initial model (i.e., *description* is significant and MLU is marginal). However, for the utterance-medial model, MLU was not significant, but both *description* and *play* were significant. In addition, we found that the predictors used were less able to predict the ratios of 'uh's. Once again, this suggests that 'uh's might be better modeled by localized factors, ones that were not included in our models (e.g., did the 'uh' precede a new versus old referent [48]).

Looking at pauses after fillers, we find that both children with TD and DLD were more likely to produce a pause after an 'um' than children with ASD, but that the three groups were equally likely to produce a pause after 'uh'. This is perhaps the most important of the pause results, in that it suggests that children with TD and DLD are learning to use 'um', although only the children with TD showed differential pausing patterns such as those seen in adults [24]. It is likely that the children with DLD, given their age inappropriate language skills, had post-filler pause lengths that more closely resemble those of younger children, in which pauses lengths did not differ between 'um' and uh' [43].

These finding are immediately relevant to several purposes. First, by illustrating the similarities and differences in use of fillers between children with TD and those with ASD and DLD, we gain a greater understanding of the mechanisms involved in dialogue production and processing, with particular insights into what aspects of dialogue may be more

social in nature. Second, this work provides additional insight into dialogue processing in ASD, leading toward a finer understanding of which dialogue skills are affected by ASD. Third, the unique pattern of filler use by children with ASD can be used to assist in diagnosis.

6.4.1 Lessons for HCI

From an HCI perspective, this work provides an improved understanding of how fillers are used during human communication, and how they might be used by SDSs to improve human-computer interaction. Specifically, it suggests that 'um' is listener-oriented and is used to inform a listener of a pending delay, or signal a speaker's uncertainty. In contrast, 'uh' appears to be speaker-oriented, and is likely an artifact of speech processing problems. Thus, SDSs can leverage speakers' use of fillers, anticipating that speakers' use of 'um' provides insight into their communicative intent and 'uh' into their cognitive load or processing state. In addition, SDSs could produce fillers, using 'um's to signal a delay or ambivalence about taking the turn and 'uh' to indicate a new or novel referent, without risking user confusion.

6.4.2 Implications for ASD and DLD

Although not the main goal of this dissertation, these findings have implications related to ASD and DLD. First, the unique pattern of filler use by children with ASD can be used to assist in differential diagnosis of ASD and DLD, which has, to date, proven difficult. Second, this work provides additional insight into dialogue processing in ASD, leading toward a finer understanding of which dialogue skills are affected by ASD. Third, by illustrating the similarities and differences in use of fillers between children with TD and DLD, and those with ASD, we gain a greater understanding of the mechanisms involved in dialogue processing, with particular insights into what aspects of dialogue may be more social in nature. This could be of particular interest in remediation of social-language deficits for children with ASD.

Chapter 7

Turn-taking Gaps, and Interactions with Questions and Disfluencies

In this chapter, we focus on turn-taking. Turn-taking is an important component of dialogue coordination, as it specifies how conversants take turns having the speaking floor. Past research has focused on how quickly one conversant starts speaking after another stops. From an HCI perspective, an understanding of what factors impact how long a user might take to respond would be beneficial to ensuring that an SDS does not engage in further prompting when the user simply needs additional time to respond. Also, knowing what factors increase a user's likelihood of initiating their utterance with less easily recognized speech (e.g., false starts, repeats, or fillers), could help designers build systems that manage the context so as to avoid these types of disfluencies.

By asking a question, a speaker places an obligation to respond on the listener [98], and this impacts how the listener responds. Both inter-turn gaps and disfluencies have been shown to be impacted by questions: when presented with a question, speakers respond more quickly [38] and are more likely to become disfluent [53]. However, it remains unclear whether the occurrence of disfluencies and the length of gaps are related.

As in the previous chapter, here we include children with TD, DLD and ASD to provide insight into the extent turn-taking is impacted by social pressure. If so, we would anticipate that the children with DLD would manage turn-taking in a manner similar to those with TD. In contrast, if the DLD group more closely resembles the ASD group, it is likely that turn-taking is more strongly impacted by processing (dis)abilities.

In this chapter we explore:

- Whether gap length is impacted by social pressure, specifically that induced by a question versus non-question, or by processing deficits.
- How responsive the different groups are to questions. We anticipate that the ASD and DLD groups will be less likely to respond to questions, due to their respective social and processing issues.
- Whether speakers are more likely to produce turn-initial disfluencies after being asked a question, and if this likelihood is impacted by the length of the preceding gap.
- Whether diagnosis (i.e., TD, DLD, and ASD) and questions influence the type of disfluency (i.e., 'um', 'uh' or other) speakers produce.

The work on gaps in this chapter subsumes that published in *"Autism and Interactional Aspects of Dialogue"* [38], on which I was the second author.[1]

7.1 Background and Related Work

7.1.1 Gap Lengths

SDSs typically predefine some "time-out" value, after which, if the user has not responded, it is assumed that a user needs re-prompting or assistance [49]. On the surface, this appears a reasonable assumption, in that people are obligated to respond to questions and strive to minimize gaps [82]. In addition, inter-turn gap length is consistent across languages and cultures, with responses generally forthcoming within 0.5 seconds [91]. Gap lengths also tend to be shorter when no visual cues of attention are present (i.e., during a telephone conversation) [96], a common context for SDSs.

[1] My contribution to that paper, in terms of gap length analyses, was to use subject means (rather than raw, unbalanced data) in the group-wise analyses, and to perform within-subject comparisons. The work in this chapter extends that work by including children with DLD, using forced alignment to refine the gap measurements, and performing more sophisticated statistical analyses. Specifically, here we use non-parametric tests of central tendency (in lieu of t-tests), conduct group-wise comparisons prior to pair-wise comparisons, and adjust p-values (e.g., bonferroni) when performing comparisons on subdivided data for which there was no significant omnibus test.

Yet, the length of gaps can vary. For example, gaps are longer when the responding utterance is longer [11], when the responder needs clarification [85], and when asked a difficult question [88]. Thus, an SDS with a predefined time-out could produce additional prompting, interrupting the user's train of thought. Instead, by better understanding when a user may take longer to respond, SDSs could tailor time-outs based on the current context.

7.1.2 Questions and Turn-taking

During a dialogue, interlocutors provide signals indicating their intent to take, keep, release, or assign the speaking turn [30, 35]. After a question, which assigns the turn to the interlocutor, Sacks et al. [82], suggest that the respondent "...has the right and is obligated to take the next turn to speak...". It has long been recognized that questions engender a social obligation to respond [98, 23, 36]. That is, if a speaker has ensured contact, perception, and understanding, then the listener is obligated to respond in some way, either answering the query, requesting clarification, or explaining why the query cannot be answered.

In work examining the way speakers respond to factual questions, Smith and Clark [88] found that when speakers are uncertain whether they know the answer, they will respond more slowly, give a non-answer more quickly, and add the fillers "um" and "uh". They suggest that a desire to preserve self-presentation prompts responders to avoid excessive delays, which might cause the questioner to view them as uncooperative or slow-witted. Instead the responder will provide signals to account for delays and indicate their level of confidence in their answer. Thus, it appears that the obligation to respond places both social and cognitive pressure on the listener.

7.1.3 Disfluencies and Turn-taking

Lickley [53] found that the highest rate of disfluent words are found in replies to wh-questions (e.g., "Who", "What"), negative replies, and instructions and clarifications. Of these, the two reply categories had the highest percentage of fillers. This work did not separate turn-initial from utterance-medial disfluencies, thus it is unclear whether

responding to a question results in generally disfluent speech, or if the effect is localized to the beginning of the utterance. However, it seems likely that these disfluencies would be more prevalent in the turn-initial position because, due to the social pressure to respond in a timely manner, a listener could start to respond before they have fully planned their response.

Although not looking at turn-taking per se, Bard et al. [12] examined how the rate of disfluencies at a turn transition are related to measures of difficulty, finding that the occurrence of disfluencies (excluding fillers) are primarily related to utterance-specific factors such as utterance length, type of referring expression, and whether the speaker is giving or receiving instruction. They conclude that a speaker's disfluencies are less related to interpersonal (e.g., familiarity of the interlocutors) and comprehension factors (e.g., complexity of the preceding utterance), and instead are more closely linked to the speech production process. However, this work did not look at whether a speaker was responding to a question and did not include fillers.

7.1.4 Use of ASD and DLD Children

Dialogue needs to be orderly to proceed smoothly, and the gaps (often termed inter-turn pauses or lapses) between turns tends to be minimal [82]. This is possible because dialogue participants provide auditory and visual cues that signal their intention to either release, keep, or take the turn [30, 35]. By recognizing these cues, speakers are able to smoothly control turns and minimize gaps within a dialogue.

In children with TD, the ability to recognize, and respond, to these turn-release cues is learned early in childhood, starting in infancy [81]. As such, we can expect that children with TD have a good sense of the rules, and can recognize the cues that indicate a speaker intends to release the turn. However, as the cues of impending turn-release tend to be paralinguistic and social in nature (e.g., prosody and gaze), it might be difficult for children with ASD to recognize them.

When discussing the ability to respond in a timely manner, it is important to note that speech timing is not regulated only by the ability to recognize when the turn is going to be released. In addition, a speaker must also have the abilities to, first, understand the

speaker's intent, and second, to quickly organize and formulate a response. These abilities are exactly those that are impaired in children with DLD, and so might also be expected to impact these children's ability to control gaps.

When responding to a question, speakers are more likely to produce disfluent speech [53]. Taken in context with the shorter gaps after questions (versus non-questions) found in children with TD [38], it seems likely that speakers may be starting to speak before fully ready. However, to the best of our knowledge, no previous work has explored whether there is an interaction between gaps and disfluencies.

7.1.5 Disfluencies and ASD

In recent work examining disfluencies in men with ASD, Lake et al. [50] found that, after being asked a question, they tended to produce more repeats and longer gaps than controls, but fewer fillers and repairs. These findings are in keeping with our previous work examining the incidence of fillers and gap lengths for children with ASD [38], but did not separate 'um' from 'uh', and examined only responses to questions. However this work does suggest that repairs, much like 'um', may be listener-oriented and social in nature.

7.2 Methods

7.2.1 Research Design

For gap lengths, the experimental design involved within-subject comparisons as a function of: (1) Question (non-question, question, consecutive questions) and (2) Activity (conversation, description, or play). An additional between-group factor was: (3) Diagnosis (TD, DLD, or ASD).

As we also wish to determine what factors might predict whether a speaker will produce a disfluency, and whether gaps and disfluencies interact, for mazes (described below) the experimental design was correlational. Here we created models to estimate the likelihood of a turn-initial maze as a function of (1) Activity, (2) Question, (3) Length (in words), (4) Gap Length, (5) Age, and (6) Sex.

7.2.2 Data and Coding

The data used in this research is described in Chapter 6. In addition to the annotations described therein, overlaps between the child's and clinician's speech was annotated by placing the overlapped regions of the transcript within angle brackets (i.e., '<>'). Speech identified as mazes (described below), was placed within parenthesis (i.e., '( )'). Here, we use data only for those turn-exchanges from clinician to child in which there is no overlap annotated at the beginning of the child's C-unit, or at the end of the clinician's C-unit.[2] Data were also excluded if the clinician's speech was annotated as incomplete.

Gaps

For gap lengths, the manual annotations were not always accurate. Hence, we used an automatic speech recognizer [93] to refine the silence durations. This was accomplished via a forced alignment between the text transcriptions for the clinician's speech proceeding the gap, and the child's speech following the gap. The resulting end and start times were then used to compute an unbiased measurement of the gap lengths. Gap lengths, in milliseconds, were then log-transformed to produce a more normal distribution and a more representative measure of central tendency. For this data, log-transforming the gaps is an acceptable practice, as we are not including overlaps in these analyses [39]. Each subject's far outliers (i.e., those outside 1.5 * inter-quartile-range) were removed. Far outliers accounted for 2.4% of the data.

Mazes

As defined by SALT [83], mazes include false starts, repetitions of a word or phrase, revisions, fillers, and stutters. Annotation guidelines for this data require that "Whatever is not within parentheses should form a complete utterance whenever possible." The annotation of mazes in our data was not second-scored, and pilot analyses suggested that

[2]Overlaps were excluded because the clinician's and child's speech was recorded on a single channel, making accurate manual annotation of the beginning and end of each speaker's speech difficult. In addition, the single channel format was not amenable to forced alignment, as we could not separate the two speech signals.

different annotators tended to be more, or less, aggressive when identifying what words are part of a maze versus the final contentful speech. This was not found to be the case for mazes occurring at the first word of a C-unit, which is the measure used here. Each C-unit was annotated as starting with, or without, a maze. For those C-units that started with a maze, the C-unit was annotated as to whether the initial word was an 'um', an 'uh', or a non-filler word.

Length

As utterance length (in terms of the number of words) has been shown to impact a subject's rate of disfluencies [72], we include the length of each C-unit as a potential predictor of turn-initial mazes. All words are included in the length measure. This was because of potential inaccuracies in the annotation of mazes, as noted above. Length was log-transformed prior to analyses.

Questions

For each turn-initial C-unit produced by a child, it was determined whether the preceding clinician C-unit was a non-question (annotated with a final '.' or '!') or question (annotated with a '?'). C-units in which the clinician utterance was incomplete were discarded. In addition, we determined whether each clinician question was a consecutive question, i.e., was immediately preceded by a clinician question with no intervening child speech.

7.2.3 Analysis

For the between-group comparisons (e.g., TD vs DLD vs ASD), we first perform Kruskal-Wallis tests, a non-parametric alternative to ANOVA, to determine if there exist significant differences between any two groups. These analyses are followed by post-hoc pairwise comparisons (e.g., TD vs DLD) of the groups using Wilcoxon rank sum tests with bonferroni correction.

For the within-subject comparisons we conducted paired statistical tests, specifically using Friedman rank sum tests to ascertain group-wise differences (e.g., for comparing between activities), and post-hoc Wilcoxon signed rank tests with bonferroni correction.

To explore what factors influence the likelihood of a child producing an utterance-initial maze, we created mixed-effect logistic regression models for each group of children, once again using the *lmer* function in the R package lme4 [14]. These models use a logit (i.e., log-odds) link function, thus the models predict the log-odds of a given binary outcome (i.e., whether or not a maze was produced). The models were iteratively refined, removing factors that did not contribute either alone or by interaction with other factors. The *lmer* function also supplies the model's predicted probability (i.e., inverse logit) for each datum. As a correlation between the binary responses and the predicted probabilities would not produce usable results, the reported r values were computed thusly: first, the predicted probabilities were sectioned into 10 bins of size 0.1 (e.g., 0-0.1, 0.1-0.2, etc.); second, the mean of the predicted probabilities was computed for each bin; third, the mean for the actual responses (i.e., 0=no maze, 1=maze) was computed for each bin; and fourth, the mean probabilities are correlated to the mean actuals. In this manner we can determine the goodness-of-fit for each model.

7.3 Results: Gap Lengths

Data for 19,624 gaps were available for analyses; 5529 for the TD group, 4767 for the DLD group, and 9328 for the ASD group.

7.3.1 Between-Group Comparisons

We first analyze the gap lengths, comparing between the groups. Figure 7.3.1 shows box-and-whisker plots representing the distribution of the children's gaps for each group. The boxes show the first through third quartiles (i.e., the inter-quartile range (IQR)), and each box's midline is that group's median. The whiskers extend to highest (above) and lowest(below) datum within 1.5 IQR. For the ASD group we see a circle below the lower whisker, indicating that one child's mean fell outside the 1.5 IQR range, and could be considered an outlier. From the plot, we can see that the children with TD are more consistent in their mean gap lengths (i.e., their IQR and whiskers are smaller), and that the ASD groups' median is higher than that of the other two groups. Comparing group-wise,

we found a significant effect of diagnosis on gap length by Kruskal-Wallis rank sum test, χ^2=6.4, df=2, p<.05. Next, comparing the groups pairwise, we did not find significant differences between the TD and DLD, or the DLD and ASD groups (both p's>0.5, NS) but did find a marginal difference between the TD and ASD groups, by Wilcoxon rank sum test (W=779, nTD=27 nASD=43, p<.052).

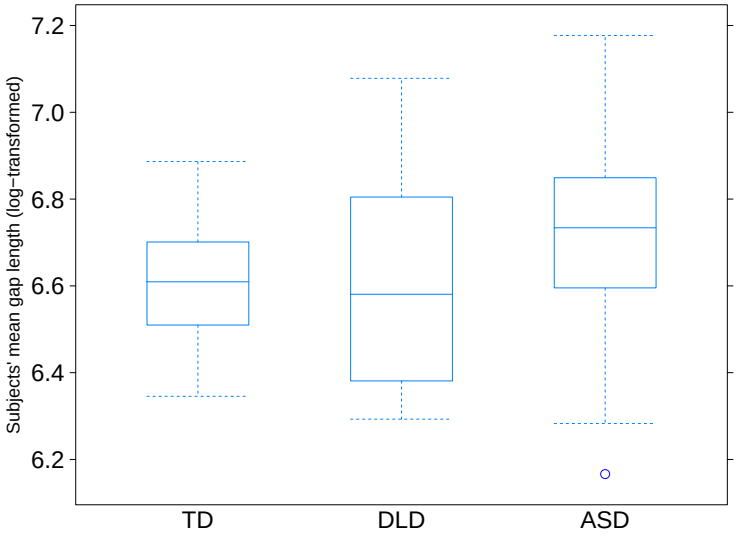


Figure 7.1: Subjects' mean gap lengths (log-transformed), displayed by group (TD, DLD, and ASD).

Gaps after Questions and Non-questions

As questions carry stronger cues of turn-release (e.g., rising intonation) and a stronger social obligation to respond, we also compare the groups gaps after the clinician issued a non-question and after the clinician asked a question.

Non-questions: As shown in Figure 7.3.1 (upper left), the groups differ little in their gaps after a non-question, with no significant group-wise difference by Kruskal-Wallis rank sum test, p>.3, NS. Shown in the remainder of Figure 7.3.1 (upper row), are analyses comparing between groups by activity. Kruskal-Wallis rank sum tests, with bonferroni correction, revealed no underlying significant difference between the groups, all p's>.08, NS.

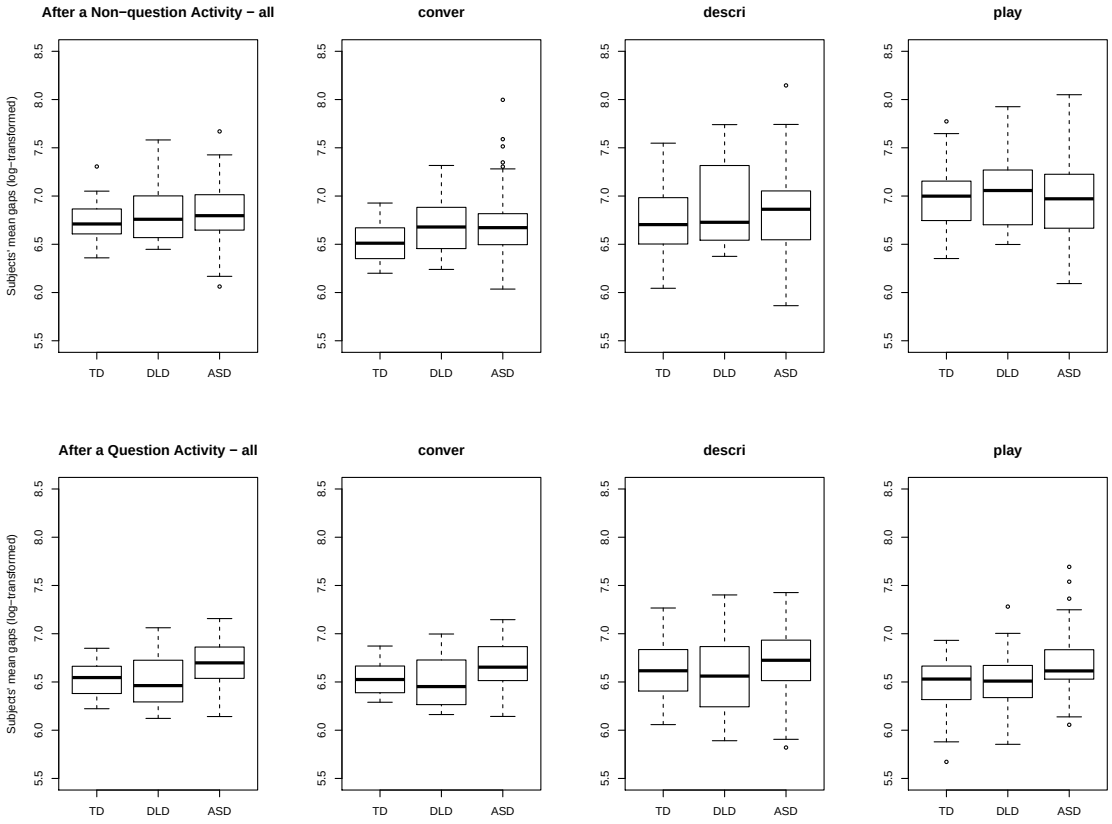


Figure 7.2: Subjects' mean gap lengths (log-transformed), both after a non-question (upper row) and after a question (lower row). Also broken down by activity.

Questions: In contrast to the non-questions, the three groups do differ in their gaps after a question as shown in Figure 7.3.1 (lower left). Here we found a significant effect of diagnosis on gap length by Kruskal-Wallis rank sum test, χ^2=10.48, df=2, p<.01. This effect is accounted for by significant differences between the group of children with ASD and the TD group (Wilcoxon rank sum test, W=821, nTD=27, nASD=43, p<.02), and a marginal difference between the ASD group and the DLD group (W=285, nASD=43, nDLD=21, p<.06).

Questions, by Activity: To further explore these differences, we next look at the gaps after questions broken down by activity, shown in Figure 7.3.1 (bottom row). For these comparisons, there are also significant group-wise differences, by Kruskal-Wallis rank sum

tests, when the children engage in *conversation* (χ^2=9.10, df=2, p<.02) and *play* (χ^2=6.44, df=2, p<.04), but not during *description*, p>.2, NS. The difference during *conversation* can be attributed to significant differences between the ASD and TD groups, p<.04 and marginal differences between the ASD and DLD groups, p<.06. The difference during *play* can be attributed to significant differences between the ASD and TD groups by Wilcoxon rank sum tests, W=781, nTD=27, nASD=43, p<.05.

Between-Group - Summary

The above results are summarized in Table 7.1. Here it it more readily apparent that the groups' gaps differed only after a question. In addition, it is clear that the ASD group differed significantly from the TD group, but that the DLD group did not.

	Activity	Group-wise	TD-DLD	DLD-ASD	ASD-TD
Non-question	All				
	Conversation				
	Description				
	Play				
Question	All	*		-	*
	Conversation	*		-	*
	Description				
	Play	*			*

Table 7.1: Between group comparison of gaps. '*' indicates a significant difference, and '-' a marginal one.

7.3.2 Within-Subject Comparisons

We now compare within-subject, looking first at how each group's gaps differ after a question versus after a non-question, as shown in Table 7.2 (the corresponding group medians are shown in Figure 7.3.1 (lower and upper left). All three groups had significantly shorter gaps after a question versus a non-question, by Wilcoxon signed rank tests (TD: V=341, p<.001, DLD: v=224, p<.001, ASD: v=736, p<.002).

	TD *	DLD *	ASD *
Non-question	6.73 (.21)	6.81 (.31)	6.82 (.32)
Question	6.53 (.16)	6.53 (.28)	6.69 (.24)

Table 7.2: Subjects' mean gap lengths (log-transformed) and standard deviation. $*$ indicates a significant within-subject difference ($p<.05$).

Questions vs Non-questions, by Activity

Next, we compare questions versus non-question within activity for each group, as shown in Table 7.3 (the corresponding group medians are shown in Figure 7.3.1). All three groups had significantly shorter gaps after questions versus non-questions during *play* by Wilcoxon signed rank test, all p's$<.0001$. Only the DLD group showed significant differences during the other activities, having significantly shorter gaps after a question during *conversation* ($p<.02$) and during *description* ($p<.005$).

	TD		DLD		ASD	
Activity	Non-question *	Question	Non-question *	Question	Non-question *	Question
Conversation	6.51 (.19)	6.53 (.17)	6.67 (.29)	6.50 (.26)	6.73 (.40)	6.68 (.25)
Description	6.75 (.35)	6.63 (.31)	6.86 (.41)	6.58 (.40)	6.83 (.44)	6.71 (.33)
Play	6.99 (.33)	6.47 (.30)	7.09 (.54)	6.55 (.35)	6.97 (.44)	6.70 (.34)

Table 7.3: Subjects' mean gap lengths (log-transformed) and standard deviation after non-questions and questions. * indicates a significant effect of activity for that group.

Questions: Comparing between Activities

We next look at gaps after questions, comparing between activities, also shown in Table 7.3 (the corresponding group medians are shown in Figure 7.3.1). We find that, for each group of children, there were no significant differences due to activity by Friedman rank sum test, all p's$>.1$, NS.

Non-questions: Comparing between Activities

Next, we explore the gaps after a non-question. All three groups showed a significant effect of activity on gap length by Friedman rank sum test, all p's$<.02$. Comparing pairwise for the activities, we found significant differences between *conversation* and *play* for all three groups, by Wilcoxon signed rank test, all p's $<.02$. However, only the TD group had

significant differences between any other activities, with gaps during *description* being significantly shorter than gaps during *play*, $p<.03$.

7.3.3 Gap Lengths - Summary

To summarize, we find that the three groups differ in their gaps, and that this difference is explained primarily by the ASD group's longer gaps after a question during *conversation*. In addition, we find no significant differences in gaps between the children with TD and those with DLD. Perhaps surprisingly, when comparing within-subject, we find that all three groups have significantly longer gaps after non-questions than after questions, and this is especially evident during *play* for all three groups. In addition, all three groups did not differ between activities after a question, but did differ between *conversation* and *play* after a non-question. In essence, all the children respond more quickly after questions, but the children with TD and DLD are more effective at shortening their gaps.

7.4 Responsiveness to Questions

For the above gap length analyses, we looked at the length of gaps whenever the child responded to the clinician. However, these results could be skewed if the children choose to postpone responding. Hence, we now take this into account.

First, we review the clinician's data, looking at how many C-units the clinician produced, how many C-units were questions, what ratio of their speech C-units consisted of questions, and how often a question was preceded by a question. This data is shown in Table 7.4. Starting at the top of the table and working down, we see that the clinician spoke more, and asked more questions, when interacting with the children with DLD and ASD than when interacting with the children with TD. In addition, the clinician had higher rates of questions and consecutive questions for the DLD and ASD groups, as compared to the TD group.

Second, we look at the children's responses to questions. Table 7.4 shows how many

times the children responded to a clinician question, and what ratio of the clinician questions were responded to by the children.[3] Here we see that the children with DLD and ASD answered more questions than the children with TD, but given the greater number of questions asked by the clinician, they responded to a lower ratio of questions.

Clinician	TD	DLD	ASD	K-W p	TD-DLD	TD-ASD	DLD-ASD
Units	561	682	704	<.001	*	*	
Questions	220	306	306	<.0001	*	*	
Questions / Units	.39	.45	.44	<.02	*	*	
Consecutive Qs / Questions	.33	.42	.42	<.0001	*	*	
Child							
Responses to Questions	130	157	155	<.01	*	*	
Responses / Clinician Qs	.59	.52	.51	<.001	*	*	

Table 7.4: Dialogue characteristics for clinicians and children (means of sessions). * indicates a significant between-group difference (p<.05).

As shown in Table 7.4, column 5, there are group-wise differences for all measures by Kruskal-Wallis (KW) rank sum test (df=2). Post-hoc Wilcoxon rank sum tests (Table 7.4 columns 6-8) showed that, for all measures, these differences could be attributed to significant differences between the sessions involving children with TD as compared to sessions involving children with DLD or ASD. No significant differences were found between the ASD and DLD groups. Essentially, when interacting with the children with ASD or DLD, the clinician spoke more, asked more questions, had a higher ratio of questions, and followed a question with another question more often. In addition, the children with ASD and DLD were less responsive to questions than the children with TD.

These results suggests that the children with ASD and DLD may be choosing to postpone or avoid responding to some questions.[4] This is relevant in that they may be using those opportunities to garner additional time to plan their response. To determine if this is the case, we compared the gaps between single and consecutive questions, as shown in Table 7.5. Although, in these grand mean values, it appears the gaps after consecutive questions may be longer than after single questions, within-subject analyses revealed no

[3] As we included only orderly turn-exchanges, which followed the pattern [clinician speech]..*gap*..[child speech], the actual percentage of child responses is likely higher than that shown here.

[4] Alternatively, the clinician may be asking more rhetorical questions.

significant differences, all p's>.2.. Thus, the children did have not systematically longer gaps after consecutive questions.

	TD	DLD	ASD
Question	6.52	6.51	6.67
Consecutive question	6.58	6.59	6.74

Table 7.5: Subjects' mean gap lengths (log-transformed) after questions and consecutive questions.

7.5 Predicting Mazes

We next create mixed-effects logistic regression models to ascertain if the likelihood of turn-initial mazes is related 1) to the child's gap length, and 2) to whether the clinician produced a non-question, a question, or asked two, or more, consecutive questions prior to the child's response. As it has been shown that the rate of disfluencies is correlated to the length of the upcoming utterance, and that activity has an effect on utterance-initial fillers, we included 3) activity, and 4) C-unit length in the models. The results are shown in table 7.6. Factors that did not contribute either alone or by interaction with other factors (e.g., sex and age) are not included in this final model.

7.5.1 By-group Comparisons

Looking at the coefficients for the by-group models in Table 7.6, we see, as anticipated, that both activity and length significantly contribute to the likelihood of a maze for all three groups, with length increasing the likelihood (TD=1.07, DLD=1.20, ASD=1.25), and the activities *description* (TD=-0.82, DLD=-0.45, ASD=-0.25) and *play* (TD=-0.20, DLD=-0.38, ASD=-0.16) decreasing the likelihood.

Looking next at the predictors of interest, we see that, having been asked a question, or a series of consecutive questions, increases the likelihood that the child will produce a maze. However, looking at the interactions between questions and gaps, we see that the two differ in that the interaction was significant only for single questions, not consecutive questions.

Factors	TD		DLD		ASD		Combined	
Intercept	-3.72	***	-3.87	***	-3.87	***	-3.68	***
Activity=*Description*	-0.82	***	-0.45	***	-0.25	**	-0.81	***
Activity=*Play*	-0.20	*	-0.38	**	-0.16	*	-0.19	.
centered(log(Gap))	0.59	***	0.19		0.44	***	0.43	***
Question	0.90	***	0.60	***	0.48	***	0.93	***
Question+	1.01	***	0.76	***	0.47	***	—	
log(UttLength)	1.07	***	1.20	***	1.25	***	1.05	***
centered(log(Gap)):Question	0.27	*	0.30	*	0.19	*	0.20	***
centered(log(Gap)):Question+	0.00		0.23		0.07		—	
centered(log(Gap)):log(UttLength)	-0.31	***	-0.19	*	-0.27	*	-0.26	***
Dx:DLD							-0.22	
DLD:*Description*							0.33	.
DLD:*Play*							0.21	
DLD:Question and Question+							-0.26	.
DLD:log(UttLength)							0.15	.
Dx:ASD							-0.22	
ASD:*Description*							0.56	***
ASD:*Play*							0.03	
ASD:Question and Question+							-0.47	***
ASD:log(UttLength)							0.21	**
Std Dev (subject)	0.55		0.43		0.55		0.52	
r	0.90		0.35		0.98		0.98	

'***'<0.001, '**'<0.01, '*'<0.05, '.'<0.1

Table 7.6: Coefficients for the models predicting the likelihood of a turn-initial maze. 'Intercept' represents the predicted log(likelihood) using the reference levels for each factor (i.e., activity=*Conversation*, Non-question, log(UttLength) =0, mean log(Gap), and, for the combined model, dx=TD). — For the combined model, collapsing Question and Question+ resulted in a better fitting model.

Next looking at gaps, we see that the likelihood of a maze increases with the length of the gap, but this effect is significant only for the children with TD and ASD. In addition, we see that the length of the C-unit and the length of the gap interact. This interaction can be interpreted as attenuating the effect of gap length, meaning that a short gap followed by a long utterance with be more likely to have a maze than a long gap followed by a short utterance.

Finally, we see that the included factors produce relatively well-fitted models for the children with TD and ASD (both r's >.90), but a less well-fit model for the children with

DLD. [5]

7.5.2 Comparing Between Groups

Looking next at the combined model in Table 7.6, in which the TD group is the reference level, we see that the children with DLD did not differ significantly from the children with TD (i.e., Dx:DLD is not significant). Comparing the children with TD to those with ASD, we see that the ASD group is more likely to produce a maze during the *description* activity (i.e., ASD:*Description*, and is more sensitive to UttLength. However, they also have a lowered likelihood of producing a maze after a question. The combined model was well-fitted to the data, with an r of .98.

In addition to the above models, which treat all mazes as a group, we also created models for 'um's and the non-filler mazes, but were unable to create models with an acceptable goodness-of-fit. Instead to determine if the use of 'um's and 'uh's differ from non-filler mazes after a question, we compared the ratios of each type of filler after questions versus consecutive questions. Figure 7.3 shows this data, with the lines representing the change in ratio for each subject. Comparing between groups, we see that the three groups did not differ substantially in their ratios. However, there are within-subject differences for all three groups in the ratio of non-filler mazes, by Wilcoxon signed rank, all p's $< .04$, and for the TD group in their ratio of 'uh's, $p<.02$.

7.6 Discussion

In terms of gap lengths, the children with DLD did not differ in any significant way from the children with TD, and in some cases did differ from the children with ASD. Thus it appears likely that the ability to minimize pausing is indeed related to social skills rather than language processing abilities. Interestingly, all three groups had shorter gaps after a question, showing that the children with ASD do respond more quickly after questions,

[5] The r computed here is particularly sensitive to bins with sparse data. For example, the low r value for the DLD group was due to a single prediction of .72, in which no maze was present in the actual data. An alternate treatment, which divides the data into slightly overlapping bins, each containing an equal number of data points, resulted in r values of over .98 for all models.

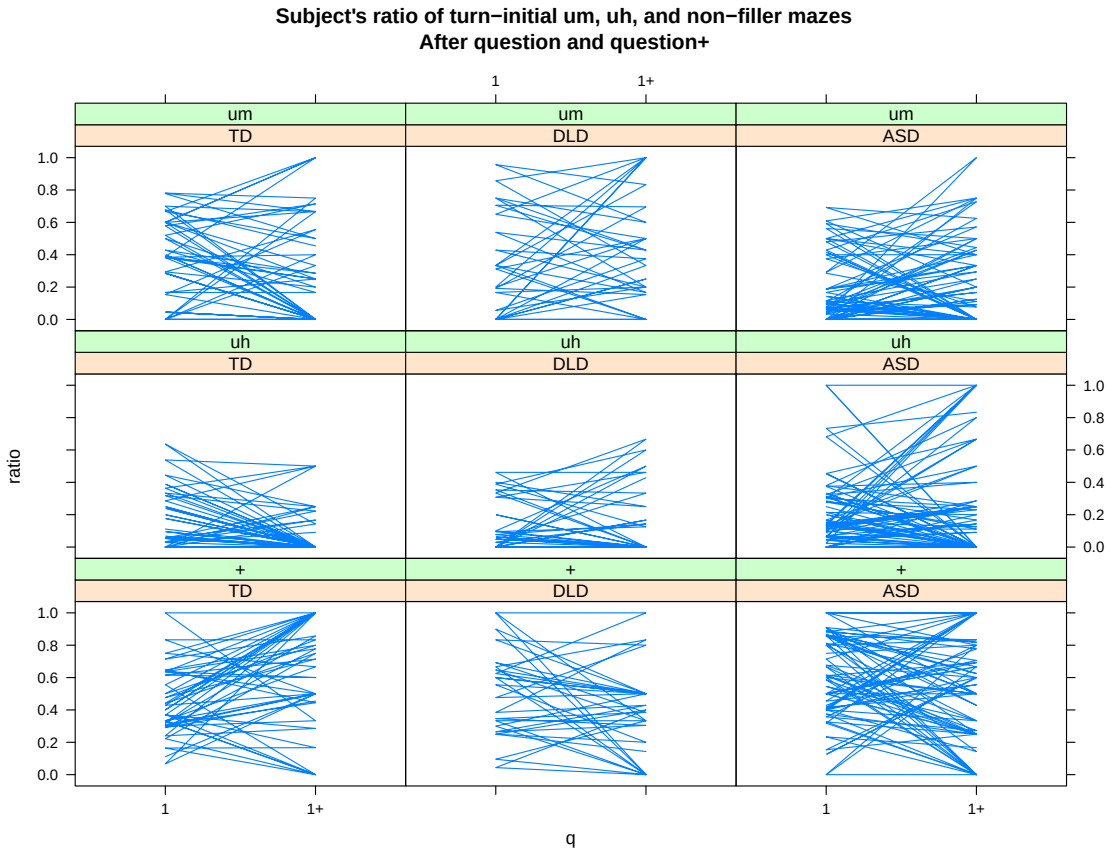


Figure 7.3: Subject's ratio of 'um's, 'uh's and non-filler mazes(+) after a question(1) or consecutive questions(1+).

but are slower to respond overall. It may be that, rather than responding to the increasing social obligation of questions, they respond more quickly to questions because they are instead better able to recognize the stronger cues that accompany questions.

Although the children with DLD produced gaps that resembled those of the children with TD, they differed in other ways. In terms of responsiveness, the children with DLD, and those with ASD, were less responsive to questions, responding to only 52% and 51% of the clinician's questions, versus 59% for the children with TD.

Looking at how the children maze (i.e., 'um', 'uh' or non-filler mazes) when responding to questions and consecutive questions, we did not find substantial differences between the groups, but did find that all three groups differed significantly in their ratio of mazes after questions versus consecutive questions. However, these differences are not all in the same

direction. The children with TD produced a significantly higher ratio of non-filler mazes after consecutive questions, and the children with DLD and ASD produced a lowered ratio. This suggests that the children with impairments may, in fact, be using non-responses as a way to hedge, thus improving the completeness of their ultimate response.

When responding to a question, it seems likely that, especially for the children with TD and DLD, the obligation to respond may lead them to begin speaking before they are ready, thus resulting in increased disfluencies at the beginning of their turn. This we did find, with all the children being more likely to produce a maze after a question and that likelihood increasing with the length of the preceding gap. However, as anticipated, the children with ASD were less likely to produce a maze after a question than the other two groups. Because the children with ASD have longer gaps, and lower mazing response to questions, it appears that the children with ASD may wait until they are ready to speak, rather than using delaying mechanisms or producing incompletely planned, disfluent speech.

7.6.1 Lessons for HCI

This research presents many opportunities for improving SDSs. First, we found that children without social impairments (i.e., those with TD and DLD) respond more quickly to questions, but that a faster response is more likely to begin with a filler or disfluency. With this in mind, SDSs could be designed that are less question-response centric, thus alleviating the social pressure to respond quickly. Alternatively, SDSs could anticipate lengthier gaps and potentially disfluent speech, thus providing a more natural dialogue flow and improved speech recognition results.

7.6.2 Implications for ASD and DLD

Although this dissertation focuses on how these dialogue mechanisms are impacted by the preceding speech, and whether they are primarily social in nature, this work also has implications for differential diagnosis of ASD and DLD. Because gap lengths and mazing appear to be impacted by social impairments, a primary differentiator between children with ASD and DLD, the differences shown here could be used as additional tools to help separate the two disorders.

Chapter 8

Using Reinforcement Learning to Create Dialogue Coordination Strategies for Diverse Users

In this chapter we investigate whether Reinforcement Learning (RL) can be used to create strategies for managing dialogue coordination. We focus on learning a strategy that can match the volume of the system to the needs of users with hearing issues, or those in environments for which a louder (e.g., in-car) or a quieter (e.g., shared office) volume would be advantageous. Speaking at an appropriate volume is related to contact, perception, and understanding, three of the four communicative functions (excluding response) described by Allwood [3]. That is, if a listener can not hear an SDS well enough to understand the speech, then the communicative requirements have not been met. Yet, from an HCI standpoint, we can expect that user's will inform an interlocutor of difficulty, providing an opportunity for the SDS to adapt. To explore how RL might create adaptive strategies, we use a simple communication channel model in which the SDS needs to determine and maintain an amplitude level that is pleasant and effective for users with differing amplitude preferences and needs.

Our long term goal is to learn how to manage the communication channel along with the task, moving away from just "what" to say and also focusing on "how" to say it. For the work in this chapter, our goals are twofold:

1. To formalize a communication channel model that encompasses diverse users, initially focusing just on explicit user actions and implicit system actions.

2. To determine whether RL is an appropriate tool for learning an effective communication channel management strategy for diverse users.

The work in this chapter was previously published as *"Using Reinforcement Learning to Create Communication Channel Management Strategies for Diverse Users"* [57].

8.1 Background and Related Work

8.1.1 Motivation

Dialogue is a social activity. As such, by entering into a dialogue, the parties accept social obligations. For a speaker, the most basic obligation is ensuring that the other participant is paying attention, can hear the speaker, can understand what the speaker is saying, and can process and respond to the speech [2, 23]. Allwood referred to this set of dialogue requirements as contact, perception, understanding and response.

To meet this obligation during human-human communication, speakers manage the communication channel; implicitly altering their manner of speech to increase the likelihood of being perceived and understood while concurrently economizing effort [54]. In addition to these implicit actions, speakers also make statements referring to breakdowns in the communication channel, explicitly pointing out potential problems or corrections, (e.g., "Could you please speak up?") [46].

As for human-computer dialogue, SDS are prone to misrecognition of users' spoken utterances. Much research has focused on developing techniques for overcoming or avoiding system misunderstandings. Yet, as the quality of automatic speech recognition improves and SDS are deployed to diverse populations and in varied environments, systems will need to better attend to possible *human* misunderstandings. Future SDS will need to manage the communication channel, in addition to managing the task, to aid in avoiding these misunderstandings.

Currently, both SDSs and Assistive Technology (AT) tend to have a narrow focus, supporting only a subset of the population. SDS typically aim to support the "average man", ignoring wide variations in potential users' ability to hear and understand the system. AT aims to support people with a recognized disability, but does not support

those whose impairment is not severe enough to warrant the available devices or services, or those who are unaware or have not acknowledged that they need assistance. However, SDS should be able to meet the needs of users whose abilities fall at, and between, the extremes of severely impaired and perfectly abled.

When aiming to support users with widely differing abilities, the cause of a user's difficulty is less important than adapting the communication channel in a manner that aids understanding. For example, speech that is presented more loudly and slowly can help a hearing-impaired elderly person understand the system, and can also help a person with no hearing loss who is driving in a noisy car. Although one user's difficulty is due to impairment and the other due to an adverse environment, a similar adaptation might be appropriate to both.

To create dialogue policies that can balance and optimize measures of task success, researchers have explored the use of reinforcement learning (RL) (e.g., see [84, 52, 40, 101]). Along these lines, RL is potentially well suited to creating policies for the subtask of managing the communication channel, as it can learn to adapt to the user while continuing the dialogue. In doing so, RL may choose actions that appear costly at the time, but lead to better overall dialogues.

8.1.2 How People Manage the Channel

When conversing, speakers implicitly adjust features of their speech (e.g., speaking rate, loudness) to maintain the communication channel. For example, speakers produce Lombard speech when in noisy conditions, produce clear speech to better accommodate a hard of hearing listener, and alter their speech to more closely resemble the interlocutor's [45, 54]. These changes increase the intelligibility of the speech, thus helping to maintain the communication channel [76]. Research has also shown that speakers adjust their speaking style when addressing a computer; exhibiting speech adaptations similar to those seen during human-human communication [16, 60].

In addition to altering their speech implicitly, speakers also explicitly point out communication channel problems [46]. Examples include; requesting a change in speaking rate or amplitude ("Could you please speak up?"), explaining sources of communication

channel interference ("Oh, that noise is the coffee grinder."), or asking their interlocutor to repeat an utterance ("What was that?"). These explicit utterances identify some issue with the communication channel that must be remedied before continuing the dialogue. In response, interlocutors will rewind to a previous point in the dialogue and alter their speech to ensure they are understood. This approach, of adapting ones speech in response to a communication problem, occurs even when conversing with a computer [90].

Both implicit speech alterations and explicit utterances regarding the communication channel often address issues of amplitude. This is to be expected, as speaking at an appropriate amplitude is critical to maintaining an effective communication channel, with sub-optimal amplitude affecting listeners' understanding and performance [10]. In addition, Baldwin [9] found that audible, but lowered, amplitude can negatively affect both younger and older subjects' reaction time and ability to respond correctly while multitasking, and that elderly listeners are likely to need higher amplitudes than younger listeners to maintain similar performance. Just as low amplitude can be difficult to understand, high amplitude can be annoying, and, in the extreme, cause pain.

8.1.3 How Systems Manage the Channel

Despite the importance of maintaining the communication channel, little work has been done in this area. One exception is the work of Martinson and Brock [63], who take advantage of the mobility and sensory capabilities of a robot to improve listener understanding in a potentially noisy environment. In this work, the robot maintains a noise map of the environment, measuring the environmental noise prior to each TTS utterance. The robot then rotates toward the listener, changes location, alters its amplitude, or pauses until the noise abates. We anticipate that a similar technique, useful for remote listeners who may be in a noisy environment or using a noisy communication medium, could analyze the signal-to-noise ratio to ascertain the noise level in the listener's environment. However, although these techniques may be useful for adjusting amplitude to compensate for noise in the listener's environment, they do not address speech alterations needed to accommodate users with different hearing abilities or preferences.

Given the need to adapt to individual users, it seems reasonable that users themselves

would simply adjust volume on their local device. However, there are issues with this approach. First, manual adjustment of the volume would prove problematic when the user's hands and eyes are busy, such as when driving a car. Second, during an ongoing dialogue speakers tend to minimize pauses, responding quickly when given the turn [82]. Stopping to alter the amplitude could result in longer than natural pauses, which systems often respond to with increasingly lengthy 'timeout' responses [49], or repeating the same prompt endlessly [100]. Third, although we focus on amplitude adaptations in this chapter, amplitude is only one aspect of the communication channel. A fully functional communication channel management solution would also incorporate adaptations of features such as speaking rate, pausing, pitch range, emphasis, etc. This extended set of features, because of their number and interaction between them, do not readily lend themselves to listener manipulation.

8.1.4 Reinforcement Learning and Dialogue

SDSs need to be built that can manage the communication channel (e.g., contact and perception) in addition to managing the task. Recently, researchers have been exploring the use of reinforcement learning (RL) to create dialogue policies for spoken dialogue systems that optimize certain measures of task success [84, 52, 101]. These policies specify what action to perform in each possible system state so that a minimum dialogue cost is achieved [101, 52]. To accomplish this, RL starts with a policy, namely what action to perform in each state. It then uses this policy, with some exploration, to estimate the cost of getting from each state with each possible action to the final state. As more simulations are run, RL refines its estimates and its current policy. RL will converge to an optimal solution as long as certain assumptions about costs and state transitions are met. RL is particularly well suited for learning dialogue strategies as it will balance opposing goals such as minimizing excessive confirmations versus ensuring accurate information.

RL has been applied to a number of dialogue scenarios. For form-filling dialogues, in which the user provides parameters for a database query, researchers have used RL to decide what order to use when prompting for the parameters and to decrease resource

costs such as database access [52, 84]. System misunderstanding caused by speech recognition errors has also been modeled to determine whether, and how, the system should confirm information [84]. However, there is no known work on using RL to manage the communication channel so as to avoid *user* misunderstandings.

User Simulation: To train a dialogue strategy using RL, some method must be chosen to emulate realistic user responses to system actions. Training with actual users is generally considered untenable since RL can require millions of runs. As such, researchers create simulated users that mimic the responses of real users. The approach employed to create these users varies between researchers; ranging from simulations that employ only real user data [40], to those that model users with probabilistic simulations based on known realistic user behaviors [52]. Ai et al. suggest that less realistic user simulations that allow RL to explore more of the dialogue state space may perform as well or better than simulations that statistically recreate realistic user behavior [1]. For our work in this chapter, we employ a hand-crafted user simulation that allows full exploration of the state space.

Costs: Although it is agreed that RL is a viable approach to creating optimal dialogue policies, there remains debate as to what cost functions result in the most useful policies [101]. Typically, these costs include a measure of efficiency (e.g., number of turns) and a measure of solution quality (e.g., the user successfully completed the transaction) [37, 84, 52]. For managing the communication channel, it is unclear how the cost function should be structured. In this work we compare two cost components, a more traditional dialogue-length cost versus a novel annoyance cost, to determine which best supports the creation of useful policies.

8.2 Communication Channel Model

Based on the literature reviewed in Section 8.1.2, we devised a preliminary model that captures essential elements of how users manage the communication channel. For now,

we only include explicit user actions, in which users directly address issues with the communication channel, as noted by Jurafsky et al. [46]. In addition, the users modeled are both consistent and amenable; they provide feedback every time the system's utterances are too loud or too soft, and abandon the interaction only when the system persists in presenting utterances outside the user's tolerance (either ten utterances that are too loud or ten that are too soft).

For this work, we wish to create policies that treat all users equitably. That is, we do not want to train polices that give preferential treatment to a subset of users simply because they are more common. To accomplish this, we use a flat rather than normal distribution of users within the simulation, with both the optimal amplitude and the tolerance range randomly generated for each user. To represent users with differing amplitude needs, simulated users are modeled to have an optimal amplitude between 2 and 8, and a tolerance range of 1, 3 or 5. For example, a user may have a optimal amplitude of 4, but be able to tolerate an amplitude between 2 and 6.

When interacting with the computer, the user responds with: (a) the answer to the system's query if the amplitude is within their tolerance range; (b) too soft (TS) if below their range; or (c) too loud (TL) if the amplitude is above their tolerance range. As a simplifying assumption, TS and TL represent any user responses that address communication channel issues related to amplitude. For example, the user response "Pardon me?" would be represented by TS and "There's no need to shout!" by TL. With this user model, the user only responds to the domain task when the system employs an amplitude setting within the user's tolerance range.

For the system, we need to ensure that the system's amplitude range can accommodate any user-tolerable amplitude. For this reason, the system's amplitude can vary between 0 and 10, and is initially set to 5 prior to each dialogue. In addition to performing domain actions, the system specifies the amount the amplitude should change: -2, -1, +0, +1, +2. Each system communication to the user consists of both a domain action and the system's amplitude change. Thus, the system manages the communication channel using only implicit actions. If the user responds with TS or TL, the system will then restate what it just said, perhaps altering the amplitude prior to re-addressing the user.

8.3 Hand-crafted Policies

To help in determining whether RL is an appropriate tool for learning communication channel management strategies, we designed two hand-crafted policies for comparison. The first handcrafted policy, termed *no-complaints*, finds a tolerable amplitude as quickly as possible, then holds that amplitude for the remainder of the dialogue. As such, this policy only changes the amplitude in response to explicit complaints from the user. Specifically, the policy increases the amplitude by 2 after a TS response, and drops it by 2 after a TL. If altering the amplitude by 2 would cause the system to return to a setting already identified as too soft or too loud, the system uses an amplitude change of 1.

The second policy, termed *find-optimal*, searches for the user's optimal amplitude, then maintains that amplitude for the remainder of the dialogue. For this policy, the system first increases the amplitude by 1 until the user responds with TL (potentially in response to the system's first utterance), then decreases the amplitude by 1 until the user either responds with TS or the optimal amplitude is clearly identified based on the previous feedback. An amplitude change of 2 is used only when both the optimal amplitude is obvious and a change of 2 will bring the amplitude setting to the optimal amplitude.

8.4 RL and System Encoding

To learn communication channel management policies we use RL with system and user actions specified using Information State Update rules [40]. Using the software and techniques described by Heeman [37], we encode commonsense preconditions rather than trying to learn them, and only use a subset of the information state for RL.

8.4.1 Domain Task

We use a domain task that requires the user to supply 9 pieces of information, excluding user feedback relating to the communication channel. The system has a deterministic way of selecting its actions, thus no learning is needed for the domain task.

8.4.2 State Variables

For RL, each state is represented by two variables; *AmpHistory* and *Progress*. *AmpHistory* models the user by tracking all previous user feedback. In addition, it tracks the current amplitude setting. The string contains one slot for each potential amplitude setting (0 through 10), with the current setting contained within "[]". Thus, at the beginning of each interaction, the string is "`-----[-]-----`", where "-" represents no known data. Each time the user responds, the string is updated to reflect which amplitude settings are too soft ("<"), or within the user's tolerance ("O"). When the user responds with TL/TS, the system also updates all settings above/below the current setting. The *Progress* variable is required to satisfy the Markov property needed for RL. This variable counts the number of successful information exchanges (i.e., the user did not respond with TS or TL). As the domain task requires 9 pieces of information, the Progress variable ranged from 1 to 9.

8.4.3 Costs

Our user model only allows up to 10 utterances that are too soft or too loud. If the cutoff is reached, the domain task has not been completed, so a solution quality cost of 100 is incurred. Cutting off dialogues in this way has the additional benefit of preventing a policy from looping forever during testing. During training, to allow the system to better model the cost of choosing the same action repeatedly, we use a longer cutoff of 1000 utterances rather than 10.

As our goal includes decreasing the amount of potentially annoying utterances (i.e., those in which the system's amplitude setting is in discord with the user's optimal amplitude), we introduce a user-centric cost metric, which we have termed *annoyance cost*. The annoyance cost (AC) assigns a cost calculated as the difference between the system's amplitude setting and the user's optimal amplitude. This difference is multiplied by 3 when the system's amplitude setting is below the user's optimal. This multiplier was chosen based on research that demonstrated increased response times and errors during cognitively challenging tasks when speech was presented below, rather than above, typical conversational levels [10]. Thus, only utterances at the optimal amplitude have no cost.

DC			
AmpHistory	System	Amp	User
-----[-]-----	$Query_1$ +0	5	TS
<<<<<[<]-----	$Query_1$ +2	7	Answer
<<<<<<-[0]---	$Query_2$ +0	7	Answer
<<<<<<-[0]---	$Query_3$ +0	7	Answer
<<<<<<-[0]---	$Query_4$ +0	7	Answer
<<<<<<-[0]---	$Query_5$ +0	7	Answer
<<<<<<-[0]---	$Query_6$ +0	7	Answer
. . .	. . .	. . .	. . .
dialogue length cost = 10			

AC			
AmpHistory	System	Amp	User
-----[-]-----	$Query_1$ +1	6	TS
<<<<<<[<]----	$Query_1$ +1	7	Answer
<<<<<<<[0]---	$Query_2$ +1	8	Answer
<<<<<<<0[0]--	$Query_3$ +1	9	Answer
<<<<<<00[0]-	$Query_4$ +1	10	TL
<<<<<<<000[>]	$Query_4$ -2	8	Answer
<<<<<<<0[0]0>	$Query_5$ +0	8	Answer
. . .	. . .	. . .	. . .
annoyance cost = 12			

Table 8.1: Comparison of DC (left) and AC (right) interactions with a user who has an optimal amplitude of 8 and a tolerance range of 3. The policies continue as shown, without changing the amplitude level, until all queries are answered.

We also utilize a second, more traditional, cost component based on the length of the dialogue. The dialogue-length cost (DC), assigns a cost of 1 for each user utterance.

8.5 Results

With the above system and user models, we trained policies using the two cost functions discussed above, eight with the DC component and eight using the AC component. All used Q-Learning and the ϵ-greedy method to explore the state space with ϵ set at 20% [92]. Dialogue runs were grouped into epochs of 100; after each epoch, the current dialogue policy was updated. We trained each policy for 60,000 epochs. After certain epochs, we tested the policy on 5000 user tasks.

For our simple domain, the solution quality cost remained 0 after about the 100th epoch, as all policies learned to avoid user abandonment. Because of this, only the dialogue-length cost (DC) and annoyance cost (AC) components are reflected in the following analyses.

8.5.1 DC-Trained Policies

By 40,000 epochs, all eight DC policies converged to one common optimal policy. Dialogues resulting from the DC policies average 9.76 user utterances long. DC policies start each dialogue using the default amplitude setting of 5. After receiving the initial user response, they aggressively explore the amplitude range. If the initial user response is TL (or TS),

they continue by decreasing (or increasing) the amplitude by -2 (or +2) until they find a tolerable volume, in which case they stop. The interaction shown in Table 8.1 (left) illustrates the above noted aspects of the policy. Additionally, if the policy receives user feedback that is contrary to the last feedback (i.e., TS after TL, or TL after TS), the policy backtracks one amplitude setting. In addition, if the current amplitude is near the boundary (3 or 7), the policy will change the volume by -1 or +1 as changing it by -2 or +2 would cause it to move outside users' amplitude range of 2-8. In essence, the DC policies are quite straightforward; aggressively changing the amplitude if the user complains, and assuming the amplitude is correct if the user does not complain.

8.5.2 AC-Trained Policies

By 55,000 epochs, AC policies converged to one of two optimal solutions, with an average annoyance cost of 7.49. As illustrated in Table 8.1 (right), the behavior of the AC policies is substantially more complex than the DC policies. First, the AC policies start by increasing the amplitude, delivering the first utterance at a setting of 6 or 7. Second, the policies do not stop exploring after they find a tolerable setting, instead attempting to bracket the user's tolerance range, thus identifying the user's optimal amplitude. Third, AC policies sometimes avoid lowering the amplitude, even when doing so would concretely identify the user's optimal amplitude. By doing so, the policies potentially incur a cost of 1 for all following turns, but avoid incurring a one time cost of 3 or 6. In essence, the AC policies attempt to find the user's optimal amplitude but may stop short as they approach the end of the dialogue, favoring a slightly too high amplitude over one that might be too low.

8.5.3 Comparing AC- and DC- Trained Policies

The costs for the AC and DC trained policy sets cannot be directly compared as each set used a different cost function. However, we can compare them using each others' cost function.

Comparison using Dialogue Cost

First, we compare the two sets of policies in terms of average dialogue-length. For example, in Table 8.1, following a DC policy results in a dialogue-length of 10. However, for the same user, following the AC policy results in a dialogue-length of 11, one utterance longer due to the TL response to Query_4.

The average dialogue-length of the DC and AC policies, averaged across users, is shown in the rightmost two columns of Figure 8.1. As expected, the DC policies perform better in terms of dialogue-length, averaging 9.76 utterances long. However, the AC policies average 10.32 utterances long, only 0.52 utterances longer. This similarity in length is to be expected, as system communication outside the user's tolerance range impedes progress and is costly using either cost component.

We also compared the AC and DC policies' average dialogue-length for users with the same optimal amplitude (i.e., each column shows the average cost across users with tolerance ranges of 1, 3 and 5), as shown in Figure 8.1. From this figure it is clear that there is little difference in dialogue-length between AC and DC policies for users with the same optimal amplitude. In addition, for both policies, the lengths are similar between users with differing optimal amplitudes.

Comparison using Annoyance Cost

Second, we compare the two sets of polices in terms of annoyance costs. For example, in Table 8.1, following the AC policy results in an annoyance cost of 12. For the same user, following the DC policy results in an annoyance cost of 36; 9 for Query_1 as it is three below the user's optimal amplitude, and 3 for each of the following nine utterances as they are all one below optimal.

As shown in the rightmost columns of Figure 8.2, DC policies average annoyance cost was 13.35, a substantial 78% increase over the average cost of 7.49 for AC policies. Figure 8.2 also illustrates that the AC and DC policies perform quite differently for users with differing optimal amplitudes. For example, users of the DC policies whose optimal is at (5), or slightly below (4), the system's default setting (5) average lower annoyance costs

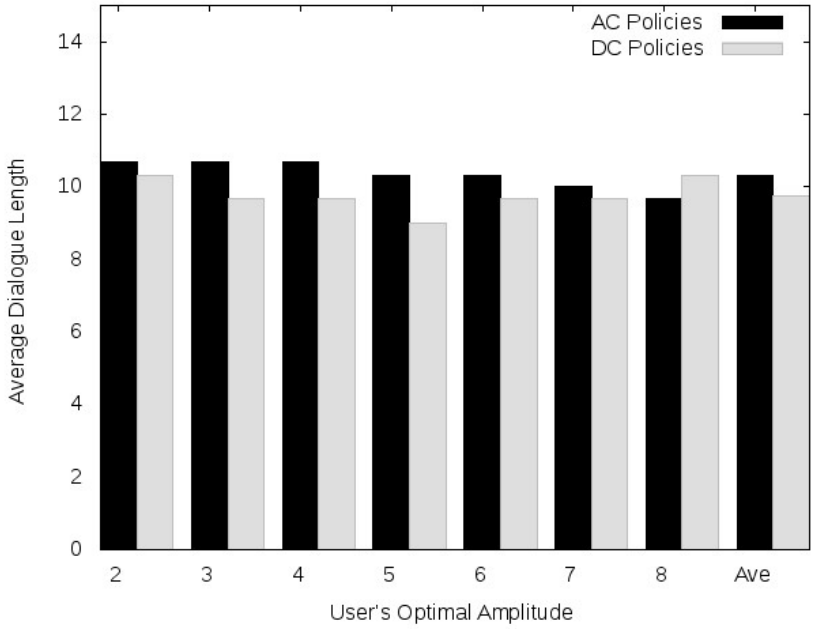


Figure 8.1: Comparison of the dialogue-length between AC and DC policies for users with differing optimal amplitudes.

than those using the AC policies. However, these lowered costs for users in the mid-range is gained at the expense of users whose optimal amplitude is farther afield, especially those users requiring higher amplitude settings. This substantial difference between users with different optimal amplitudes is because, for DC policies, the interaction is often conducted at the very edge of the users' tolerance. In contrast, the AC policies risk more intolerable utterances, but use this information to decrease overall costs by better meeting users' amplitude needs. As such, users of the AC policies can expect the majority of the task to be conducted at, or only one setting above, their optimal amplitude.

8.5.4 Comparing Hand-crafted and Learned Policies

Each of the two hand-crafted policies were run with each user simulation (i.e., optimal amplitude from 2-8 and tolerance ranges of 1, 3, or 5). In addition, we varied the domain task size, requiring between 4 and 10 pieces of information. DC and AC policies were also trained for these domain task sizes.

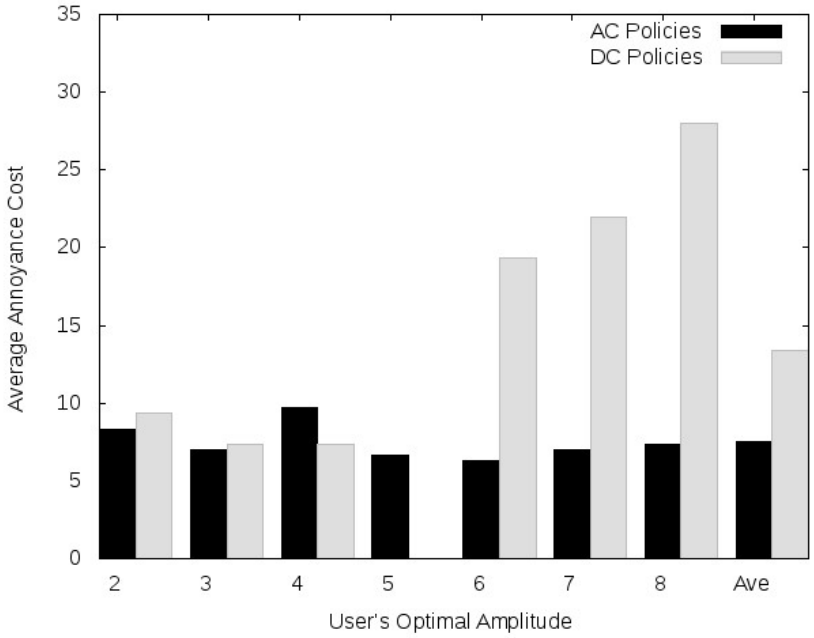


Figure 8.2: Comparison of the annoyance cost between AC and DC policies for users with differing optimal amplitudes.

As shown in Figure 8.3, The no-complain policy's annoyance costs ranged from 7.81 for dialogues requiring four pieces of information to 14.67 for those requiring ten pieces. The cost increases linearly with the amount of information required, because the no-complain policy maintains the first amplitude setting found that does not result in a user response of TS or TL. This ensures the amplitude setting is tolerable to the user, but may not be the user's optimal amplitude.

In contrast, the find-optimal policy's annoyance costs initially increase from 9.67 for four pieces of information to 12.24 for seven through ten pieces. The cost does not continue to increase when the amount of information required is greater than seven because, for dialogues long enough to allow the system to concretely identify the user's optimal amplitude, the cost is zero for all subsequent utterances.

Figure 8.3 also includes the mean annoyance cost for the DC and AC policies. Although one might expect the DC trained policies to resemble the no-complain policy, the learned policy performs slightly better. This difference is because the DC policies learn the range

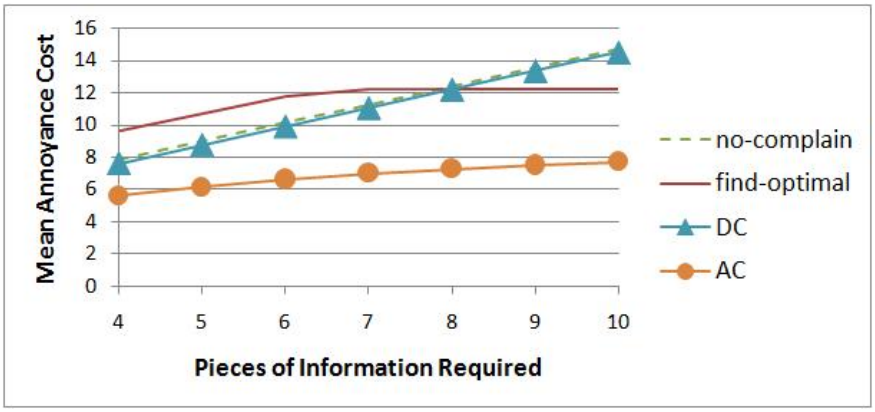


Figure 8.3: Average user annoyance costs for hand-crafted, DC and AC policies across dialogues requiring differing amounts of information.

of users' optimal amplitude settings (2-8), and do not move the amplitude below 2 or above 8. In contrast, the no-complain policies behave consistently regardless of the current setting, and thus will incur costs for exploring settings outside the range of users' optimal amplitudes. Similarly, AC policies could be anticipated to closely resemble the find-optimal policy. However, the AC policies average cost is lower than the costs for either hand-crafted policy, regardless of the amount of information required.

This difference is, in part, due to differences in behavior at the ends of the users' optimal amplitude range, like the DC policies. However, additional factors include the AC policies' more varied use of amplitude changes and their balancing of the remaining duration of the dialogue against the cost to perform additional exploration, as discussed in subsection 8.5.2.

8.6 Discussion

The first objective of this work was to create a model of the communication channel that takes into account the abilities and preferences of diverse users. In this model, each user has an optimal amplitude, but will answer a system query delivered within a range around that amplitude, although they find non-preferred, especially too soft, amplitudes annoying. When outside the user's tolerance, the user provides explicit feedback regarding

the communication channel breakdown. For the system, the model specifies a composite system action, pairing a domain action with a possible communication channel management action to change the amplitude. By modeling explicit user actions, and implicit system actions, this model captures some essential elements of how people manage the communication channel.

The second objective was to determine whether RL is appropriate for learning communication channel management. Towards this end, we compared handcrafted solutions to learned policies. As expected, the learned policies found and maintained a tolerable amplitude setting and eliminated user abandonment. In addition, we found that the learned policies performed better than the hand-crafted policies, regardless of domain task size. This was primarily due to RL's ability (especially for the AC policies) to balance two sets of opposing goals: 1) the effort to find the user's optimal amplitude versus the dialogue-length, and 2) the needs of diverse users. This illustrates the strength of RL for solving the communication channel management problem.

An added benefit of RL is that it optimizes the system's behavior for the users on which it is trained. In this work, we purposely used a flat distribution of users, which caused RL to find a policy (especially when using annoyance costs) that does not penalize the outliers, which are usually those with special needs. In fact, we could modify the user distribution, or the simulated users' behavior, and RL would optimize the system's behavior automatically.

In this study, we contrasted dialogue length (DC) against annoyance cost (AC) components. We found that the AC and DC policies share the objective of finding an amplitude setting within the user's tolerance range because both incur stepwise costs for intolerable utterances. But, AC policies further refine this objective by incurring costs for tolerable, but non-optimal, amplitudes as well. AC policies are using information that is not explicitly communicated to the system, but which none-the-less RL can use while learning a policy.

As this was exploratory work, the user model does not yet fully reflect expected user behavior. For example, as the system's amplitude decreases, users may misunderstand the system's query or fail to respond at all. In future work we will use an enhanced user model

that includes more natural user behavior. In addition, because we wanted the system to focus on learning a communication channel management strategy, the domain task was fixed. In future work, we will use RL to learn policies that both accomplish a more complex domain task, and model connections between domain tasks and communication channel management. Ultimately, we need to conduct user-testing to measure the efficacy of the communication channel management policies. We feel confident that learned policies trained using a communication channel model which reflects the range of users' abilities and preferences will prove effective for supporting all users.

8.6.1 Lessons for HCI

In terms of HCI, this work shows that policies can be learned that engage in dialogue coordination. In addition, it illustrates that by simulating a broader range of users, and employing non-traditional costs, policies created using RL can adapt to diverse users.

Chapter 9

Summary and Conclusions

The goal of this dissertation was to identify dialogue coordination mechanisms and assess how they can be used to improve human-computer interaction (HCI). In chapter 1, we suggested three ways in which this dissertation work can contribute to improved SDSs. We now revisit those suggestions in light of the work presented in this dissertation.

> First, SDS designers can simply acknowledge that aspects of people's communication are realized through dialogue coordination mechanisms and design systems that react accordingly. For example, SDS designers could recognize that speakers typically identify to whom they are addressing a question or request, and build systems that respond only when being addressed.

In chapters 3 and 4, we explored dialogue coordination mechanisms people use to differentiate a computer from human addressee during human-computer and multi-party contexts. We found that increased speech amplitude is a reliable indicator of speech addressed to a computer, regardless of context. Interestingly, this contrasted with existing work on human-human cues of addressee, which found cues other than amplitude (e.g., gaze) to be most informative when identifying addressee. In chapter 5 we explored this dichotomy, finding that when human observers could see the speaker, they were much more likely to assume the computer (rather than a human) was being addressed. Thus, it appears that gaze is an expected cue of addressee, but is less reliable when a speaker is addressing a computer.

These findings point the way toward SDSs that can react when addressed, without the need for users to explicitly indicate a desire to engage the system. Instead, users would be

able to speak naturally, with the system recognizing that increased amplitude indicates speech addressed to the system. Follow-on work to that included in this dissertation found that users were able to successfully engage the system 86% of the time [73]. However, it is important to note that the user's increased amplitude is likely to be a natural response to perceiving the computer as a less capable interlocutor. Thus, systems that provide cues of attentiveness (e.g., embodied agents which emulate gaze), may find other dialogue coordination mechanisms more salient.

> Second, SDSs can be designed to use dialogue coordination mechanisms that are appropriate for the current context. For example, an SDS could backchannel (e.g., "uh-huh") to indicate that it is able to understand the user's speech, but refrain when the message is unclear.

In chapter 6 we examined the use of the fillers 'um' and 'uh', finding that 'um' is used significantly less often by children with social impairments (ASD) than by children with no impairments (TD) or language processing impairments (DLD). However, the three groups of children did not differ in their ratio of 'uh's. In addition, for the TD group, the rate of 'um's increased with age and significant differences were found in the likelihood and length of pauses after 'um' versus 'uh'. These findings suggest that the appropriate (i.e., adult) use of 'um' is learned and listener-oriented. In contrast, the use of 'uh' appears to be unrelated to social skills, suggesting that 'uh' is instead speaker-oriented.

From an SDS design perspective, these findings show that the fillers 'um' and 'uh' should not be treated as interchangeable. Instead, 'um' can be used as a dialogue coordination mechanisms, used to indicate an SDS's desire for additional time to respond or to indicate uncertainty. However, based on both this dissertation work, and the work of others [13, 48], it appears that 'uh' might best be used inside the utterance, to indicate a new, or uncommon, referent.

> Third, SDS designers could design systems that anticipate how the system's dialogue coordination mechanisms will impact a user's speech. For example, systems could be designed that anticipate longer inter-turn pauses when the

user is asked a question, or proactively adapt their actions so as to minimize a user's cognitive load.

In chapter 7 we examined turn-taking, analyzing what factors effect the timeliness and fluency of responses produced by children with TD, DLD, and ASD. In this work, we compared responses after a question to those after a non-question. Here we found that all three groups of children responded more quickly after a question than after a non-question, but that children with ASD were slower to respond in general. In addition, we found that all the children were more likely to produce a maze after a question, and that the likelihood of a maze increased with the length of the preceding pause, but that children with ASD were less likely to produce a maze after a question than the other two groups of children. This work suggest two important points: 1) Questions confer a social obligation to respond in a timely manner; and 2) that speakers may respond to this obligation by speaking before fully prepared, thus producing more mazes.

From an SDS perspective, chapter 7 shows that future systems should be able to anticipate the effect of the system's dialogue coordination mechanisms. This could be accomplished in two ways. First, after a system query, the system could anticipate that user's may produce more variable pauses lengths and a higher rate of mazes after longer pauses, perhaps using language models that better accommodate disfluencies. Second, SDSs could proactively anticipate situations in which a user might become disfluent due to an obligation conferred by the system, and avoid doing so. For example, the system could replace a wh-question (e.g., "What radio station genre would you like?") with a list of options ("I have four radio station genres; rock, country, classical, and indie.").

Finally, to demonstrate that SDSs could be trained to take advantage of dialogue coordination mechanisms, in chapter 8, we use Reinforcement Learning to train a dialogue policy that can adapt the system's loudness based on a user's dialogue coordination mechanisms.

Secondary Contributions

We now discuss a number of secondary contributions of this dissertation.

This dissertation exposes the previously un-acknowledged high rate of self-directed speech that can be expected when users are cognitively challenged. Although this self-talk may present a challenge for SDSs deployed in environments in which user's are already cognitively loaded (e.g., in-car or educational), self-talk also provides an opportunity for evaluation of interfaces. As self-directed speech implies user difficulty, and the content of self-directed speech includes those aspects of the task that the user is finding difficult, analyses of self-talk could indicate aspects of an interface that should considered for redesign.

This dissertation also showcases the use of a wide range of approaches. Herein, these differing approaches allowed us to address questions that could not have been answered using a single approach. Using a wizard-of-oz system allowed us to examine how speaker's behavior differs when addressing a computer versus human, a result that could not be inferred from human-human interaction. Using a perceptual study, we were able to determine cues that people use the differentiate speech addressed to a computer versus a nearby human. Using data collected during human-human interaction, we were able to investigate what mechanisms people naturally use to coordinate dialogue, how these mechanisms interact, and what mechanisms are driven by social pressure, results that could not be inferred from human-computer interaction. Finally, we used Reinforcement Learning to create dialogue policies that incorporate dialogue coordination mechanisms.

In addition, the work herein also provides information relevant to cognitive models of dialogue. Showing that 'um' and 'uh' are likely produced by different cognitive processes helps to explain other's work showing that listeners do not interpret 'um' and 'uh' in the same way [13].

This work also provides insights into autism and potential diagnostic criteria. We have shown differences in the dialogue coordination mechanisms, and responses to dialogue coordination mechanisms, of children with TD, DLD, and ASD. These findings suggest that the social impairments in children with ASD effect not only their ability to produce a coherent response, but also effect the dialogue coordination aspects of their speech. Perhaps more importantly, we found easily identifiable differences between the children with ASD and DLD, for whom differential diagnosis is often time-consuming and difficult.

9.1 Conclusion

In conclusion, we have shown that speaker's dialogue coordination mechanisms are regular, and are thus amenable to computer recognition and use. We have also shown that although we can use human-human cues as a starting point, we cannot assume that the dialogue coordination mechanisms used when interacting with a computer will be the same as those used during human-human interaction. In addition, we must be careful to not assume that regular dialogue behaviors, such as fillers, are dialogue coordination mechanisms. Finally, future SDS should adapt their behavior, accounting for the effects of their own dialogue coordination mechanisms, either proactively, so as to minimize effects on the user's response, or retroactively, allowing for differences in the user's behavior when the system places pressure on the user.

Bibliography

[1] Ai, H., Tetreault, J. R., and Litman, D. J. Comparing User Simulation Models for Dialog Strategy Learning. In *NAACL-HLT* (Apr. 2007).

[2] Allwood, J. Obligations and Options in Dialogue. *THINK Quarterly 3* (1994), 9–18.

[3] Allwood, J., Nivre, J., and Ahlsén, E. On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics 9*, 1 (Jan. 1993), 1–26.

[4] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR)*. Washington DC, 2000.

[5] Arnold, J. E., Fagnano, M., and Tanenhaus, M. K. Disfluencies Signal Theee, Um, New Information. *Journal of Psycholinguistic Research 32*, 1 (Jan. 2003), 25–36–36.

[6] Arthur, A. M., Lunsford, R., Wesson, M., and Oviatt, S. Prototyping novel collaborative multimodal systems: simulation, data collection and analysis tools for the next decade. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (New York, NY, USA, 2006), ACM, pp. 209–216.

[7] Baayen, R. H. *Analyzing linguistic data: A Practical Introduction to Statistics using R*. Cambridge University Press, 2008.

[8] Bakx, I., Van Turnhout, K., and Terken, J. Facial orientation during multi-party interaction with information kiosks. In *Proceedings of the Interact Conference 2003* (2003), Academic Press, pp. 163–170.

[9] Baldwin, C. L. Impact of age-related hearing impairment on cognitive task performance: evidence for improving existing methodologies. In *Human Factors and Ergonomics Society Annual Meeting; Aging* (2001), pp. 245–249.

[10] Baldwin, C. L., and Struckman-Johnson, D. Impact of speech presentation level on cognitive task performance: implications for auditory display design. *Ergonomics 45*, 1 (2002), 62–74.

[11] Bard, E. G., Aylett, M., and Bull, M. More than a Stately Dance: Dialogue as a Reaction Time Experiment. In *Society For Text & Discourse* (2000).

[12] Bard, E. G., Lickley, R. J., and Aylett, M. P. Is disfluency just difficulty? In *DISS'01* (2001), ISCA Tutorial and Workshop, pp. 97–100.

[13] Barr, D. J. Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In *Oralité and gestualité: Interactions et comportements multimodaux dans la communication* (2001), C. Cavé, I. Guaïtella, and S. Santi, Eds., L'Harmattan, pp. 597–600.

[14] Bates, D. *Linear mixed-effects models using S4 classes*, Aug. 2011.

[15] Batliner, A., Zeissler, V., Nöth, E., and Niemann, H. Prosodic Classification of Offtalk: First Experiments. In *TSD '02: Proceedings of the 5th International Conference on Text, Speech and Dialogue* (London, UK, 2002), Springer-Verlag, pp. 357–364.

[16] Bell, L., Gustafson, J., and Heldner, M. Prosodic adaptation in humancomputer interaction. In *Proceedings of ICPhS 03* (2003), vol. 1, pp. 833–836.

[17] Berk, L. E. Why children talk to themselves. *Scientific American 271*, 5 (1994), 78–83.

[18] Bishop, D. V., Whitehouse, A. J., Watt, H. J., and Line, E. A. Autism and diagnostic substitution: evidence from a study of adults with a history of developmental language disorder. *Developmental medicine and child neurology 50*, 5 (May 2008), 341–345.

[19] Boersma, P. Praat, a system for doing phonetics by computer. *Glot International 5*, 9/10 (2001), 341–345.

[20] Bunt, H. Dialogue Control Functions and Interaction Design. In *Dialogue in Instruction* (1995), pp. 197–214.

[21] Bunt, H. *Dynamic Interpretation and Dialogue Theory*, vol. 2. Johns Benjamin, 2000, pp. 139–188.

[22] Buxton, W. Integrating the periphery and context: A new taxonomy of telematics. In *Proceedings of the Graphics Interface Conference* (1995), Morgan Kaufman, pp. 239–246.

[23] Clark, H. H. *Using Language*. Cambridge University Press, Cambridge, May 1996.

[24] Clark, H. H., and Fox Tree, J. E. Using uh and um in spontaneous speaking. *Cognition 84*, 1 (May 2002), 73–111.

[25] Comblain, A. Working Memory In Down Syndrome : Training The Rehearsal Strategy. *Down's Syndrome: Research and Practice 2*, 3 (1994), 123–126.

[26] Corley, M., and Stewart, O. W. Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Language and Linguistics Compass 2*, 4 (2008), 589–602.

[27] Coulston, R., Oviatt, S., and Darves, C. Amplitude Convergence in Children's Conversational Speech with Animated Personas. In *7th International Conference on Spoken Language Processing (ICSLP'02)* (2002), vol. 4, pp. 2689–2692.

[28] Czaja, S. J., and Lee, C. C. *Designing computer systems for older adults*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003, ch. Designing computer systems for older adults, pp. 413–427.

[29] Duncan, R. M., and Cheyne, J. A. Private speech in young adults: Task difficulty, self-regulation, and psychological predication. *Cognitive Development 16* (2002), 889–906.

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

[31] Escera, C. An electrophysiological and behavioral investigation of involuntary attention towards auditory frequency, duration and intensity changes. *Cognitive Brain Research 14*, 3 (Nov. 2002), 325–332.

[32] Fox Tree, J. E. Interpreting Pauses and Ums at Turn Exchanges. *Discourse Processes 34*, 1 (2002), 37–55.

[33] Gina-Anne Levow, . Prosodic Cues to Discourse Segment Boundaries in Human-Computer Dialogue. In *5th SIGdial Workshop on Discourse and Dialogue* (2004), pp. 93–96.

[34] Goldman-Eisler, F. *Psycholinguistics: Experiments in spontaneous speech*. Academic Press, 1968.

[35] Gravano, A., and Hirschberg, J. Turn-Yielding Cues in Task-Oriented Dialogue. In *Proceedings of the SIGDIAL 2009 Conference* (London, UK, Sept. 2009), Association for Computational Linguistics, pp. 253–261.

[36] Grice, H. P. Logic and conversation. In *Syntax and Semantics 3: Speech acts*, P. Cole and J. L. Morgan, Eds. Academic Press, New York, 1975, pp. 41–58.

[37] Heeman, P. Combining Reinforcement Learning with Information-State Update Rules. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* (Rochester, NY, Apr. 2007), pp. 268–275.

[38] Heeman, P. A., Lunsford, R., Selfridge, E., Black, L., and van Santen, J. Autism and Interactional Aspects of Dialogue. In *SIGdial* (Sept. 2010), pp. 249–252.

[39] Heldner, M., and Edlund, J. Pauses, gaps and overlaps in conversations. *Journal of Phonetics 38*, 4 (2010), 555–568.

[40] Henderson, J., Lemon, O., and Georgila, K. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Comput. Linguist. 34*, 4 (Dec. 2008), 487–511.

[41] Hirschberg, J., and Grosz, B. Intonational features of local and global discourse structure. In *HLT '91: Proceedings of the workshop on Speech and Natural Language* (Morristown, NJ, USA, 1992), Association for Computational Linguistics, pp. 441–446.

[42] Hirschberg, J., and Nakatani, C. H. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th annual meeting of the Association for Computational Linguistics* (Morristown, NJ, USA, 1996), Association for Computational Linguistics, pp. 286–293.

[43] Hudson Kam, C. L., and Edwards, N. A. The use of uh and um by 3- and 4-year-old native English-speaking children: Not quite right but not completely wrong. *First Language 28*, 3 (Aug. 2008), 313–327.

[44] Iyengar, G., and Neti, C. A Vision-Based Microphone Switch for Speech Intent Detection. In *RATFG-RTS '01: Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01)* (Washington, DC, USA, 2001), IEEE Computer Society, pp. 101+.

[45] Junqua, J. C. The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America 93*, 1 (Jan. 1993), 510–524.

[46] Jurafsky, D., Shriberg, L., and Biasca, D. Switchboard: SWBD-DAMSL Coders Manual, 1997.

[47] Katzenmaier, M., Stiefelhagen, R., and Schultz, T. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces* (New York, NY, USA, 2004), ACM, pp. 144–151.

[48] Kidd, C., White, K. S., and Aslin, R. N. Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science* (2011), 1–10.

[49] Kotelly, B. *The Art and Business of Speech Recognition.* Addison-Wesley, Jan. 2003.

[50] Lake, J., Humphreys, K., and Cardy, S. Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders. *Psychonomic Bulletin & Review 18*, 1 (Feb. 2011), 135–140.

[51] Levelt, W. Monitoring and self-repair in speech. *Cognition 14*, 1 (July 1983), 41–104.

[52] Levin, E., Pieraccini, R., and Eckert, W. A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies. *IEEE Transactions on Speech and Audio Processing 8*, 1 (2000), 11–23.

[53] Lickley, R. J. Dialogue Moves and Disfluency Rates. In *DiSS '01* (2001), ISCA Tutorial and Workshop, pp. 93–96.

[54] Lindblom, B. *Explaining phonetic variation: A sketch of the H & H theory.* Kluwer Academic Publishers, 1990, pp. 403–439.

[55] Lord, C., Risi, S., Lambrecht, L., Cook, E., Leventhal, B., DiLavore, P., Pickles, A., and Rutter, M. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autium. *Journal of Austism Developmental Disorders 30*, 3 (June 2000), 205–223.

[56] Lunsford, R. Private speech during multimodal human-computer interaction. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (New York, NY, USA, 2004), ACM, p. 346.

[57] Lunsford, R., and Heeman, P. A. Using Reinforcement Learning to Create Communication Channel Management Strategies for Diverse Users. In *1st Workshop on Speech and Language Processing for Assistive Technologies* (Los Angeles, June 2010).

[58] Lunsford, R., Heeman, P. A., Black, L., and van Santen, J. Autism and the use of fillers: Differences between 'um' and 'uh'. In *DISS-LPSS* (Sept. 2010), pp. 107–110.

[59] Lunsford, R., and Oviatt, S. Human perception of intended addressee during computer-assisted meetings. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (New York, NY, USA, 2006), ACM, pp. 20–27.

[60] Lunsford, R., Oviatt, S., and Arthur, A. M. Toward open-microphone engagement for multiparty interactions. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (New York, NY, USA, 2006), ACM, pp. 273–280.

[61] Lunsford, R., Oviatt, S., and Coulston, R. Audio-visual cues distinguishing self- from system-directed speech in younger and older adults. In *Proceedings of the 7th International Conference on Multimodal Interfaces* (New York, NY, USA, 2005), ACM, pp. 167–174.

[62] Luria, A. R. *The Role of speech in the regulation of normal and abnormal behavior*. Liveright, NY, USA, 1961.

[63] Martinson, E., and Brock, D. Improving human-robot interaction through adaptation to the auditory scene. In *HRI '07: Proceedings of the ACM/IEEE international conference on Human-robot interaction* (New York, NY, USA, 2007), ACM, pp. 113–120.

[64] Meichenbaum, D., and Goodman, J. Reflection-impulsivity and verbal control of motor behavior. *Child Development 40* (1969), 785–797.

[65] Messer, D. J. The identification of names in maternal speech to infants. *Journal of Psycholinguistic Research 10*, 1 (Jan. 1981), 69–77.

[66] Messer, S. B. Reflection-impulsivity: A review. *Psychological Bulletin 83*, 6 (1976), 1026–1052.

[67] Mouridsen, S. E., and Hauschild, K. M. A longitudinal study of autism spectrum disorders in individuals diagnosed with a developmental language disorder as children. *Child: Care, Health and Development 35*, 5 (2009), 691–697.

[68] Neti, C., Iyengar, G., Potamianos, G., Senior, A., and Maison, B. Perceptual Interfaces For Information Interaction: Joint Processing Of Audio And Visual Information For Human-Computer Interaction. In *In Proceedings of the International Conference on Spoken Language Processing, volume III* (2000), vol. 2000, pp. 11–14.

[69] Oviatt, S., Cohen, P., Fong, M., and Frank, M. A Rapid Semi-Automatic Simulation Technique for Investigating Interactive Speech and Handwriting. In *International Conference on Spoken Language Processing* (1992), pp. 1351–1354.

[70] Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., and Carmichael, L. Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (New York, NY, USA, 2003), ACM, pp. 44–51.

[71] Oviatt, S., Darves, C., and Coulston, R. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *TOCHI: ACM Transactions on Compututer-Human Interaction 11*, 3 (2004), 300–328.

[72] Oviatt, S., MacEachern, M., and Levow, G. A. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication 24*, 2 (1998), 87–110.

[73] Oviatt, S., Swindells, C., and Arthur, A. Implicit user-adaptive system engagement in speech and pen interfaces. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 969–978.

[74] Paek, T., Horvitz, E., and Ringger, E. Continuous listening for unconstrained spoken dialog. In *Proceedings of the 6th International Conference on Spoken Language Processing* (2000), vol. 1, pp. 138–141.

[75] Paul, R., Orlovski, S. M., Marcinko, H. C., and Volkmar, F. Conversational Behaviors in Youth with High-functioning ASD and Asperger Syndrome. *Journal of Autism and Developmental Disorders 39*, 1 (2009), 115–125.

[76] Payton, K. L., Uchanski, R. M., and Braida, L. D. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and

impaired hearing. *The Journal of the Acoustical Society of America 95*, 3 (Mar. 1994), 1581–1592.

[77] Rajendran, G., and Mitchell, P. Cognitive theories of autism. *Developmental Review 27*, 2 (June 2007), 224–260.

[78] Reich, D., Putze, F., Heger, D., Ijsselmuiden, J., Stiefelhagen, R., and Schultz, T. A real-time speech command detector for a smart control room. In *InterSpeech* (2011), pp. 2641–2644.

[79] Rice, M. L., Warren, S. F., and Betz, S. K. Language symptoms of developmental language disorders: An overview of autism, Down syndrome, fragile X, specific language impairment, and Williams syndrome. *Applied Psycholinguistics 26*, 01 (2005), 7–27.

[80] Rieks, N. J., Jovanovic, N., Den Akker, R. O., and Nijholt, A. Addressee Identification in Face-to-Face Meetings. In *In 11th Conference of the European Chapter of the ACL (EACL* (2006).

[81] Rutter, D. R., and Durkin, K. Turn-Taking in Mother-Infant Interaction: An Examination of Vocalizations and Gaze. *Developmental Psychology 23*, 1 (1987), 54–61.

[82] Sacks, H., Schlegoff, E. A., and Jefferson, G. A simplest systematic for the organization of turn-taking for conversation. *Language 50*, 4 (Dec. 1974), 696–735.

[83] Segmenting utterances into C-Units, Accessed - 2010.

[84] Scheffler, K., and Young, S. J. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of Human Language Technology* (San Diego CA, 2002), pp. 12–18.

[85] Schober, M. F., and Bloom, J. E. Discourse Cues That Respondents Have Misunderstood Survey Questions. *Discourse Processes A Multidisciplinary Journal 38*, 3 (Nov. 2004), 287–308.

[86] Schröger, E. A neural mechanism for involuntary attention shifts to changes in auditory stimulation. *Journal of Cognitive Neuroscience 8*, 6 (1996), 527–539.

[87] Shriberg, E. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association 31*, 01 (2001), 153–169.

[88] Smith, V. L., and Clark, H. H. On the course of answering questions. *Journal of Memory and Language 32*, 1 (1993), 25–38.

[89] Stenström, A.-B. *An introduction to spoken language interaction.* Longman, 1994.

[90] Stent, A., Huffman, M., and Brennan, S. E. Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication 50*, 3 (Mar. 2008), 163–178.

[91] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences 106*, 26 (June 2009), 10587–10592.

[92] Sutton, R. S., and Barto, A. G. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[93] Sutton, S., Cole, R., de Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, R., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, M., and Cohen, M. Universal speech tools: the CSLU toolkit. In *International Conference on Spoken Language Processing* (Sydney Australia, Nov. 1998), pp. 3221–3224.

[94] Svirsky, M. A., Lane, H., Perkell, J. S., and Wozniak, J. Effects of short-term auditory deprivation on speech production in adult cochlear implant users. *The Journal of the Acoustical Society of America 92*, 3 (Sept. 1992), 1284–1300.

[95] Swerts, M. Filled pauses as markers of discourse structure. *Journal of Pragmatics 30*, 4 (Oct. 1998), 485–496.

[96] ten Bosch, L., Oostdijk, N., and de Ruiter, J. Durational Aspects of Turn-Taking in Spontaneous Face-to-Face and Telephone Dialogues. In *Text, Speech and Dialogue*, P. Sojka, I. Kopecek, and K. Pala, Eds., vol. 3206 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2004, ch. 71, pp. 563–570.

[97] Terken, J., Joris, I., and De Valk, L. Multimodal cues for addressee-hood in triadic communication with a human information retrieval agent. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces* (New York, NY, USA, 2007), ACM, pp. 94–101.

[98] Traum, D. R., and Allen, J. F. Discourse obligations in dialogue processing. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 1994), ACL '94, Association for Computational Linguistics, pp. 1–8.

[99] van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces* (New York, NY, USA, 2005), ACM, pp. 175–182.

[100] Villing, J., Holtelius, C., Larsson, S., Lindström, A., Seward, A., and Aaberg, N. Interruption, Resumption and Domain Switching in In-Vehicle Dialogue. In *GoTAL '08: Proceedings of the 6th international conference on Advances in Natural Language Processing* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 488–499.

[101] Walker, M. A. An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email. *Journal of Aritificial Intelligence Research 12* (2000), 387–416.

[102] Welkowitz, J., Feldstein, S., Finklestein, M., and Aylesworth, L. Changes in vocal intensity as a function of interspeaker influence. *Perceptual and Motor Skills 35* (1972), 715–718.

[103] Wilpon, J. G., and Jacobsen, C. N. A study of speech recognition for children and the elderly. In *ICASSP '96: Proceedings of the Acoustics, Speech, and Signal Processing, 1996* (Washington, DC, USA, 1996), IEEE Computer Society, pp. 349–352.

[104] Winsler, A., Abar, B., Feder, M. A., Schunn, C. D., and Rubio, D. A. Private Speech and Executive Functioning Among High-Functioning Children with Autistic Spectrum Disorders. *Jounal of Autism & Developmental Disorders 37* (2007), 1617–1635.

[105] Winsler, A., and Naglieri, J. Overt and covert verbal problem-solving strategies: Developmental trends in use, awareness, and relations with task performance in children aged 5 to 17. *Child Development 74*, 3 (2003), 695–678.

[106] Xiao, B., Girand, C., and Oviatt, S. Multimodal integration patterns in children. In *International Conference on Spoken Language Processing* (2002), pp. 629–632.

[107] Xiao, B., Lunsford, R., Coulston, R., Wesson, M., and Oviatt, S. Modeling multimodal integration patterns and performance in seniors: toward adaptive processing of individual differences. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (New York, NY, USA, 2003), ACM, pp. 265–272.