Brion van Over, Elizabeth Molina-Markham, Sunny Lie, and Donal Carbaugh

7 Managing interaction with an in-car infotainment system

Abstract: This work investigates trouble in multimodal turn exchange between an in-car infotainment system and human interactants. The trouble is linked to a lack of crystallization of norms surrounding the turn status of non-speech sounds as well as misalignment on culturally constituted and variable indicators of upcoming transition relevance places. Four interactional adaptations employed by users in order to accomplish their goals for the interaction despite the trouble are identified, as well as norms governing user interaction with the system, and cultural premises that inform that interaction. The work concludes with a discussion of considerations for future multimodal system design.

7.1 Introduction

Imagine a conversation wherein you ask a friend if they would like to see a movie. The friend replies “what concert did you want to see?” Your first thought might be, “What? How is that relevant to the question I just posed? Why might they have said that? Did they mishear me? Perhaps they were implying they would rather go to a concert than a movie?” Humans routinely use this kind of answer, one that flouts Grice’s (1975) maxim of relevance to accomplish communicative goals like conversational implicature. However, what are we to make of the following interaction between a human user and an in-car multimodal infotainment system?

Instance 1: Context FM Radio.

```
1   Participant: (Participant touches microphone button)
2   System:      (audible ding)
3                (0.6)
4   Participant: phone ca[ll
5   System:              [which station or channel do you want to
                 hear?
```

Here we see much the same oddity as in the hypothetical conversation between friends; the user asks to make a phone call and the system replies with a question about what radio station they would like to hear. Except unlike our conversation between friends, here, the user is not likely to wonder if the system would prefer to listen to a radio station or might be attempting a conversational implicature. So how, as a user of a system like this, do we make sense of this interaction when the framework we use to interpret similar human speech no longer works? How did we get into this situation to begin with, and what do we do about it now that we are here?

Users of multimodal systems like this are often faced with these kinds of communicative challenges because machines are imperfect interactants and frequently fail to follow basic rules and principles for the governance of communication that human interactants generally follow with each other. This poses a problem for studying these human-machine interactions through the application of many existing theories of social interaction because of their reliance on the assumption of a model interactant who is competent in the culturally distinctive ways humans have developed for communicating with one-another. This model interactant follows particular rules for the organization and ongoing management of conversation that are based on the assumption that humans interact from a set of what some have suggested are universal principles (Sacks 1974; Sidnell 2001; Stivers et al. 2009). One prominent example of a theory that relies on the model interactant is Brown and Levinson's (1987) Politeness Theory.

Politeness Theory posits a Model Person (MP) that consists in a "willful fluent speaker of a natural language, further endowed with two special properties – rationality and face." (p. 58) This MP is "rational" to the extent that they have goals for their interaction, and a process through which the optimal means of achieving these goals are known and pursued. The MP has "face" to the extent that all speakers have wants, "roughly to be unimpeded and the want to be approved of in certain respects" (p. 58).

Brown and Levinson's MP is informed by Grice's (1975) model interactant who follows what he dubs the "cooperative principle." Interactants who obey the cooperative principle act in accordance with the following rule: "make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." Grice goes on to specify four conversational maxims that he suggests govern conversation that occurs under the cooperative principle, including the maxim of quantity (give the most helpful amount of information), the maxim of quality (do not lie), the maxim of relation (make your contributions relevant), and the maxim of manner (make your contributions clear, brief and orderly). Of course, Grice does not claim that all speakers follow these rules. In fact, Grice suggests these maxims can be violated where the rule is simply ignored, or they can be flouted, where the rule is broken for a sought conversational effect.

The relevance of these theories to the present discussion consists in noticing that machines in interaction with humans routinely violate behaviors expected of the Model Person, or the interactant operating under the cooperative principle. This is because machines do not currently exhibit the kind of rationality presumed by Brown and Levinson, nor share the “face” concerns of human interactants. They may seek to operate within the bounds of the cooperative principle, if programmed to attempt to do so, but may violate maxims in obvious ways similar to how a human interactant might flout a maxim in order to accomplish a conversational implicature (Grice 1975), though the machine has no intention of doing so and users likely know that.

This means that some of our fundamental assumptions about social interaction become unreliable in human-machine interaction. And by extension, machines themselves may be found to be unreliable interactants for these very reasons. Two questions then become essential for us to pursue in order to better understand the dynamics at work in human-machine multimodal interactions. (1) How do humans interact with machines that are not assumed to be, nor able to operate as, fully culturally competent interlocutors? and (2) how do we manage moments when things inevitably go wrong in these interactions?

One area designers of such systems have often overlooked in their attempts to create ever more human-like interactants is the structure of turn exchange in conversation between humans and machines, with an eye toward the distinctive ways turn exchange may be managed across cultures. Even less well understood, is how human-machine turn exchange is accomplished in interactional contexts where multiple potential communicative modalities are at play. Much more attention has been paid to how systems use sound (Brewster 1998; Rinott 2008), recognize and produce speech that invokes human emotion (Busso et al. 2004; Cahn 1990; Mor 2014; Oakley et al. 2000), and more recently, operate in a multimodal capacity (see Dumas et al. 2009; or Wechsung 2014, for a review).

This piece seeks to address this gap through an analysis of trouble in turn exchange in human-computer multimodal interaction in an in-car infotainment system. One central question we examine is how, in these interactions, where participant expectations of their interlocutor’s competence as a model interactant may not hold, does repair get done, and what does the trouble and its repair tell us about the culturally distinctive ways to do it “right”?

In what follows, we review the concepts and theoretical framework employed in the analysis and research design that produced the data set we analyze here, though the theory and methodology as adapted for the study of in-car communication is more fully detailed elsewhere (Carbaugh et al. 2012).

Next, we analyze a number of instances to determine the source of trouble experienced by many participants in interaction with the system, and the variety of methods users employed for accomplishing their goals despite this trouble. Based on this analysis, we highlight cultural norms and premises governing interaction in this communication situation (Hymes 1972). We conclude with a discussion of the implications for multimodal system design.

7.2 Theoretical framework and related literature

In the tradition of Conversation Analysis (see Heritage (2010), for a summary of principles), it has long been accepted that conversation occurs in a sequential fashion, organized through the managed exchange of turns at talk, though debate exists over seemingly contradictory cases (Reisman 1974; Sidnell 2001). Generally, this organization is taken to be fundamental to meaning-making in interaction. Schegloff (2000) captures this stance in the following:

> The orderly distribution of opportunities to participate in social interaction is one of the most fundamental preconditions for viable social organization. ... One feature that underlies the orderly distribution of opportunities to participate in conversation, and of virtually all forms of talk-in-interaction that have been subjected to disciplined empirical investigation, is a turn-taking organization. The absence of such an organization would subvert the possibility of stable trajectories of action and responsive action through which goal-oriented projects can be launched and pursued through talk in interaction ... (p. 1)

The preponderance of articles on the topic of turn-based organization preclude a thorough review here, but for the seminal work of Sacks, Schegloff and Jefferson (1974). Therein, the authors propose a Turn Constructional Unit (TCU) and a Turn Allocation Component. The idea of the TCU suggests that interactants' turns are constructed in such a way as to make the kind of turn it is, the action the turn seeks to accomplish, available to fellow interlocutors such that the projection of the coming end of the turn can be anticipated. That such a function exists in conversation is evidenced by the ability of interlocutors to cut-in before a turn is fully completed, having projected what the completed utterance may likely have contained. The Turn Allocation Component suggests that interlocutors actively manage the exchange and allocation of turns at talk through a variety of practices that are designed to select a next speaker, or self-select as next speaker, and signal when a speaker's turn is completed, or about to be, through a transition-relevance place (TRP). Schegloff (1992) conceptualizes TRPs as "discrete places in the developing course of a speaker's talk (...) at which ending the turn or continuing it, transfer of the turn or its retention become relevant" (p. 116).

The sequential organization of speaking turns also provides the foundation for the interpretation of the meaning of talk in interaction. For instance, returning to our hypothetical conversation between friends in the introduction, if "what concert do you want to see?" is the first turn in a conversation, its meaning may be heard as an invitation to a concert. However, following a prior turn where a friend asks “do you want to see a movie?” that same speech may now be interpreted as indication of a mishearing of the prior turn, or a rejection of the invitation to a movie and a counter-invitation to a concert. Since meaning is reliant on the position of an utterance in relation to surrounding utterances, the timing of conversational turns becomes significant in managing the mutual intelligibility of the interaction, without which the interaction cannot continue without repair. How humans and machines in interaction manage this exchange of turns, and the careful timing required to do so in order to assure mutual intelligibility is of primary concern to this work.

One way interactants can signal that a turn is complete is the use of pause (Maynard 1989). An interlocutor may use a pause at the end of an utterance to invite another speaker’s participation, but they may also simply be pausing for breath, to develop their next utterance, because they were distracted, etc. The trouble, then, is determining whether a gap in speech is an invitation to exchange speaking turns, or merely a period of silence where the speaker intends to maintain the floor. Among other strategies, like audible in-breaths, or other disfluencies (Corley & Stewart 2008) such as “umm” or “annnd,” culturally competent interlocutors come to know the length of time an interactant might pause to signal a TRP and actively monitor for these in conversation, though syntax (Sacks 1974), prosody (Couper-Kuhlen & Selting 1996), pointing (Mondada 2007) and other pragmatic information (Ford & Thompson 1996) are also potential cues of the coming completion of a turn.

Indeed, in instances of intercultural interaction differences in the use of silence can often create trouble as speakers evaluate the meaning of silence from a cultural vantage. Carbaugh (2005), documents a moment of such trouble in an introductory meeting with a future colleague in Finland, Jussi Virtanen. To Carbaugh’s surprise, Virtanen would respond to each of Carbaugh’s turns at talk with a 10–20 second pause. From the vantage of an American English speaker from the Northeast, these pauses were exceptionally long and, for Carbaugh, signaled something untoward in the interaction. As he later discovered, the pauses were the result of a confluence of factors including a Finnish customary practice of long (from the American view) pauses after sentences, Virtanen’s personal use of longer pauses (from the Finnish view), Virtanen’s careful use of English as a second language, and finally the use of long pauses as a means of signifying the respect one has for the occasion and its significance. Here, then, inter-turn pause is both the result of situational factors, but also a motivated use of cultural means for communicating respect and appreciation.

Scollon and Scollon (1981) note trouble with culturally distinctive pause lengths in conversation in their study of interaction between Athabaskan-English speakers and US American-English speakers. The authors find that in conversation, Athabaskans are often overrun by English speakers because of a preference for pausing between the exchange of speaking turns for about a half second longer than typical for US English speakers. This means that English speakers often set the topic of conversation, and then proceed to dominate (from the Athabaskan view) the remainder of the conversation as Athabaskans monitor for TRP's at the end of the English speaker's turn, only to find that the English speaker has started speaking again before they had a chance to take their turn. This leads to negative evaluations of the conversation from both the US English and Athabaskan-English speaker's view, based primarily on cultural variation in inter-turn pause length and the meaning of pauses that last relatively longer or shorter.

Because turn-taking practices serve as the foundation for mutual intelligibility in conversation, and these practices are subject to cultural variation (Tannen 2012), understanding the cultural norms and premises governing the exchange of turns at talk is essential. Attending to issues of variation in norms for turn taking can help illuminate trouble in human-computer interaction in the same way such an analysis can illuminate trouble between members of different cultures. As a result, in the analysis that follows, we make use of concepts from Conversation Analysis, reviewed above, as well as a framework for analyzing the culturally distinctive ways talk is patterned in interaction – Cultural Discourse Analysis (Berry 2009; Carbaugh 1988, 2007, 2012; Scollo 2011).

Cultural Discourse Analysis (CuDA) is a development of the Ethnography of Communication (Hymes 1962, 1972, 1974), that seeks to describe, interpret, compare and critique culturally patterned communication practices. It does so first through the conceptualization of communicative phenomena as a communication "act", "event" or "situation". Communication "acts" are the smallest unit of analysis and may include a single utterance or turn at talk. Communication "events" are bound by a clear initial act and closing act, and contain a structured sequence of acts that are "directly governed by rules or norms for the use of speech" (Hymes 1974, p. 52). The communication "situation" is not bound by particular communication acts as in the communication event, but rather by other boundaries including spatial, such that we might conceive of communication "on the front porch," or "in the corner bar" (Philipsen 1992) as bounded situations that inform the kind of communication in that place. Thus, here, we conceptualize communication in the car as a communication situation, containing a number of communication events, and made up of a number of communication acts.

So conceived, CuDA invites us to investigate these communication acts, events, and situations for "radiants of meaning" found in messages about personhood, social relations, emotion, place, and communication itself. Presumed and enacted in these messages are "cultural premises" which serve as a resource for the interpretation and production of meaning in interaction. Cultural norms may also be identified, which are implicit or explicit rules that govern the moral domain of social action. In our analysis, we employ these concepts in understanding the distinctive premises and norms that both system designers, and users of systems, may employ as they seek to accomplish their respective goals for the interaction, as well as the ways they may sometimes be misaligned, the consequences of that misalignment, and what users do to get back on track.

7.3 Methodology

As discussed in previous work (Carbaugh et al. 2012, 2013; Molina-Markham et al. 2014, 2015) data for the analysis below were collected from the driving sessions of 26 (14 female, 12 male) participants during the study of an in-car infotainment system conducted in Western Massachusetts. During the driving session, participants would use their own car, which had been outfitted with a prototype infotainment system that they would interact with through a dashboard mounted tablet computer. Participants were asked to drive on mostly rural roads of their choosing for one and a half to two hours on average. During this time participants were invited to make use of a variety of the voice capabilities of the system as they would in the normal operation of their own vehicle were they to have such a system. Touch interaction with the system was permitted for starting or ending interactions through use of the microphone button, which started a voice command event, or the “end” button, which could be used to close an ongoing action.

Two researchers were present in the vehicle at all times, one in the passenger seat who made observations and conducted interviews at predetermined stop intervals, and who was ready to take control of the vehicle in case of an emergency. The other researcher was seated in the back seat of the vehicle serving the role of “Wizard,” using a laptop to interact with the system to fulfill user directives. Since some of the voice interaction functionality we wanted to test is not yet in production, the human Wizard served the role of the “brain” of the system, interpreting participant’s speech and executing directives they had made to the system. We have found no evidence that participants realized the researcher in the back was operating the system.

At the beginning of the driving session participants were instructed to explore the system’s multimodal abilities through both touch and voice interaction. After participants felt comfortable using the system, we proceeded to an off-road test where participants drove around a parking lot and further explored voice interaction with the system. Following this we proceeded to an on-road test where participants were instructed to do their best to ignore our presence and use the system as they normally would, which participants generally seemed able to do. At the midpoint of the drive, we conducted a short interview to hear about their experience of the system, answer questions, and suggest functionality they may not have explored or been aware was possible. A similar interview was conducted after the driving session had completed.

Being interested in the sequential organization of multimodal interaction in this context, the way users managed turn exchange with the system, and any potential misalignment between the system's behavior and users' cultural norms and premises operating in this communication situation, we identified interactional sequences where the system's talk overlapped with the user as a sign of potential turn exchange trouble. We found that when users engage in a task-switching event (asking the system to do a task that is not part of the current task the system is performing) a disproportionate amount of overlapping talk occurred relative to other interactional sequences, like directing the system to perform a new task. Thus, the data for this analysis are taken from overlapping talk that occurred during user attempts to switch tasks. Not all users made use of the task switching functionality of the system, which produced a corpus of 7 participants that did, each of whom experienced some degree of overlapping talk with the system on their first use of the task-switch event, making this a regular and robust phenomena for investigation.

7.4 Prompt timing and misalignment – A formula for interruptions

When the system was engaged in the ongoing performance of a given task (playing the radio, making a phone call) and the user pressed the microphone button to initiate a voice command, the system was programmed to respond by providing an audible "ding" sound to confirm that the system received the users request to initiate a voice command and was now in an "on" state. The system would then play a task-relevant prompt. We can already see at play a number of interactional modalities users are negotiating in the opening of this interaction, including touch, audible non-speech in the form of the "ding", and speech from both the system and user, making this a rich and complex interactional context.

For instance, while in the radio task, if the microphone button is pressed, the system would ding and then ask "which station or channel do you want to hear?" It was then the user's turn to talk. However, the sequence never went according to design in the first instance. Below is an example of a typical way this interaction occurred.

Instance 1: Context FM Radio (Participant 11–51:44 minutes into the session).

```
1  P:  (participant touches microphone button)
2  S:  (audible ding)
3      (0.6)
4  P:  phone ca[ll
5  S:          [which station or channel do you want to hear?
6  P:  phone call
```

In this instance, the participant presses the microphone button while listening to an FM radio station. The system responds to the participant's touch with an audible ding. There is then 0.6 seconds of silence before the participant begins her directive to the system – "phone call." During this directive the system overlaps her talk with its own prompt "which station or channel do you want to hear?" The participant responds by restating her directive from the prior turn, "phone call," on line 6, which can be understood as a corrective action since the system's turn was not responsive to her command "phone call". In this way, the system violates Grice's maxim of relevance, issuing an utterance that is not sequentially relevant to the participant's prior turn. In this instance, the user opts to treat the system's violation as the result of a mishearing, and reissues her command.

Why would the system respond in a way that is so badly unresponsive to the participant's directive? At least part of this misalignment, we suggest, is the result of differing meanings for the audible ding the system plays on line 2, and implicates the difficulty of designing multimodal interfaces wherein the meaning or function of certain modes (an audible ding) is not well established. The system is designed to play a ding, followed by a task-relevant prompt, "which station or channel do you want to hear?" The ding, then, from the system's design functions as an acknowledgement of the participant's request to initiate a voice command, a "wake up" chime. The system, however, is designed to take a turn following this ding, verbally prompting the user to provide information relevant to the ongoing task at hand. It plays this verbal prompt, in this instance, 1.5 seconds after the audible ding. The system never manages to play the verbal prompt faster than 1.3 seconds after the audible ding, leaving a sizeable pause between ding and prompt.

Because the system is designed to take the first turn, it is not listening for the participant's voice between the audible ding and its first turn, the verbal prompt. Therefore it cannot move to cut-off its turn in recognition of the participant-issued directive as a human interlocutor might (Scheglof 2000). Because the participant waits only 0.6 seconds before beginning their turn, they issue a command that is not heard by the system, and which causes the system's turn (its first turn from the system's view) to be badly non-responsive to the sequential position of the interaction at that juncture.

Now that we understand what the system thinks is happening we might ask, why does the participant take their turn at 0.6 seconds? One possible explanation is misalignment between the system and participant on the meaning of the audible ding on line 2. The participant may understand this ding in a number of ways. We first suggest that the participant might understand the ding as a summons response, borrowing from the interactional form of the telephone conversation.

In 1964, Sacks (Jefferson & Scheglof 1995) pioneered studies of telephone call interactions and concluded that the ringing telephone functions as a summons and the answering of the phone with "hello" functions as a response to the summons (Sacks 1974), ostensibly two turns at talk have been exchanged. The next turn, then, wherein the topic of the conversation is set, belongs to the actor who did the summoning, the caller. In this case, the participant issues the summons through touching the microphone button, and the system responds with an audible ding. If this interaction were following the routine form of the telephone call then the participant would take the next turn and set the topic. Instead, here, the system attempts to set the topic by asking what station the participant wants to hear. It is possible, then, that participants are modeling interaction with the system after the routine form of the telephone call, wherein the participant takes the first turn at talk, and that this understanding informs their move to initiate their directive before the system's verbal prompt, since they do not expect the system to be taking a turn in this position.

Also at work are cultural norms for the amount of time that passes in a gap between turns before that gap signifies a Transition Relevance Place (TRP) where a conversational turn may be understood to be over or relinquished. If a cultural norm exists for the participant that turns are exchanged after roughly 0.6 seconds of silence, then the system will routinely take too long to take its first turn as users proceed to interpret the system's silence as yielding the speaking floor. As we see in the following instances this appears to be the case.

Instance 2: Context My Music (P8 – 28:06).

```
1  P:  (Participant touches microphone button)
2  S:  (audible ding)
3      (1.0)
4  P:  next
5  S:  what artist would you like?
6  P:  next
```

In this instance, like the last, the user is in a task, in this case listening to their downloaded music library, when she decides to touch the microphone button. After doing so the system dings, and after a 1 second gap she issues her directive "next." The system's next turn, which sequentially would be heard as a reply to the participant's directive, asks the participant what artist she would like to hear. This verbal prompt is of course not responsive to the participant's directive, and so the participant restates it on line 6, again treating the system's turn as the result of a mishearing in need of correction. The amount of time it takes the system to respond to the microphone button press was measured at 2 seconds, this is the longest the system takes to issue a prompt. This allows the participant to wait 1 second and then issue her directive in the remaining 1 second before the system plays its prompt. Because of this the user does not experience overlapping talk with the system, but is perhaps presented with an even more confusing response, since the overlap itself can function to let the participant know that something is wrong. Without the benefit of the overlap, the user is left to wonder what the system's turn means and what should be done next?

This 1 second pause, like the 0.6 second pause in the prior instance, is long enough to indicate to the participant that the system has yielded the floor, and it is now her turn. This is not the case however, from the system's design, and the system proceeds to issue what it takes to be its first turn, leaving the participant to conclude that the system has either not heard, or misheard her command. This may negatively impact participant perception of the competence of the system as a voice interaction partner. We can see the persistence of this pattern in Instance 3 below.

Instance 3: Context XM Radio (P12 – 23:13).

```
1  P:  (Participant touches microphone button)
2  S:  (audible ding)
3      (1.5)
4  P:  ca[ll
5  S:    [what xm channel?
6      (1.7)
7  P:  call tom
```

In this instance, the participant is listening to an XM radio station when she presses the microphone button. The system responds to the press with an audible ding, at which point the participant waits 1.5 seconds before beginning to issue her directive "call." However, as she begins to issue this directive, the system overlaps her speech with its own question, "what xm channel?" After this, the participant waits 1.7 seconds and then restates the directive she appeared to be beginning on line 4.

The participant in this instance treats 1.5 seconds as sufficient time to signify a TRP, inviting her to take her turn. However, here, the participant has waited long enough that the system begins its prompt almost simultaneously with the start of her directive. Unlike the prior instances the participant here abandons her turn, surrendering the floor to the system in an overlap resolved after one beat (Scheglof 2010). Overlaps which do not end after one beat may be the beginning of an indication of "competitive production" wherein two or more interlocutors vie for the floor in a competitive move. One can imagine that the ideal design of the system would not be such as to enter into competition with its users for turns at talk, being more preferably oriented to user-satisfaction and compliance. However, here, the system produces a turn wherein the user is forced to either competitively co-produce talk until the system completes its turn, or abandon their turn thereby deferring to the system's "right to speak". This is open to being heard by participants as a dominant interactional move, which is likely not a desirable position for the system and user.

Across these last three instances, participants encountered the same interactional trouble, attempting to initiate a directive to the system that the system responds to with a sequentially non-responsive verbal prompt, and/or in most cases, the participant's talk is overlapped by the system's forcing the participant to compete for the floor or abandon the turn.

Participants in the larger corpus from which these instances are taken varied in the amount of time they waited before speaking after the audible ding from the system from 0.6 seconds to 1.5 seconds. Since the average time the system takes to generate a prompt following the audible ding is 1.7 seconds, this means those participants who wait around 1.5 seconds to give the system a directive will almost certainly be interrupted by the system, while those who begin a directive immediately following the ding may be able to complete their directive utterance, only to be met with a question that seems irrelevant to the directive they have issued. A seemingly simple fix for this trouble is an anti-overlap feature to assure that if the system hears the user talking, it hold its turn until the system can decide what next action to take that would be relevant to the user's speech. However, listening for user speech all the time when it has no reasonable expectation that the user is about to speak, like after a TRP, means lots of mistaken "hearings" on the system's part that could lead to even more trouble.

The patterning of the interaction above is suggestive of a norm for the management of turn exchange in conversational positions where turn allocation is ambiguous. This norm treats pauses of longer than 0.6 seconds, and no longer than 1.5 seconds, to be indicative of the passing of a turn. The system's routine pause length of 1.3–2 seconds, then creates a misalignment in the turn-taking management of interaction between the participant and system. As a result, participants are forced to abandon their turn, or competitively produce a turn in overlap with the system. Participants must further make sense of the system's turn, which given its late positioning in the interaction relative to the position it was designed to inhabit, appears non-responsive to the participant's directive. In the instances above, participants treated the system's turn as a mishearing in need of correction through repetition of the initial directive, though this was not the only way participants managed to negotiate the difficulty of this misalignment. After having encountered this misalignment some number of times participants would generally adjust their interaction with the system in one of four ways, which we review in the following section.

7.5 Interactional adaptation

After a participant experienced the system overlapping their directive, and/or responding to their directive in non-responsive ways, they appeared to learn, at different rates, that the system will be taking a turn after the audible ding in task-switch events, and that this turn will take place after some notable pause. Given their apparent noticing that this is the case, users proceeded in future interactions with the system in one of four ways.

(1) One way participants adapted to the system was to sustain a "competitive production" (Schegloff 2000). In the instance below this is accomplished through the extra-ordinary elongation of the vowel sound in "too," sustained until after the system's turn had completed.

Instance 4: Context FM Radio (P10 – 1:24:33).

```
1  P:  (Participant touches microphone button)
2  S:  (audible ding)
3      (0.8)
4  P:  change st[ation too:::::::::::::::::::::::::::::::owuh (1.2)
5  S:           [what radio station do you want to hear?
6  P:  ninety seven point three
7      (4.5)
8  S:  could you repeat that please
```

In this case, the participant, unlike the participant in instance 3, refuses to abandon their turn to the system and makes a bid to hold the floor through the elongation of the vowel sound in "too" on line 4. A fellow human interactant would then be forced to choose to continue their own turn in a sustained overlap, or yield the floor to the other speaker. Because the system is not listening when it's playing its own prompt, it is incapable of knowing that the participant is speaking and therefore incapable of deciding to abandon its turn. This means the system, in instances of competitive production, will always sustain overlap until its turn is complete. Somewhat ironically, however, the system can never “win” since human interactants engaged in competitive production can project the incipient end of a turn shape and adjust their strategy for elongating their turn to assure it lasts longer than the system’s. This is the case in the instance above.

Despite the participant “winning” the competitive production, the directive to tune to 97.3 cannot be understood by the system because it was not listening to the participant’s utterance during the overlap. Even if the system had been listening it would not be able to understand a directive including the extraordinary stretched vowel seen here. A bitter-sweet victory. The implications for the outcome of this competition among human interactants would likely include messages about the status of the relationship between interactants. As Tannen (1993) points out, however, the meaning of this overlap to participants is not set *a priori* as Schegloff’s (2000) use of the term “competitive production” might suggest. Overlap may also be understood by interactants as a move to solidarity, though it does appear in this instance that the participant intends to outlast the system. Regardless, the possible interpretations of the meaning of the overlap to participants, one thing is certain, the system will take no implications about social relations from the interaction, though the participant may.

This is one way participants have borrowed interactional strategies (competitive production) from human interaction for use in dealing with an invasive conversational partner (the system) but where the social effects of which may not carry over. This seems to not discourage their use here however.

(2) The second way users adapted to the system’s overlap was to wait out the long pause for the prompt and then speak. In this strategy, participants allow for a longer pause than they generally had in the past, giving the system the opportunity to play its prompt. The participant would then give their directive, which was necessarily shaped as a corrective, since quite often the directive they gave the system was not related to the task-specific prompt the system played.

Instance 5: Context My Music (P8 – 31:08).

```
P:  (Participant touches microphone button)
S:  (audible ding)
    (1.3)
S:  What artist would you like?
P:  FM Radio
S:  Just a second
```

Here the user touches the microphone button and the system dings in reply, 1.3 seconds pass, and the system asks the participant what artist she would like to hear. The participant, apparently not wanting to hear an artist replies “FM

Radio". Structurally, the exchange of turns has gone smoothly here (no overlap) though there are two things to note. First, this participant experienced trouble with a turn exchange performing a task switch 4 minutes prior (Instance 2), experiencing the overlap phenomenon common among all users. As a result, we suggest that her decision to not take a turn during the 1.3 seconds gap after the system's "ding" is an adaptation to the overlap trouble from her prior task-switch experience. This, then, is another way users have developed to deal with turn-exchange difficulty, as ultimately the participant has goals for the interaction she would like accomplished and needs to find a way to get back on track in order to do so. In this case, this is accomplished through the participant's adaptation to the system's norm for a 1.3–2 second pause before its first turn. This participant has then moved through a process to identity the trouble (the system intends to take a turn at talk after the ding, and has a relatively long pause before it does so), develop a possible solution (wait until the system speaks) and implement that solution, though it contradicts her and other participants' routine norm for managing turn exchange (a 0.6–1.5 second pause).

Despite having adjusted the timing of her initial turn to accommodate the system, she is still placed in the position of having to respond to the system's prompt with a move to reject the system's offer. Since the system opts to take a guess at what the participant might want, asking "what artist would you like?" the system constructs its turn as to prefer a response that chooses an artist. Any response from the participant that is not the name of an artist is thereby shaped as a dispreferred response (Heritage 1983; Levinson 1983; Pomerantz 1984). This restricts the available next actions for the participant to either a response to the question that selects an artist, or an outright rejection of the system's offer to play an artist, which people would generally rather not have to do.

Another participant adopted this same strategy, also 5 minutes after experiencing an overlap with the system during a task-switch event.

Instance 6: Context XM Radio (P12 – 28:45).

```
P:  (Participant touches microphone button)
S:  (audible ding)
    (1.3)
S:  what XM channel?
P:  bridge
    (7.5)
S:  hold on
```

Given the proximity of this instance to the participant's prior system overlap, it is likely here that the user has adjusted her turn timing to accommodate what she now knows to be the system's long pause following the audible ding. As it happens in this sequence, allowing the system to take its turn after the extended pause provides the system a chance to guess that the user might want another XM channel. Often this guess is wrong, as in the last instance, which leads to the necessity of a participant's rejection of the system's guess, but in this case the guess is right as the user does not wish to switch tasks, only channels. This is, then, the best case scenario, though it requires a deviation from the user's established interactional norms for pause that indicates a TRP in order to accomplish.

(3) Another option participants developed for dealing with the overlap involved an abandonment of the use of the task switch capacity. In this sequence, participants opted to switch tasks by first touching the "End" button to stop the current task (playing the radio), and then pressed the microphone button to initiate a new voice command. When the microphone button is pressed outside of an ongoing task, like when the user is on the system's home screen, the system plays an audible ding and then waits for the participant's command. Possible overlap is then avoided by selecting an interactional path that does not include the system taking a spoken turn. One benefit, then, of multimodal systems is the ability of the user to adapt to verbal interactional trouble by employing alternate modes that avoid the trouble.

Instance 7: Context XM Radio (P9 – 53:05).

```
1   P:  (Participant touches End Radio button)
2       (2.0)
3   S:  (Radio stops playing, screen shifts to home)
4       (.5)
5   P:  (hand begins move toward radio)
6   P:  (.7)
7   P:  (finger touches mic button)
8   P:  gimme dubbelyu=efem (.) ehhhn give me doubelyu:::::
        whatsitcalled effseear
```

Here the user begins by ending an ongoing task, the playing of the radio by touching the End Radio button. It takes the system 2 seconds to comply with the user's directive to stop playing the radio and return to the home screen on the display. Within 0.5 seconds the user begins to move his hand back toward the radio and makes contact with the microphone button 0.7 seconds later. He then gives the radio a directive to play WFCR, all in less time than it took the system to comply with his initial directive to end the radio.

It is likely, then, that the user intended to change radio stations when he hit the end radio button, but why not just press the microphone button and tell the system to change stations, making use of the system's task-switch function? The answer we propose that best accounts for the participant's actions here is that in prior interactions the user had difficulty with the turn exchange, particularly, during a task switch event 40 minutes prior, whereafter he ceased to use the task-switch functionality, opting instead to explicitly end all ongoing tasks through touch before initiating a voice interaction to issue a new command. Abandoning the line of interaction that produced the turn-exchange difficulty is then one method a user developed for accomplishing the task sought, despite trouble with the timing of turn exchanges in the task-switch event. Doing so is likely not the ideal case however, as the task switch function allows users to achieve their goal in as little as one button press and one voice command, while the strategy adopted by this participant will require a minimum of two button presses and a voice command. This is not ideal from the perspective of system designers either, since the minimization of the use of touch while driving is preferred for safety reasons.

(4) Not all users did develop new methods for dealing with the overlap trouble. One participant continued to repeat the pattern observed in instances 1–3 (issue command, system overlaps, reissue command) 6 times repeatedly, one after the other, over the course of her drive with the first occurrence at minute 20 and the last 42 minutes later. During this time the participant never adjusted the pattern of their interaction, continually experiencing overlap with the system each time she performed a task switch. We have included one of these instances here for illustration, though the patterning is identical to instances 1–3 reviewed above. The following is taken from the fourth recurrence of this pattern with this participant.

Instance 8: Context FM Radio (P11 – 54:19).

```
P:  (Participant touches microphone button)
S:  (audible ding)
P:  phone call
S:  which station or channel do you want to hear
P:  phone call↑
S:  okay (1) who would you like to call
```

That this participant persists across a number of instances to issue a directive prior to the system's turn is likely the result of the participant never identifying that the system intends to take the first turn at verbal interaction and as a result is not listening as it prepares its turn. Instead, the participant through the repetition of her initial directive, treats the system's prompt as a mishearing of her initial directive in need of repetition. If the participant had identified that the system was not listening, persisting with issuing the command before the system's prompt would serve no purpose and would likely have discontinued. This is suggestive that not all participants have equal access to the resources required to interpret the source of trouble with the system's behavior. This instance further highlights the trouble of poorly crystalized norms surrounding the meaning and turn status of the audible "ding" in multimodal interaction.

It is also possible that if the participant was not able to ultimately accomplish the task she sought as a result of this trouble then she may have proceeded to explore other strategies. However, in this instance, the participant is ultimately able to get the system to follow her directive on line 6 when the system acknowledges her directive to make a phone call asking "who would you like to call." This participant then appears willing to accept some trouble so long as the task is accomplished in the end.

7.6 Norms and premises

The analysis above suggests some normative ways that participants approach interacting with the system, as well as certain premises that inform this use. In the instances presented above, each participant experienced either an overlap of talk with the system and/or a seemingly non-responsive reply to their directive. The regularity with which this phenomenon occurred throughout the corpus suggests that the amount of time these participants understand as evidence of, or opportunity for, a turn exchange, in contexts where next speaker is ambiguous, is less than the 1.7 second average time the system takes to produce its verbal prompt. We believe this amount of time to be a cultural norm for managing the interactional exchange of conversational turns when the next speaker is ambiguous. The system, then, is engaged in a kind of norm violation when it produces its overlapping prompt that carries the interactional force of an interruption, violating the moral order of turn-taking and politeness that is generally expected between human interactants in social interaction. This norm can be more explicitly formulated as: *In contexts where next speaker is ambiguous, if an interactant wishes to take a turn, they should do so between 0.6 and 1.3 seconds after the prior action, in order to be a proper interactant.*

Misalignment between the system and participants was not restricted solely to the normative timing of the exchange of speaking turns, but also in the meaning of particular actions within a communication event, as in the case of the audible ding. Whereas the system was designed with the audible ding's intended meaning being an alert of the system's status, akin to announcing the system is on, participants treated the audible ding as a summons response akin to the organization of telephone calls. Depending on which meaning of the audible ding one employed, a different next speaker would be appropriate. From the participants' vantage the audible ding occupied the space of an interactional turn, and therefore the system was understood to have passed the turn back to the participant for their first spoken turn.

In all but one case, participants chose to adjust their interactional strategies for accomplishing the task they sought, with one user persisting in the original pattern of overlap, likely doing so as the result of failing to identify that the system was not listening in the gap after the audible ding. This means that all participants who became aware of the source of the trouble opted to make adjustments in order to accomplish the task.

It is not automatically the case that this should be so. In human interaction, an interlocutor behaving in the way the system does would likely be called to account for both their repeated interruptions, but also for violations of the maxim of relevance for no apparent conversational purpose. However, the participants in the instances collected above never call the system to account for its behavior, nor exhibit any animus toward the system for what might be cause for an argument with a human interactant. The system, in effect, gets a pass. This is not to say participants will, or do, find interacting with a system under these conditions pleasing, only that the system itself appears not to be held responsible for its behavior in this context. This is likely because participants understand that the system lacks the fully fledged capabilities of a culturally competent human interactant, and therefore cannot be held responsible for these sorts of issues, but likely not trusted either.

A premise of and for communication can then be identified in the participants' interaction that *those who are not fully competent interactional partners cannot be held responsible for certain interactional blunders.* An accompanying premise of personhood can then also be formulated as *voice interactive machines are not fully competent interactional partners.* And finally, an additional norm can be identified for proper behavior given the above premises, *since machines are not fully competent interactional partners, human users ought to adjust to the system in order to accomplish their goals.* These premises likely inform the level of tolerance users have for interacting with systems that routinely violate human interactional norms, without which interaction with systems at this level of capability would not be possible. This does not mean that the above premises are universal or automatic, as one can imagine alternate premises that systems such as these can be held responsible for interactional blunders, such that repeated violations of the interactional order result in discontinued use of the system. This did not, however, appear to be the case in participant interaction with the system during task-switch events in this corpus. It is unclear whether the research context in which these data were collected resulted in more persistent attempts to continue using the system than might have occurred had the user been alone in their own vehicle.

7.7 Implications for design

The analysis above can be used to make particular recommendations for the improvement of multimodal interactive systems in the future. First, the task-relevant prompt is problematic as many participants noted in interviews that it was unnecessary, inappropriate, or too long. Some participants suggested that no prompt was needed at this stage of the interaction at all, citing that when someone presses the microphone button while in a task they likely have something they would like to do in mind, and have pressed the microphone button in order to give the system that command. As a result, the system need not offer any prompt, but rather just listen for the participant's command.

Second, system designers need better understand the role of non-speech sounds in multimodal interaction and their turn taking relation to other modalities such as touch and speech. In the design of this system, the audible ding is treated as if it occupies no conversational position – it takes no turn. This is clearly not how participants in the above interactions understand the ding. A turn-based analysis of the interaction suggests that the ding does function as a turn-at-talk, with the first turn being the participant's touch of the microphone button, the second turn being the system's reply to the touch through audible ding, and the third turn then passing back to the user for first topic. However, because the system design does not account for the audible ding as an interactional turn, it presumes users will wait for the system to respond to the microphone press with a verbal prompt. This appears to not be the case as users hear the audible ding as the response to the microphone press and proceed to take their turn. Some research on the role of non-speech sounds in human-computer interaction is already underway (Brewster 1997; Brewster 2002; Hereford & Winn 1994), but does not incorporate an analysis of the sequential position of this mode in the organization of interaction.

Some participants did report a desire to have a system prompt after the microphone button was pressed as a sign that the system is "listening," but thought the prompt that was offered was simply too long for regular use. Participants suggested alternate prompts including "yes?" or "what would you like?" which are likely better alternatives as they are task independent and do not require users to reject the system's wrong guess, which as indicated above is a dispreferred action in conversation.

Ultimately, however, what holds multimodal interactive systems back the most in the instances analyzed above is the system's inability to listen for user speech and act accordingly. Cases of overlap in human interaction are resolved in a variety of ways (Schegloff 2000) but all require monitoring of the ongoing turn by both interactants. In order to properly model human interaction, the system must be able to listen to users' ongoing turns and adapt, as we do with them. Research on the broader phenomena of overlapping speech in HCI, sometimes referred to as "barge-in," is also underway, examining the frequency and context of "barge-in" cross-culturally (Wang, Winter & Grost 2015)

We further advocate attention be paid to the cultural nature of the management of turn-exchange both in the amount of time interlocutors normatively wait as indication of a TRP, but also in the practices employed managing turn exchange, and the strategies adopted to accomplish interactants' goals. The analysis above suggests two cultural norms and two premises of and for communication and personhood that may vary culturally and influence the way users interact with these kinds of systems, particularly surrounding the resolution of trouble and the meaning of that trouble.

Abbreviations

FM	(radio)
MP	Modal Person
TCU	Turn Constructional Unit
TRP	Transition-relevance Place
CuDA	Cultural Discourse Analysis
XM	(radio)
WFCR	(radio station)
HCI	Human Computer Interaction

References

Berry, M 2009, 'The social and cultural realization of diversity: An interview with Donal Carbaugh', *Language and Intercultural Communication*, vol. 9, pp. 230–241.

Bossemeyer, RW & Schwab, EC 1990, 'Automated alternate billing services at Ameritech', *Journal of the American Voice I/O Society*, vol. 7, pp. 47–53.

Brewster, SA 1997, 'Using non-speech sound to overcome information overload', *Displays*, vol. 12, no. 3–4, pp. 179–189.

Brewster, SA 1998, 'Using non-speech sounds to provide navigation cues', *ACM Transactions on Computer-Human Interaction*, vol. 5, no. 2, pp 224–259.

Brewster, SA 2002, ‘Nonspeech auditory output’, in *The human-computer interaction handbook: Fundamentals, evolving technologies*, eds A Sears & J Jacko, CRC Press, pp. 221–237.

Brown, P & Levinson, SC 1987, *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge.

Busso, C, Deng, Z, Yildirim, S, Bulut, M, Lee, CM, Kazemzadeh, A, Lee, S, Neumann, U, & Narayanan, S 2004, ‘Analysis of emotion recognition using facial expressions, speech and multimodal information’, *Proceedings of the 6th International Conference on Multimodal Interfaces*, State College, PA, USA, pp. 205–211.

Cahn, JE 1990, ‘The generation of affect in synthesized speech’, *Journal of the American Voice I/O Society*, vol. 8, July, pp. 1–19.

Carbaugh, D 1988, *Talking American: Cultural discourses on DONAHUE*, Ablex, Norwood, NJ.

Carbaugh, D 2007, ‘Cultural Discourse Analysis: Communication practices and intercultural encounters’, *Journal of Intercultural Communication Research*, vol. 36, no. 3, pp. 167–182.

Carbaugh, D 2012, ‘A communication theory of culture’. *Inter/Cultural Communication: Representation and Construction of Culture*, ed A Kurylo, Sage, Thousand Oaks, pp. 69–87.

Carbaugh, D, Molina-Markham, E, van Over, B & Winter, U 2012, ‘Using communication research for cultural variability in HMI design’, in *Advances in human aspects of road and rail transportation*, ed N Stanton, CRC Press, Boca Raton, FL, pp. 176–185.

Carbaugh, D & Poutiainen, S 2005, ‘Silence, and third-party introductions: An American and Finnish dialogue’, in *Cultures in conversation*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 27–38.

Carbaugh, D. Winter, U, van Over, B, Molina-Markham, E & Lie, S 2013, ‘Cultural analyses of in-car communication’, *Journal of Applied Communication Research*, vol. 41, no. 2, pp. 195–201.

Corley, M, Stewart, OW 2008, ‘Hesitation disfluencies in spontaneous speech: The meaning of um’, *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602.

Couper-Kuhlen, E & Selting, M 1996, ‘Towards an interactional perspective on prosody and a prosodic perspective on interaction’, in *Prosody in Conversation*, eds Couper-Kuhlen & Selting, Cambridge University Press, Cambridge, pp. 11–56.

Dumas, B, Lalanne, D & Oviatt, S 2009, ‘Multimodal interface: A survey of principles, models and frameworks’, in *Human Machine Interaction: Lecture Notes in Computer Science*, eds D Lalanne & J Kohlas, pp. 3–26.

Ford, CE & Thompson, SA 1996, ‘Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns’, in *Interaction and grammar*, ed EA Schegloff & SA Thompson, Cambridge University Press, CAmbridge, pp. 135–84.

Grice, P 1975, ‘Logic and conversation’, in *Syntax and semantics: Vol. 3 speech acts*, eds P Cole & JL Moran, Academic Press, New York, pp 41–58.

Hereford, J & Winn, W 1994, ‘Non-speech sound in human-computer interaction: A review and design guidelines’, *Journal of Education Computing Research*, vol. 11, no. 3, pp. 211–233.

Heritage, J 1983, *Garfinkel and Ethnomethodology*, Polity, Oxford.

Heritage, J 2010, ‘Conversation analysis: Practices and methods’, in *Qualitative research* (3rd ed), ed D Silverman, Sage, London, pp. 208–230.

Hymes, DH 1972, ‘Models of the interaction of language and social life’, in *Directions in sociolinguistics: The ethnography of communication*, eds JJ Gumperz & D Hymes, Holt, Rinehart & Winston, New York, pp. 35–71.

Hymes, DH 1974, *Foundations in sociolinguistics: An ethnographic approach*, University of Pennsylvania Press, Philadelphia.

Jefferson, G & Schegloff, E 1995, *Lectures on conversation*, Willey Blackwell.

Levinson, SL 1983, *Conversational structure*, Cambridge University Press, Cambridge.

Maynard, SK 1989, *Japanese conversation: Self contextualization through structure and interactional management*, Ablex, Norwood, NJ.

Molina-Markham, E, van Over, B, Lie, S & Carbaugh, D 2014, '"You can do it baby": Non-task talk with an in-car speech enabled system.' Manuscript submitted for publication.

Molina-Markham, E, van Over, B, Lie, S & Carbaugh, D 2015, '"OK, talk to you later": Practices of ending and switching tasks in interactions with an in-car voice enabled interface', in *Globalizing personas: Employing local strategies research to understand user experience*, ed T Milburn, Lexington Books.

Mondada, L 2007, 'Multimodal resources for turn-taking: pointing and the emergence of possible next speakers', *Discourse Studies*, vol. 9, no. 2, pp. 194–225.

Mor, Y 2014, 'The future of human-machine interaction: It's not what you say, it's how you say it', *Wired*. Retrieved from: http://www.wired.com/2014/02/future-human-machine-interaction-say-say/. [21 February 2014].

Oakley, I, Brewster SA & Gray, PD 2000, 'Communicating with feeling', in *Proceedings of the First Workshop on Haptic Human-Computer Interaction*, pp. 17–21.

Philipsen, G 1992, *Speaking culturally*, State University of New York Press, Albany, New York.

Pomerantz, A 1984, 'Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes', in *Structures of Social Action*, eds JM Atkinson & J Heritage, Cambridge University Press, Cambridge, pp. 57–101.

Reisman, K 1974, 'Contrapuntal conversations in an Antiguan village', in *Explorations in the ethnography of speaking*, eds R Bauman & J Sherzer, Cambridge University Press, Cambridge, pp. 110–124.

Rinott, M 2008, 'The laughing swing: Interacting with non-verbal human voice', *Proceedings of the 14th International Conference on Auditory Display*, Paris, France, June 24–27, 2008.

Sacks, H, Schegloff, EA & Jefferson, G 1974, 'A simplest systematics for the organization of turn-taking for conversation', *Language*, vol. 50, no. 4, pp. 696–735.

Schegloff, EA 1992, 'To Searle on Conversation: A note in return', in *(On) Searle On Conversation*, eds H Parret & J Verschueren, Benjamins, Amsterdam, pp. 113–128.

Schegloff, EA 2000, 'Overlapping talk and the organization of turn-taking for conversation', *Language in Society*, vol. 29, no. 1, pp. 1–63.

Scollo, M 2011, 'Cultural approaches to discourse analysis: A theoretical and methodological conversation with special focus on Donal Carbaugh's Cultural Discourse Theory', *Journal of Multicultural Discourses*, vol. 6, pp. 1–32.

Scollon, R, & Scollon, S, 1981, *Narrative, literacy, and face, in interethnic communication*, Ablex, Norwood, NJ.

Sidnell, J 2001, 'Conversational turn-taking in a Caribbean English creole', *Journal of Pragmatics*, vol. 33, no. 8, pp. 1263–1290.

Stivers, T, Enfield, NJ, Brown, P, Englert, C, Hayashi, M, Heinemann, T, & Levinson, S 2009, 'Universals and cultural variation in turn-taking in conversation', *Proceedings of the National Academy of Sciences*, pp. 106–126.

Stokes, R & Hewitt, J 1976, 'Aligning actions', *American Sociological Review*, no. 41, pp. 46–42.

Tannen, D 1993, ‘The relativity of linguistic strategies: Rethinking power and solidarity in gender and dominance’, in *Gender & Discourse*, Oxford University Press, New York & Oxford, pp. 19–52.

Tannen, D 2012, ‘Turn-taking and intercultural discourse and communication’, in *The Handbook of Intercultural Discourse and Communication*, eds CB Paulston, SF Kiesling & ES Rangel, Blackwell Publishing, pp. 135–157.

Wang, P, Winter, U & Grost, T 2015, ‘Cross cultural comparison of users’ barge-in with the in-vehicle speech system’, in *Design, user experience, and usability: Interactive experience design*, Lecture Notes in Computer Science, ed A Marcus, Springer International Publishing, Switzerland, pp. 529-540.

Wechsung, I 2014, *An evaluation framework for multimodal interaction: Determining quality aspects and modality choice*, Springer International Publishing, Switzerland.