Report Number: WUCS-93-33

1993-01-01

The DIM system: Turn-Taking in Dyadic Telephone Dialogues

Umesh Berry and Anne Johnstone

The analysis of human conversations has revealed that the design of interfaces using spoken dialogue must differ radically from those using written communication. Such characteristics as prosody, confirmations, echoes, and other speech phenomena must be considered. This work is a step in that direction. Prosodic, syntactic and semantic information from actual human dialogues has been used to build a turn-taking model empirically for dydadic telephone dialogues. The ability to predict completion of turns has been the biggest motivating factor in the development of this model. The design and evaluation of the model are presented in this report.

... Read complete abstract on page 2.

The DIM system: Turn-Taking in Dyadic Telephone Dialogues

Umesh Berry and Anne Johnstone

Complete Abstract:

The analysis of human conversations has revealed that the design of interfaces using spoken dialogue must differ radically from those using written communication. Such characteristics as prosody, confirmations, echoes, and other speech phenomena must be considered. This work is a step in that direction. Prosodic, syntactic and semantic information from actual human dialogues has been used to build a turn-taking model empirically for dydadic telephone dialogues. The ability to predict completion of turns has been the biggest motivating factor in the development of this model. The design and evaluation of the model are presented in this report.

The DIM system :

Turn-Taking in Dyadic Telephone Dialogues

Umesh Berry and Anne Johnstone

WUCS-TR-93-33

August 15, 1993

Department of Computer Science
Campus Box 1045
Washington University
One Brookings Drive
St. Louis, MO 63130.

Abstract

The analysis of human conversations has revealed that the design of interfaces using spoken dialogue must differ radically from those using written communication. Such characteristics as prosody, confirmations, echoes, and other speech phenomena must be considered. This work is a step in that direction. Prosodic, syntactic and semantic information from actual human dialogues has been used to build a turn-taking model empirically for dyadic telephone dialogues. The ability to predict completion of turns has been the biggest motivating factor in the development of this model. The design and evaluation of the model are presented in this report.

Table of Contents

List of Tables

List of Figures

1. Introduction

As a species, we developed speech long before we developed writing. As individuals, we learn to speak before we learn to write, and speech remains for most of us our primary means of communication with each other. However, in communicating with machines, the sequence of development from speaking to writing was reversed. Early natural language interfaces were designed based on the assumption that the interaction between the machine and the user would be carried out using the keyboard and the terminal. Automatic Speech Recognition (ASR) technology was not developed enough for speech to be considered as a feasible means of interaction. Consequently, the theories underlying the design of natural language interfaces were developed with a text-centered view.

In text, information is encoded as a sequence of words, together with a few punctuation marks and layout details. If the words are known, text can be reproduced exactly. By contrast, in speech, information is encoded in a variety of ways, and just knowing which words were used does not allow a spoken utterance to be reproduced. Setting aside considerations of information content, there are marked differences between speech and text in the way the word sequences are presented. A printed word is separated by a space from its neighbors, and, apart from capitalization at the beginnings of sentences, a given word has the same form irrespective of the identities of its neighbors or of its position in the sentence. In speech there are no gaps necessarily between words, nor indeed any distinct mechanism to indicate word boundaries. The ability of speech to encode information in a set of "prosodic features": in timing, including pauses, in loudness, in pitch, to name a few, has no counterpart in typed language. In this report, we will use the terms "prosodic" and "intonational" interchangeably, and they refer to any acoustic phenomena present in speech. Prosodic cues indicate emphasis on a particular word and reflect speaker attitudes too. Even though the formal rules of grammar underlying the two modes of communication are generally thought to be the same, the styles of language appropriate for writing and speaking are different. In terms of these formal rules, spoken language is more error prone, partly because we have much less time to plan and polish our spontaneous speech than we have for our writing, though many so-called errors in speech may actually be observances of different rules. The words commonly used to link or separate ideas in spontaneous speech are generally different from those used in text. Words like *moreover, nevertheless, consequently* and many others are common in text but rare in speech whereas words like *OK, right, well* and others are more common in speech and are very rarely used in text.

Recent advancements in speech recognition technology have made it possible for speech to be considered as a medium of interaction. From the foregoing discussion, it is evident that speech is more than just an audible version of text and has led natural language interface designers to modify or discard previous theories as they fail to account for a wide range of spoken utterances.

Definitions will emerge in the following chapters, but for the present **conversation** may be taken to be that familiar pre-dominant kind of talk in which two or more participants freely alternate in speaking, which generally occurs outside specific institutional settings like law courts, classrooms, religious services and the like, a **turn** in a conversation may be considered as a speaker's stream of speech bounded by the other

participant's speech and **turn-taking** being the process through which the party doing the talk of the moment is changed.

A striking difference between human-computer typed interaction and human-computer spoken interaction is the way in which the computer and the human user can alternate turns during the course of the interaction. When the interaction is to be carried out using the keyboard and the video terminal, the user can type in the intended command or query and signal a release of turn (the point at which the speaker has finished saying what she wanted to say, for the moment) to the computer by using the return key. The user can have the turn throughout the time period until the return key is hit. There is a clean alternating organization of turns and the intention of releasing or keeping the turn can be simply and unambiguously signaled.

Spoken interaction too, as can be observed in human-human conversations, is characterized by turn-taking: one participant talks, stops; the other starts, talks, stops; and so we obtain an alternating organization of turns across the two participants. But how such a distribution is actually achieved is not that obvious. Research in Conversation Analysis has revealed that less (and often considerably less) than 5 per cent of the speech stream is delivered in overlap (both participants speaking simultaneously), yet gaps between one person speaking and another starting are frequently measurable in just a few micro-seconds (Levinson, 1983). How is this orderly transition from one speaker to another achieved with such precise timing and so little overlap? Whatever the mechanism responsible, it must be capable of operating in quite different circumstances: the number of parties may vary from two to more than two; persons may enter and exit the pool of participants; and turns at speaking can vary from minimal utterances to many minutes of continuous talk. In addition the same system seems to operate equally well both in face-to-face interaction and in the absence of visual monitoring, as on the telephone.

In this report, we will focus on one aspect of two-party spoken interactions (or dyadic spoken interactions) over the telephone - the organization of the conversation into turns. We will limit our research to dyadic conversations, for it provides a good framework to conduct objective research and the findings are fairly generalizable to multi-party conversations. In a similar vein, the findings for telephone conversations can be generalized to face-to-face conversations, with suitable modifications.

1.1 Turn-Taking Knowledge: Why is it necessary?

The idea of a computer with a cognitive ability to match a human has always appealed to researchers in Artificial Intelligence. This, more than anything else, has been a motivation for all the work that has been done on building natural language interfaces for computers. The basic premise for developing natural language interfaces is to help users get their task done, be it a database query or a command for an action, in a way such that there are no constraints on the way users express the desired task. A user should not have to learn an unfamiliar language to interact with a computer. Or, to put it concisely, it should be *natural* for the user to interact with the computer. The word *natural* has had varying interpretations to it as can be observed by the progression from machine languages to the different generations of programming languages to the two modalities of natural languages (written and then spoken). Spoken conversation, for the most of us, is the most natural way of communicating with others. For it to be a natural means of communicating with a computer, it is imperative that the computer possess certain cognitive abilities that humans possess. Humans are able to recognize the words that are

being said, to process the recognized words to interpret their intended meaning, and to keep track of the ongoing dialogue for the interpretation of utterances in context. Another basic cognitive ability that humans possess, that has often been overlooked by spoken natural language interface designers, is the ability to alternate turns in a conversation. This is true even for conversations which are not face-to-face, for instance telephone conversations. Most human conversations have a remarkable orderliness in them. There may be interruptions and brief periods of overlap but it is generally quite clear which speaker has the turn at any particular moment during the conversation. Moreover, that speaker usually gives up the turn voluntarily and this is recognized by the hearer. There is enough evidence to suggest that participants in a conversation subscribe to a set of principles which governs the turn-taking process. These rules seem to be quite complex, operating mainly on prosodic patterns, coupled with lexical cues (syntactic and semantic, to be precise).

It is clear from the foregoing discussion that in order for a computer to be a *natural* partner in a conversation, it should have a knowledge of the turn-taking rules and the patterns (mainly prosodic) on which they operate, for how else would the computer know when a user has finished what she wanted to say or vice versa. One can devise other mechanisms to achieve the same purpose, but these would be at the cost of either the naturalness or the efficiency of the interaction. For instance, both the computer and the user could use a specific word, phrase or any other sound pattern to signal the end of a turn unambiguously (similar to the use of the word *over* in radio communication), but this would mean adding a constraint on the way the interaction can be carried out and would be going against the basic premise that we started with. Another possibility is to rely on periods of silence to signal end of turns but this would make a conversation quite inefficient as far as the time required to complete the dialogue and at the same time would make it unnatural because a large majority of turn-switches in human conversations happen without any pause. Another problem is that on many occasions, a pause is not intended as a signal to release the turn (by prefacing it with a word that indicates hesitation or less frequently by the speaker explicitly stating that the turn is not meant to be released) and is perceived as such by the other participant.

Thus, for a spoken natural language interface to really be natural, it is necessary (but not sufficient) that it detect the prosodic and lexical turn-taking signals that are used so frequently in human conversations and at the same time it should have a working knowledge of the turn-taking mechanism (operating on these signals) that humans employ. In order to test this hypothesis, we ran an experiment (Johnstone, Berry, and Nguyen, 1993a) where we compared conversations across two groups. Group I consisted of subjects talking to a human operator across a telephone line. Group II consisted of subjects talking to a simulated computer system across a telephone line, with the communication channel replaced by an impoverished one, where the turn-taking information was removed.

Preliminary analysis of the data shows that although the subjects in the Group II used approximately 76% of the number of words and 50% of the number of exchanges that subjects used in Group I, subjects in Group II needed approximately *twice* the amount of time to complete the task as compared to the time needed by subjects in Group I. Closer examination of the dialogues from Group II has revealed that most of the pauses occurred when the turn switched from the users to the simulated computer (the "computer" did not receive the turn-taking cues being given by the users), whereas

exchange of turns from the computer to the users were quick (the users did hear everything that was said by the computer).

Illustrative data in this report are drawn where possible from published sources; in these cases the source heads each extract; otherwise, data headed by a number are drawn from the data collected by the team working on the DIalogue Manager (DIM) system (Berry and Johnstone, 1992).The transcription system used in this work is similar to the one which is most commonly used in Conversation Analysis (Sacks et al, 1974). Some of the conventions for the transcriptions are given below. For a complete description, see Appendix A.

::	within a word indicates that the articulation of the sound for the word is noticeably lengthened
>	preceding a word indicates that the word following it was spoken at a noticeably lower volume
/	following a word indicates that the word was spoken with a rising pitch
=	between two words indicates that the words were spoken without any pause in between
[	indicates beginning of overlapping talk
]	indicates end of overlapping talk
{short}	indicates end of overlapping talk
{long}	indicates a longer duration of silence (approximately twice the duration of a short pause).

The following dialogue fragment is from Group II:

(5) *NT-27 (Johnstone, Berry, and Nguyen, 1993b)*

S: hello {short} this is southwestern bell's voice phone service {short} can I help you {short}

U: hhh yes I would like to call a uh {short} call forwarding schedule p::lease {long}

S: the call forward schedule i::s {short} monday through friday from 7 30 am {short} to 6 o'clock pm the calls are not forwarded {short} at all other times {short} the calls are forwarded {short} to 2 2 4 {short} 0 0 {short} 9 >9 {long}

U: I would like to make some changes on that p::lease {short} on m::onday {short} hhh I would like the c::alls from 2:: to 4 pm {short} hhh to go to 7 7 6 {short} 1 2 3 >4 {long}

S: on monday your calls are forwarded from 2 o'clock pm {short} to 4 o'clock pm {short} to 7 7 6 {short} 1 2 {short} 3 4:: {short}

U: I would also like on f::riday {short} hhh for the c::alls to be forwarded {short} to {short} hhh 3 5 6 {short} 4 9 {short} 3 0:: {short} hhh from 7 30 am {short} to 12 o'clock >n::oon {long}

S: on f::riday ....

In the Group II, there were numerous instances (especially in the beginning of dialogues) where the users would say what they desired, pause for a while (probably expecting a response) and finding none, they would continue speaking and give the remaining information or complete the request. Sometimes, the user and the computer would start speaking at the same time after a pause from the user (each interpreting the pause in a different way). In general, conversation was not smooth (probably due to the erratic cues being used). This seems to suggest that the absence of a turn-taking mechanism operating on *prosodic* and lexical patterns leads to the conversation being inefficient and unnatural and hence the experiment clearly demonstrates the need for a turn-taking model in any spoken natural language interface.

1.2 Problem Statement

The aim of this work is to develop a computational model that predicts when a speaker intends to *release*, *keep* or *take* the turn in a spoken conversation.

1.3 Research Methodology

The study of conversations has followed one of the two major approaches: **Discourse Analysis (DA)** and **Conversation Analysis (CA)**. Both approaches are centrally concerned with giving an account of how coherence and sequential organization in discourse is produced and understood. But the two approaches have distinctive styles of analysis. The fundamental difference between the two approaches lies in the amount of actual data that they work on. In DA there is a tendency to take one (or a few) instances of data to analyze and to develop theories. Utterances are viewed as means to advance the mental plans of the speaker and intentions of the speaker are derived from what the speaker says. In contrast, CA is a rigorously empirical approach. The methods are essentially inductive; search is made for recurring patterns across many records of naturally occurring conversations. There is little appeal as possible to intuitive judgements; the emphasis is on what can actually be found to occur. The tendency is to avoid analyses based on single texts. Instead, as many instances as possible of some particular phenomena are examined across texts in order to discover the systematic properties of the organization of talk.

We have combined the two approaches in our work, with more emphasis on the CA approach. The team working on the DIM (DIalogue Manager) project (Berry and Johnstone, 1992) collected forty-five telephone dialogues between users and a human operator as part of an experiment to compare conversation styles of people when they are talking to other people as opposed to computers. These dialogues are transcribed in Appendix B. We studied, in detail, thirty-four of these dialogues, looking for recurring patterns across the dialogues (in accordance with the CA approach) in order to develop a

computational model for turn-taking in a dyadic telephone conversation. The analysis of the data was guided by common sense reasoning (akin to the DA approach). We used the remaining eleven to evaluate the performance of the model. This experiment, which provided all the data required for this work, is documented in (Johnstone, Berry, and Nguyen, 1993a).

Chapter 2 provides a background for this area of research. It highlights the work done in turn-taking by several researchers across different disciplines.

Chapter 3 gives an overview of the DIM system, and how this work represents a part of the whole voice-driven dialogue system.

Chapter 4 presents the computational model and the step-by-step empirical process by which it was developed. This chapter analyses the different turn-taking cues in terms of their frequency of usage and their ability to be recognized unambiguously. Also discussed are the assumptions that have been made in the development of the model.

Chapter 5 evaluates the model on actual dialogues. The results of testing the model are given. Once again the different cues are analyzed for their frequency of usage and their ability to be recognized unambiguously.

Chapter 6 concludes this work indicating possible directions for future work.

2. Previous work

A basic empirical finding about conversation, one that has been discovered independently by different investigators (Allen and Guy, 1974; Argyle, 1969; Duncan, 1972; Duncan, 1974; Goffman, 1967; Jaffe and Feldstein, 1970; Sacks, Schegloff, and Jefferson, 1974; Schegloff, Jefferson and Sacks, 1977; Yngve, 1970), and that can be seen by even casual inspection of almost any fragment of conversation, is that talk within it proceeds through a sequence of turns. Perhaps the aspect of interactive conversation that most distinguishes it from other kinds of discourse production is the choreographing of the switch in roles from speaker to hearer and vice versa. How does a hearer get to take the floor and become a speaker? Similarly, how does a speaker let a hearer know that she has no more to say (for the moment) and is expecting a response? Since we are not usually conscious of having to resolve these problems while we are carrying on a conversation (whether small talk, or executive boardroom decisions) the question arises: Why does conversation seem to flow so smoothly? Only a very small portion of a participant's conversation overlaps another's, and gaps between different speakers' turns are generally measured in fractions of a second (Ervin-Tripp, 1979). Several researchers across different disciplines have attempted to provide answers to these questions. In this section, we will summarize the basic findings of the research that has been conducted in the previous years. All the research has been done from a linguistic point of view, with little consideration for the computational feasibility.

2.1.1 Sacks, Schegloff and Jefferson

The organization of turn-taking in conversation has been most extensively investigated by Sacks, Schegloff, and Jefferson (1974). They believe that turn-taking is central to conversational activity, irrespective of the nature or social setting of the conversation. They suggest that the mechanism that governs turn-taking is a set of rules with ordered options which operates on a turn-by-turn basis, and can be termed a *local management system*. One way of looking at the rules is as a sharing device, an *economy* operating over a scarce resource, namely control of the *floor*. Such an allocational system operates over minimal units, units from which turns at talk are constructed. The end of such units constitute a point at which speakers may change - it is a *transition relevance phase* (TRP). At a TRP the rules that govern the transition of speakers then come into play, which does not mean that speakers will change at that point but simply that they might do so. Operating on the turn-units are the following rules, where C is current speaker, N is next speaker, and TRP is the recognizable end of a turn-unit:

Rule I - applies initially at the first TRP of any turn

(a) If C selects N in current turn, then C must stop speaking, and N must speak next, transition occurring at the first TRP after N-selection

(b) If C does not select N, then any (other) party may self-select, first speaker gaining rights to the next turn

(c) If C has not selected N, and no other party self-selects under option (b), then C may (but need not) continue (i.e. claim rights to a further turn-unit)

Rule II - applies at all subsequent TRPs

> When Rule I (c) has been applied by C, then at the next TRP Rules I (a) - (c) apply, and recursively at the next TRP, until speaker change is effected

These rules provide for the basic observations already noted. Only one speaker will generally be speaking at any one time in a single conversation and overlaps can be predicted to be, at least in the great majority of cases, precisely placed: overlaps will either occur as competing first starts, (as allowed by Rule I (b) and illustrated in (1), where D and L begin speaking together after J has finished) or they will occur where TRPs have been misprojected for systematic reasons, e.g. where a tag or address term has been appended (as illustrated in (2), where U, after a short pause appends the speech with a tag, which overlaps with the speech of S who had interpreted the short pause as a possible transition point), in which case overlap will be predictably brief. These rules thus provide a basis for the discrimination between inadvertent overlap as in (1) or (2) and violative interruption as in (3):

(1) *Sacks, Schegloff and Jefferson, 1978*

J: twelve pounds I think wasn't >it =

D: = can you believe it/
[ ]
L: twelve pounds on the weight watchers' >scale

(2) *C-4D*

S: o::k {short}

U: a::nd that is i::t {short} my schedule
[ ]
S: that that's fine = well thank you = for using the service {short}

(3) *C-4C*

U: ok = so we'll just {short} so all I'll have to do is just call y::ou {short}

S: that's
[ ]
U: and and say >um for this um {short} for the time being just take it o::ff {short}

Although these rules provide an insight into the way people regulate turns in a conversation, they do not specify how a TRP (the end of a turn-unit) is signaled or perceived by humans, nor do they define what a speaker-selects-next technique is. These rules specify what happens at a TRP, once it is reached.

2.1.2 Jaffe and Feldstein

Jaffe and Feldstein (1970) provide the simplest version of what is perhaps the most common hypothesis, the proposal that turn-transition is cued by a discrete signal on the part of the speaker:

> An explanation for the switch of roles is still required, however. We look to the cues operative at the boundary between time domains. The utterance of each speaker is presumably terminated by an unambiguous "end of message" signal, at which point the direction of the one-way channel (and the transmitting and receiving roles) are simply reversed.

In essence, conversation is argued to be like short-wave radio communication, with the production of some equivalent of "over" at the end of each turn signaling to the recipient that she should now take the floor.

2.1.3 Duncan

The turn-taking system proposed by Duncan (1972; 1974) was based on the hypothesis that turn-transition is cued by signals. In this system, the speaker cues her recipient that she is about to relinquish the floor by producing a "turn-yielding" signal. On the basis of empirical observation, Duncan describes six specific turn-yielding signals: rising or falling (but not sustained) pitch at the end of a phonemic clause, elongation of the final syllable of a phonemic clause, the termination of a hand movement used during the turn, a number of stereotyped expressions such as *you know* which may be accompanied by a drop in pitch, and the termination of a grammatical clause. Though the hearer may take the floor after one or more of these signals, she is not required to do so. The more signals displayed at a specific moment, the greater the possibility of the hearer taking the floor. However, the speaker has the ability to neutralize any floor-yielding signals she is displaying with an *attempt-suppressing signal.* This signal consists of the speaker maintaining gesticulation of her hands during the turn-yielding signals. Duncan's work thus provides detailed and important analysis of many phenomena occurring at points of speaker transition.

2.1.4 Beattie

Beattie (1983) evaluated the models proposed by previous researchers, principally the models proposed by Sacks et. al. (1974) and Duncan (1974). In a study employing objective analysis of the temporal properties of natural telephone conversations, he discovered that turn-taking on the telephone was remarkably smooth, quick and efficient. Speakers were found to exchange the floor with minimum delay and with little simultaneous speech. This is compatible with both models since both place a good deal of emphasis on information carried in the auditory channel for turn-taking. However, this does not agree with earlier psychological accounts such as that of Kendon (1967), which placed total emphasis on the role of visually-transmitted signals in the regulation of conversation. Beattie, in another study, found that the turn-yielding cues proposed by Duncan were a good predictor of the smooth exchange of turns but not exactly as Duncan had suggested. The linear relationship between the number of turn-yielding cues and the probability of a listener making a turn-taking attempt suggested by Duncan was not observed by Beattie. He found that special cue combinations were the best predictor of smooth turn-taking attempts by the listener. Clause completion accompanied by a falling

intonation with drawl on the stressed syllable and the termination of a hand gesture seem to operate effectively in conversation to inform the listener that it is her turn to speak.

2.1.5 Goodwin

Goodwin (1981) studied extensively how turns are constituted through mutual interactions of speaker and hearer. His work studied the following phenomena: display of hearership, phrasal breaks (restarts, pauses) and the ordering of mutual gaze within a turn. He proposed how mutual orientation between speakers and hearers is achieved during a turn, how separate actions of speakers and hearers are coordinated in a turn, how speakers keep their talk appropriate for different hearers and how hearership is displayed. Previous work in CA assumed that phrasal breaks occur due to speech defects, however Goodwin showed how they are really tools which speakers employ for continued hearership from the listeners. The main result of Goodwin's work is that it demonstrates how a particular state of gaze is relevant to a turn and how participants use systematic procedures (e.g. phrasal breaks) to achieve or remedy this state. Although his work deals with face-to-face conversations, it has served to highlight the importance of certain phenomena (e.g. pauses filled with *ums)* in regulating turns.

2.1.6 Edmondson

Edmondson (1981), in evaluating the work of Sacks et. al. (1974) correctly observes that a specification is needed of what a possible transition-point is, and what a speaker-selects-next technique is. Further, some criteria is needed on which it can be decided whether the occurrent sequence of turns in a given conversation is indeed the result of the application of the claimed turn-taking rules. Edmondson claims that turn-taking procedures are subject to the control of the speaker and/or hearer, such that turn-selection or assignment is distinct from turn-taking:

> .....one cannot predict at any one point in time that a change of speaker role *will* occur, though one may well be able to distinguish on the basis of what is said, how it is said, and concomitant behaviors such as eye-movement and body-shift, between for example different types of silence.

Although it is true that one cannot predict at any one point in time whether a turn-switch *will* occur, it is quite possible that one can predict when a speaker wants to yield a turn to the listener and when she wants to keep the turn. This distinction is important and could explain why a turn is not accepted by a hearer as is demonstrated in this example:

(4) *Atkinson and Drew, 1979*

A: is there something bothering you or not {long}

A: yes or no {long}

A: eh/

B: no

2.1.7 Green

Green (1989) observed that speakers who have self-selected regularly select the previous speaker as the next speaker by the use of brief questions like *where* or *who did.* By the same token, she proposes that speakers who are not ready to give up the floor, but who sense that a pause will be interpreted as indicating a TRP, may indicate that they are not done speaking by prefacing the pause with a conjunction (e.g. *and, but, or, yet),* which indicates that a sentence is in the process of being uttered, so that taking the floor will be construed as an interruption. Or, the speaker may vocalize the pause (with *uhhh*) while she collects her thoughts; as long as one is vocalizing, one has the floor, even though any words may not be uttered.

2.1.8 Cutler and Pearson

Cutler and Pearson (1986) conducted a controlled experiment in which they demonstrated that a major cue used by humans in regulating turns is the fundamental frequency contour of an utterance. They found that a downstep in pitch is a good turn-yielding cue but a pitch upstep is a good turn-holding one. They acknowledge the fact that other prosodic and vocal quality features are also important, apart from the frequency. They concluded this after observing that many of the utterances which their subjects found ambiguous also had upstepped or downstepped pitch.

2.1.9 French and Local

French and Local (1986) studied the phenomena of interruptions and observed that violative interruptions are signaled and perceived by speakers and hearers through the use of a rising pitch and rising volume at the beginning of the interruption (until the interruption attempt succeeds). This is a significant discovery as it gives a procedure (rather than a description) to differentiate interruptions from back-channel responses.

2.1.10 Keller

Keller (1981) identified some special expressions that exist in English to signal intentions and wishes concerning participants' turns in a conversation:

(i) [I want to have a turn]: "May I interrupt you for a moment," "Can you spare a minute," "I'd like to say something," "I have something to say on that too."

(ii) [I want to keep my turn]: "Wait a second," "Well, let's see now," "What I would say is... ."

(iii) [I want to abandon my turn]: "That's all I have to say on that," "That's about it."

(iv) [I don't want to take a turn]: "I have nothing to say on that," "I'll pass on that."

(v) [Why don't you take a turn]: "So, what do you think of that?," "And what about you?," "What have you got to say on that?."

(vi) [I want to leave the conversational group]: "It's been nice talking to you," "I'd better not take up any more of your time."

Keller mentions that such overt turn-taking signals are of varying frequency. In informal dyadic discourse, they appear to be quite rare, since the intention to take a turn is usually signaled non-verbally. However, intentions to keep a turn are relatively more frequent.

2.1.11 Traum and Hinkelman

Traum and Hinkelman (1992) address a series of problems in the structure of spoken language discourse, including turn-taking and grounding. They view turn-taking as composed of fine-grained actions, which resemble speech acts both in resulting from a computational mechanism of planning and in having a rich relationship to the specific linguistic features which serve to indicate their presence. They give a hierarchical classification of conversation acts, with the turn-taking acts forming the bottom level of the hierarchy.

They propose a series of low level acts to model the turn-taking process. The basic acts are **keep-turn**, **release-turn** (with a subvariant, **assign-turn**) and **take-turn**. In their work, they write:

> There may be several turn-taking acts in a single utterance. The start of an utterance might be a take-turn action (if another party initially had the turn), the main part of the utterance might be keeping the turn, and the end might release it. Conversants can attempt these acts by any of several common speech patterns, ranging from propositional (e.g. "let me say something") to lexical (e.g. "umm") to sublexical. Many turn-taking acts are signalled with different intonation patterns and pauses.

They have employed some utterance-final features using the Pierrehumbert pitch description system (Pierrehumbert and Hirschberg, 1990).

However, in their work, Traum and Hinkelman do not give an account of how the intonational features (and other features) can be used to determine whether a speaker has started an utterance, is in the middle of it or has reached the end of the utterance. This is the main focus of our work. We show how the different intonational and lexical features can be used to determine whether a conversant wants to say something (take-turn), is in the middle of saying something (keep-turn) or has finished speaking (release-turn).

3. The DIalogue Manager (DIM) System: An Overview

The aim of the DIM project is to develop a habitable, speech-driven dialogue system which will interact via naturally spoken English in a limited domain. The idea is that the task-oriented domain itself should lead to natural restrictions on the vocabulary and speech of the users. This rather than constraints imposed by the system designers would create the impression of habitability and an inherent robustness. The DIM system will have to interpret a user's spoken utterance with respect to various goals the user might be trying to achieve. The goals are limited to a certain domain, in the current case, the custom-calling features provided by Southwestern Bell Telephone.

Before we began work on the DIM system, we wanted to have a clear understanding of user requirements and of the spoken language phenomena that occur in a typical task-oriented domain. It is important to research such issues in a real-world environment precisely because people adapt to their environment. Much work has been done on the theoretical aspects of discourse understanding and plan recognition (Litman and Allen, 1987; Lambert and Carberry, 1991) but the results have not been fully exploited or tested because, until now, underlying technologies such as speech and natural language processing were not adequate. If plans are an essential part of human action and cooperation (Pollack, 1991), then it is important to examine their use in complex and dynamic real-world environments. Therefore an important aspect of our work is the combination of experimental and computational exploration with theoretical analysis. We began a series of experiments, using the Wizard of Oz (WOZ) methodology (Kelley, 1983; Fraser and Gilbert, 1991), designed to collect dialogues which would serve as the basis for developing the DIM system. The first experiment was conducted in the summer of 1991. A key role of this experiment was to study the dialogue style and vocabulary used by users when they interact with a computer. We observed that the vocabulary set was small (305 words for 210 dialogues), users accomplished their tasks in uniform, predictable ways, they used mostly short command-like constructs and most important of all, they adapted their style of interaction to that of the computer as their conversation progressed. Details of the experiment can be found in (Balentine, Berry, and Johnstone, 1992). We conducted the second experiment in the series in December, 1992. The aim of this experiment was threefold. Firstly, to show the need for including turn-taking knowledge in any spoken natural language interface. Secondly, to study in detail using actual dialogues, how people take turns in a dyadic conversation over the telephone, and thirdly to compare and contrast how people speak to a computer as opposed to other people. Details of the experiment and the results can be found in (Johnstone, Berry, and Nguyen, 1993a).

Based on an analysis of the dialogues, together with a review of the current literature, we have designed a generic speech-driven dialogue system. An important feature of this system, and one which we believe to be original, is its independence of the type of speech input, whether isolated phrases or continuous speech. This means that it is not necessary to have full natural language processing capability in order to exploit the many advantages of dialogue processing. The primary reasons for this are as follows. DIM's main function is to relate a user's utterance to intended domain actions via a representation of likely goals and plans. The types of discourse plans and the constraint satisfaction procedures used to make these links (Allen, 1987) are the same regardless of the types of phrases used to signal the intention. Therefore, we can use the same dialogue

processing techniques with either isolated phrases or more complex natural language constructions. Since we do not require a complete syntactic and semantic analysis of each utterance, it is not required of the speech recognizer to recognize each word in an utterance. The major components of the system are outlined below:

- The Plan Recognizer (PR) forms the core of the system and has been implemented based on the plan recognition algorithm proposed by Litman and Allen. They assumed the input to the system to be text-based. We believe that the same algorithm can be used effectively for speech input too (Johnstone and Balentine, 1992). The PR is *domain-independent*, allowing users to rely on conventional techniques for information exchange. Details of the PR can be found in (Berry and Groner, 1992).

- The Natural Language Processor (NLP) obtains its input from the speech recognizer and uses keyword spotting to analyze the utterance syntactically and semantically. Details of the implementation of the NLP can be found in (Balentine, Johnstone, and Mathias, 1992).

- The Turn-Taking Module (TTM) regulates the conversation by monitoring the turn throughout the interaction. There is a bi-directional flow of information between the NLP and the TTM as syntactic and semantic information is used in deciding turns and at the same time the NLP needs to know the end of a turn to send the processed utterance to the PR. This part of the system will be implemented based on the computational model .

- The Speech Recognizer (SR) forms the low-level component of the system which generates word hypotheses and detects prosodic cues in the spoken utterance.

4. A Computational Model

4.1 Introduction

When a conversation breaks down, the problem can often be traced to a failure in the turn-taking procedure, i.e. the smooth interchange of speaking turns between conversational partners. For a conversation to function successfully, each speaker's turn should not go too long, and should be accomplished without overlaps; and at the end of one speaker's turn another speaker should take over without too long an intervening pause. Of course, at what point an inter-turn pause becomes "too long" may depend upon the particular conversational circumstances. For any given conversation, however, it is usually obvious whether or not it is proceeding smoothly.

The ability to predict completion of turns has been the biggest motivating factor in the development of the computational model that we will present in this chapter. To take over a turn at the appropriate moment, without undue hesitation, it is obviously useful to be able to decide as early as possible that the previous speaker has finished. In order to predict the completion of turn by a speaker, the model has to keep track of non-completions, or points in the conversation when the speaker intends to keep the turn. The model is also able to distinguish between violative and non-violative interruptions (the former being characterized by the fact that the interrupter intends to take the turn in the conversation). Thus, the model recognizes the following turn-taking acts (Traum, 1991): Release-Turn (RT), Keep-Turn (KT) and Take-Turn (TT). The model has been developed as a passive entity, i.e. it monitors the conversation between two participants and predicts when a turn is kept, released or taken by either of the participants. It will be fairly straightforward to modify the model so that it plays an active part in the conversation, i.e. it represents one of the participants in the conversation.

The model has been built as a finite state machine (FSM). The different states in the machine represent the different states of the conversation, in so far as the turns are concerned. The different states obtain their input in the form of *utterance units*, appended by information (cues) from three dimensions: prosodic, syntactic and semantic. In the following sections in this chapter, we will discuss what these cues are for each dimension. Utterance units are defined as the words that are present in the speech stream, the words being grammatical words that can be found in a dictionary as well as certain sound patterns that are prevalent in speech (e.g. *um, uh-huh,* in-breaths). Currently, the FSM works on inputs that represent speech-internal cues only, i.e. all the cues that can be found in the speech signal. There are more speech-internal cues which are used to regulate turns, but those are not currently used by the model, the most important one is reference to the discourse context. There exist a variety of speech-external cues which speaker may employ to inform hearers where the current turn will end. For example, speakers often look away from the interlocutors while speaking, but look towards them again as they finish talking (Kendon, 1967), especially if speaker and interlocutor do not know each other well (Rutter et al, 1978). Termination of hand gesture has also been claimed to be associated with turn-final utterances (Duncan, 1972). However, since we are dealing with telephone conversations, these cues are not relevant and hence have not been considered.

Since the Turn-Taking Module (TTM) is part of a system with many components, there is a lot of inter-dependence between the many components and consequently it is difficult to build or evaluate it as an independent unit. We have made a few assumptions in the development of the model. We have assumed that the speech recognizer is able to detect the utterance units as they are spoken and is also able to analyze the utterance in terms of the prosodic information contained therein. Further, we have assumed that the natural language processing component is able to detect the syntactic and semantic cues that we mentioned before and will discuss in the following sections. There is enough evidence in the literature to suggest that the detection of all the cues can be automated.

4.2 Methodology

We have adopted the methodology that is used by researchers in Conversation Analysis. The first step was to collect actual data. The dialogues between subjects and the operator, that were collected as part of the experiment, provided the data for this work. There were forty-five dialogues in all, and we used thirty-four of them to build the model and used the remaining eleven to evaluate its performance.

Once the recordings of the conversations were available, the next step was to transcribe them. In the first stage of the transcription process, a professional transcriber, whose services were provided by Southwestern Bell Telephone, was used in order to transcribe the words (both grammatical and non-grammatical) and periods of silence in the dialogues. In the second stage, we appended all the prosodic, syntactic and semantic information to the transcriptions. This stage was an iterative one. We heard the recordings in detail in order to identify the various cues, followed by appending them to the transcriptions. This procedure was done repeatedly, until we could not find any more relevant cues. The transcriptions were cross-checked by another transcriber and suitable modifications (to the agreement of both transcribers) were made.

The final set of transcriptions was then used to empirically develop the computational model. The development of the model was an inductive process. In accordance with the CA methodology, we searched for recurring patterns across the volume of data. The emphasis was on what could be actually found to occur, rather than premature theory construction based on intuitive judgements.

In the next three sections, we will discuss the different prosodic, syntactic and semantic cues that have been used in the computational model.

4.3 Prosodic Cues

There is enough evidence in the literature on turn-taking to suggest that prosody contains active cues for marking boundaries in speech. Broadly speaking, prosody includes pitch (the fundamental frequency of sound), intensity (the amplitude of the sound, which is perceived as loudness), duration (perceived as length) and timing (perceived as the distribution of speech into segments, marked on the boundaries by periods of silence). Based on the experimental studies reported in the literature (see Chapter 2) and on the actual conversations that we collected, we have identified three turn-yielding cues (Lengthening, Down-stepped Amplitude and Rising Pitch), one turn-keeping cue (Contiguous Speech) and one interruptive cue (Rising Pitch and Amplitude). A brief discussion of each follows:

i) Lengthening: This refers to the phenomena of articulating an utterance unit with a noticeable elongation. It is also referred to as segmental lengthening or drawl in the literature. Lehiste (1975) reported that segmental lengthening occurs as a marker of sentence finality, paragraph finality and conversation turn finality.

ii) Down-stepped Amplitude: The amplitude (or loudness) of speech varies over time. An utterance unit is often spoken at a higher volume in order to put stress on it. It has also been observed by Cutler and Pearson (1986) that an utterance unit spoken with a lowered volume (compared to previous utterance units) is associated with turn-finality. A down-stepped pitch has also been found to be associated with turn-completions, but in this work we have considered down-stepped amplitude only.

iii) Rising Pitch: An utterance unit spoken with a rising pitch (or frequency) contour has been shown to be used by conversational participants to indicate a question (Gussenhoven, 1986), especially if the utterance does not have an interrogative form in the grammatical sense. This phenomena can be observed quite frequently in tag questions too. For example, the phrase *that's all* can be articulated with a rising pitch to serve as a question, and with a sustained or falling pitch to serve as a reply to the same question.

iv) Contiguous Speech: There are many instances when the utterance units comprising the speech are spoken with no break between them (either by the same speaker or different speakers) to the extent that to a listener they sound like one utterance unit. We will refer to this phenomenon as contiguous speech. This was observed (in the dialogues we collected) to be used by speakers, especially at possible turn-completion points, to signal that they do not wish to relinquish their turn yet as they have more to say.

v) Rising Pitch and Rising Amplitude: French and Local (1986) observe that the intention to interrupt someone's speech is frequently signaled by speaking with a rising pitch and a rising volume until the other person realizes the intention and stops speaking. The same tool is used by speakers in order to keep a turn in the face of an interruption attempt by another speaker.

Periods of non-speech are frequently used in conversations to mark boundaries. Lehiste (1979) observed that there is a trading relationship in the perception of boundaries between the length of pauses and other prosodic cues. So a relatively short pause is perceived as a boundary if it is coupled with, for instance, a lengthening of the preceding utterance unit. In my work, pauses have been classified as being either short or long. Currently, this distinction has been made on the basis of the judgement of the transcribers, with consistency across the dialogues being the prime consideration. A rigorous timing analysis will eliminate the errors that could have been introduced due to human limitations.

The prosodic cues that have been discussed are speaker dependent. For instance, in order to classify an utterance unit as being spoken with a down-stepped amplitude, it has to be compared to the amplitudes of other utterance units spoken by the *same* speaker. Similarly, pauses have to be classified differently for different speakers with varying speech-rates.

It is significant to note here that these cues do not occur in isolation. In other words, speakers frequently use more than one of them to signal their intent. For instance, lengthening of an utterance unit is frequently coupled with a drop in amplitude or a rise in pitch. There are some combinations which are very rare, for example the use of a rising pitch and a lowered volume. These prosodic cues can also occur in combination with cues from the syntactic or the semantic dimension. This issue will be discussed further in section 4.6. It is also worth mentioning here that amongst these prosodic cues, all but the last two are associated with turn-finality and it is their absence that is associated with turn-mediality (within a turn) and interruptions respectively.

4.4 Syntactic Cues

Syntax plays an important role in the regulation of turns in a conversation. Termination of a grammatical clause has been found to be a turn-yielding cue (Duncan, 1974). If one studies the points at which a speaker voluntarily relinquishes a turn, one can see that in a large majority of the cases the utterances are syntactically complete. The dialogues that we collected strongly support this observation.

We have identified two syntactic cues (both are turn-yielding): syntactic completion and utterances which have an interrogative syntactic form. Note that the latter is really a special case of syntactic completion. The end of a grammatically well formed utterance or an elliptical utterance is defined to be a point of syntactic completion. For example, in the following conversation, syntactic completion points are marked with |synl and end of utterances with an interrogative form are marked with |quesl:

(5) *C-1A*

S: hello |synl this is southwestern bell's phone service |synl can I help you |synl |quesl

U: hhh yes |synl I would like to a::dd |synl {short} a number |synl a name |synl to my speed call d::irectory |synl

Note that in the second utterance, "a name" is the end of an elliptical utterance ("I would like to add a name") and hence is a point of syntactic completion. Note also that "help" in the first utterance and "like" in the second utterance are transitive verbs and hence are not syntactic completion points.

If a telephone number is being spoken, then the entire number consisting of seven digits is considered as one object. In other words, syntactic completion can occur only at the seventh digit, and not in between. In a similar way, if a name is being spelt after it has been uttered, then the list of letters used to spell the name is considered as one object. For example:

(6) *C-1A*

U: hhh t::he number is 7 2 7 {short} 7 5 0 7 |syn| {short}
S: 7 2 7 {short} 7 5 0 7 |syn|
U: y::es |syn| {short}
S: ok = |syn| = and the n::ame/ |syn| {short}
U: hhh mike |syn| {short} m i k e |syn| {short}

Any form of grounding (for e.g. *OK, uh-huh, yes,* verbatim repetition of speech by the previous speaker) is also considered as a point of syntactic completion, as can be observed in the second, third and fourth utterances in the above example.

All forms of closings (the variants of *thank you* and *good-bye* that occur at the end of conversations) are syntactically complete units. For example:

(7) *C-1A*

S: mike |syn| {short} ok = |syn| = I've made that change |syn| {short}
U: thank you |syn| {long}
S: byebye |syn|
U: bye |syn|

As with the prosodic cues, the two syntactic cues considered here are associated with turn-finality and their absence is associated with turn-mediality. As mentioned before, the syntactic cues do not necessarily occur in isolation. They are frequently found in combination with prosodic and semantic cues. The issue of combinations of a syntactic cue with another syntactic cue is not present because an interrogative form necessarily implies syntactic completion and these are the only two syntactic cues being used.

4.5 Semantic Cues

Semantic cues are meant to represent the intention of what has been said. Semantic completion invariably occurs simultaneously with syntactic completion and hence semantic cues have been identified only at points of syntactic completion. We have used two categories of turn-yielding cues (Grounding and Closing) and one category of turn-keeping cues (Hesitation).

i) Grounding: This refers to the various ways in which information is confirmed and clarified in spoken dialogues. In the following example, syntactic completion points where grounding is being done are marked with |grn|:

(8) *C-1A*

U: hhh t::he number is 7 2 7 {short} 7 5 0 7 {short}
S: 7 2 7 {short} 7 5 0 7 |grn|
U: y::es |grn| {short}

ii) Closing: This is mostly uttered at the end of conversations in order to thank the other participant or to bid the other participant good-bye. In the following example, they are marked with |thnl and |byel respectively:

(9) *C-2B*

U: hhh ok very good thank you |thnl {short}

S: thank you |thnl

U: uh huh = bye |byel

iii) Hesitation: In spoken conversation, participants have limited time to convert their thoughts into words and consequently it is marked by numerous pauses which the participants utilize to verbalize their thoughts. In order to indicate to the other participant that they do not intend to give the turn away, they either fill the pause with a filler like *um* or they explicitly indicate that they want to keep the turn. A noticeable difference between the two is that the latter is used to keep the turn for a longer duration of time. In the following example, they are marked as |hesl and |holdl respectively:

(10) *C-2B*

U: hhh on f::riday hhh I'm going to be at 3 5 6 {short} 4 9 3 0 hhh um |hesl {short} from 7 30 to n::oon {short}

S: o::k {short} that's 3 5 6 {short} 4 9 3 >0 {short}

U: uh huh = hhh 7 30 to >noon and on {short} >let's see |holdl {long} hhh a::nd {short} I'll be at my grandma's = let's see |holdl the number is 3 3 9 {short} 2 3 2 3 {short}

The semantic cues are mutually exclusive, i.e. combinations of semantic cues with themselves are not found. However, they are found in conjunction with cues from the other two domains.

4.6 The Computational Model

Having identified the various cues that influence turn-taking, the next step was to study the distribution of these cues and their combinations over the conversations in order to develop rules which operate on them. In other words, for each combination, we computed the number of times it occurred at a point where the turn was switched, kept and forcibly taken. For the sake of completion, we had considered all possible combinations of cues, not only within one dimension, but across different dimensions. The total number of such combinations was 194. Some examples are:

i) an utterance unit

ii) an utterance unit which marks syntactic completion

iii) an utterance unit which is lengthened and marks syntactic completion

iv) an utterance unit which is lengthened, is spoken at a lowered volume and marks syntactic completion

v) an utterance unit which is lengthened, is spoken at a lowered volume, marks syntactic completion and is followed by a short pause

vi) an utterance unit which is lengthened, is spoken at a lowered volume, marks syntactic completion and is followed by two short pauses (i.e., a long pause)

It is important to note that all the cues (prosodic, syntactic and semantic) that the combinations are composed of are such that they can be detected and made available simultaneously. Consequently, there is a one-to-one mapping from what is said to each of the combinations. For instance, i) implies that the utterance unit does not have any cues associated with it at all and ii) implies that the utterance unit has one syntactic cue (syntactic completion) and does not have any prosodic or semantic cues associated with it. iv) represents an utterance unit that is a combination of one syntactic cue (syntactic completion) and another combination (a combination of two prosodic cues: lengthening and down-stepped amplitude).

It is clear that it would have been extremely cumbersome to try to develop rules that would operate on each of the combinations separately. Hence, the next step was to group the combinations in such a way that each group contained combinations with similar distributional characteristics. Let's consider an example of this process of grouping. Table 1 lists some combinations along with their distributional pattern. A quick reference to the various shorthand notations used is given below the table. A complete listing is given in Appendix C.

Table 1. Distributional Characteristics of some Combination of Cues

	Release-Turn	Keep-Turn	Take-Turn
uu	0	3834	0
uu sp	2	113	0
*uu sp sp	2	0	0
uu_hes	0	111	0
uu_hes sp	1	38	0
**uu_hes sp sp	0	10	0
*uu_hes sp sp sp sp	2	0	0
uu_da	0	29	0
uu_da sp	1	2	0
uu_da sp sp	0	0	0
uu_da_hes	1	0	0
uu_da_hes sp	0	0	0
uu_da_hes sp sp	0	0	0
uu_sl	0	39	0
uu_sl sp	0	24	0
uu_sl sp sp	0	1	0
uu_sl_hes	0	0	0
uu_sl_hes sp	0	0	0
uu_sl_hes sp sp	0	0	0
uu_rp	0	0	0
uu_rp sp	0	0	0
uu_rp sp sp	0	0	0
uu_rp_hes	0	0	0
uu_rp_hes sp	0	0	0
uu_rp_hes sp sp	0	0	0
uu_da_sl	0	0	0
uu_da_sl sp	0	1	0
uu_da_sl sp sp	0	0	0
uu_da_sl_hes	0	0	0
uu_da_sl_hes sp	0	0	0
uu_da_sl_hes sp sp	0	0	0
uu_sl_rp	0	0	0
uu_sl_rp sp	0	0	0
uu_sl_rp sp sp	0	0	0
uu_sl_rp_hes	0	0	0
uu_sl_rp_hes sp	0	0	0
uu_sl_rp_hes sp sp	0	0	0
uu_da_sl_rp	0	0	0
uu_da_sl_rp sp	0	0	0
uu_da_sl_rp sp sp	0	0	0
uu_da_sl_rp_hes	0	0	0
uu_da_sl_rp_hes sp	0	0	0
uu_da_sl_rp_hes sp sp	0	0	0

uu: utterance unit hes: hesitation sl: segmental lengthening
sp: short pause da: downstepped amplitude rp: rising pitch

It can be observed from Table 1 that all combinations represent points of syntactic *in*completion and moreover all but two (marked with * in the table) of them occur at point where the turn is kept. One of these two, consists of two short pauses (equivalent to a long pause) following an utterance unit (uu sp sp). Thus as a first attempt, one could conclude that speakers do not intend to relinquish their turns at syntactically incomplete points, except if these points are followed by a long pause. Closer observation reveals that there is an exception to the above rule. If an utterance unit has some form of hesitation associated with it, and two short pauses (a long pause) occur after it (uu_hes sp sp, marked with ** in the table), then the turn is kept by the same speaker. However, a turn cannot be held for too long, and hence if the pause following the utterance unit with hesitation is significantly long (equivalent to four short pauses) (uu_hes sp sp sp sp), then the turn is indeed switched, as can be seen from the table.

Thus, from the foregoing discussion, it is evident that the combinations can be classified into two major groups: one which contains utterance units which occur at syntactically incomplete points and have some form of hesitation associated with them and the second which contains utterance units which occur at syntactically incomplete points but have no form of hesitation associated with them. Thus, we developed the following rule which operates on the first group:

R1 "If an utterance unit is tagged with zero or more prosodic cues, and is tagged with hesitation, then it keeps the turn for a duration less than four short pauses. A pause longer than or equal to that releases the turn."

Similarly, the following rule operates on elements of the second group:

R2 "If an utterance unit is tagged with zero or more turn-yielding or turn-keeping prosodic cues then it keeps the turn for a duration less than two short pauses. A pause longer than or equal to that releases the turn."

The same technique was applied to the other combinations in order to group them and to develop rules for such groups. It is important to note that the absence of any cue (or a combination of cues) in the left hand side of the rules specifically implies that they are not present. Table 2 gives the distributional pattern for the groups that were identified. This table gives the number of times these groups were used at points where the turn was kept, released or taken. For a description of the shorthand notations used, see Appendix C. Based on these numbers, the following rules, which operate on the groups, were developed:

R3 "A pause whose duration is more than two short pauses, releases the turn. An exception to this rule is when the pause is preceded by an utterance unit tagged with hesitation or hold."

R4 "A pause whose duration is less than or equal to a short pause, keeps the turn. An exception to this rule is when the pause is preceded by an utterance unit tagged with syntactic completion."

R5 "If an utterance unit is tagged with contiguous speech and zero or more other cues (prosodic, syntactic or semantic), then it keeps the

turn for less than two short pauses. A pause longer than or equal to that releases the turn."

R6 "If an utterance unit is tagged with syntactic completion then it keeps the turn for a duration less than one short pause. A pause longer than or equal to a short pause, releases the turn."

R7 "If an utterance unit is tagged with one or more turn-yielding prosodic cues (lengthening, down-stepped amplitude or rising pitch), and is tagged with syntactic completion, then it releases the turn immediately."

R8 "If an utterance unit is tagged with zero or more turn-yielding prosodic cues, is tagged with syntactic completion, and is tagged with one or more turn-yielding semantic cues (grounding or closing), then it releases the turn immediately."

R9 "If an utterance unit is tagged with zero or more turn-yielding prosodic cues, is tagged with an interrogative form and is tagged with zero or more turn-yielding semantic cues, then it releases the turn immediately."

R10 "If an utterance unit is tagged with zero or more prosodic cues, is tagged with syntactic completion, and is tagged with hold, then it keeps the turn indefinitely."

R11 "If an utterance unit is tagged with rising pitch and rising amplitude and is tagged with any other cues (prosodic, syntactic and semantic), then it takes the turn immediately (even if the other speaker hasn't relinquished the turn)."

Rule 10 states that when the *hold* semantic cue is detected, it keeps the turn with the speaker indefinitely. We did not observe any instances where a speaker signaled a hold in the conversation and did not re-initiate the conversation, leading the hearer to take the turn. In the absence of any empirical data, we decided to let the *hold* keep the turn indefinitely. It would be fairly easy to modify the model in order to set a time-limit for the turn to be kept in a situation like this.

These eleven rules and the groups of cues that they operate on, form the basis for the development of the finite state machine, which is given in a tabular form in Figure 1. The finite state machine is characterized by ten states, each one of them representing the state of the conversation, as far as the turn is considered. The inputs to the machine are the various groups of cues that were given in Table 2. The finite state machine has been built as a Moore machine (the output is associated with a state). In Figure 1, the states and the inputs have been represented by the numbers 1 to 10, and the alphabets A to K respectively.

Table 2. Distributional Characteristics of the Groups of Cues for the Training Set

	KT	RT	TT
uu_*_cs, t < 2 sp	220	0	0
uu_*p_syn_hold	6	0	0
uu_rp_ra_*sy_*se, t < 2 sp	0	0	10
uu_*typ_syn_bye	0	39	0
uu_*typ_*tkp, t < 2 sp	4043	3	0
t >= 2 sp	0	2	0
uu_*p_hes, t < 4 sp	159	2	0
t >= 4 sp	0	2	0
uu_*typ_syn_ques_*tyse	1	73	0
uu_syn, t < sp	402	7	0
t >= sp	16	68	0
uu_*typ_syn_thn	2	62	0
uu_+typ_syn	13	210	0
uu_*typ_syn_grn	44	235	0

	A	B	C	D	E	F	G	H	I	J	K	L
1	2	2	9	3	8	1	10	10	10	10	10	2
2	2	2	9	3	8	4	10	10	10	10	10	2
3	2	2	9	3	8	5	10	10	10	10	10	2
4	2	2	9	3	8	10	10	10	10	10	10	2
5	2	2	9	3	8	6	10	10	10	10	10	2
6	2	2	9	3	8	7	10	10	10	10	10	2
7	2	2	9	3	8	10	10	10	10	10	10	2
8	2	2	9	3	8	10	10	10	10	10	10	2
9	2	2	9	3	8	9	10	10	10	10	10	2
10	2	2	9	3	8	10	10	10	10	10	10	2

A : uu_*_cs
B : uu_*typ_*tkp
C : uu_*p_syn_hold
D : uu_*p_hes
E : uu_syn
F : sp

G : uu_+typ_syn
H : uu_*typ_syn_grn
I : uu_*typ_syn_thn
J : uu_*typ_syn_bye
K : uu_*typ_syn_ques_*tyse
L : uu_rp_ra_*sy_*se

Figure 1. The Finite State Machine

Conversation begins in the **start** state (1). As soon as one of the speakers starts speaking, the conversation enters into one of the other states. For instance, if an utterance unit is tagged with zero or more prosodic cues, syntactic completion, and hold, then the machine enters a state such that any length of pause does not give the turn away.

The **keep-turn** state (2) represents that state of the conversation when the speaker does not intend to give the turn away. The transition to this state always occurs on an utterance unit which is tagged with contiguous speech (and zero or more other cues) or on an utterance unit that is tagged with zero or more prosodic cues.

An utterance unit that is tagged with zero or more prosodic cues, and is tagged with hesitation, makes the machine enter the **hesitation** state (3). The speaker still has the turn, like in the previous state, but the pause that can follow without there being a turn-switch is longer in this state.

The next four states of the machine are used to count the number of short pauses that have occurred since some transition. As the name suggests, the **after-one-sp-from-keep-turn** state (4) is reached after one short pause from the keep-turn state. Similarly, **after-one-sp-from-hesitation** (5), **after-two-sp-from-hesitation** (6), and **after-three-sp-from-hesitation** (7) are states that are reached after one, two and three short pauses respectively from the hesitation state. Each of the states corresponds to the turn not being relinquished.

Any time that an utterance unit tagged with syntactic completion is encountered, the machine enters the **syntactic-completion** state (8). This is the only state that interprets a short pause as a released turn. Of course, as can be observed from the Table 2, there are other cues which, if they occur, can signal a released turn.

If an utterance unit that is tagged with zero or more prosodic cues, syntactic completion, and hold, is encountered, then the machine enters the **hold** state (9). Conversation is put on a hold, i.e. the speaker who had the turn previously keeps the turn indefinitely during an ensuing pause, until she starts speaking again, leading to a transition to the **keep-turn** state.

The **transition-relevance-phase (TRP)** state (10) is reached whenever a turn-yielding signal is encountered from any state, i.e. the turn is predicted to switch after this state is reached. The TRP state could be considered to be the final state of the finite state machine, although there are possible transitions out of it. It absorbs any pause that occurs (after the signal is given) during the transition of the turn.

It is significant to note here that at any time an utterance unit (from the speaker who does not currently have the turn) is encountered, the turn is forcibly taken by this speaker, in spite of the fact that the previous speaker did not have the intention to release the turn. This transition does not go through the TRP state. Thus, the finite state machine monitors the conversation continuously and predicts when a turn will be released, when it will not be released, and when it is forcibly taken.

Let us consider the working of the finite state machine with an example. Figure 2 gives a portion of the finite state machine as a state diagram. Conversation begins in the start state and reaches the Keep-Turn state whenever an utterance unit with contiguous speech or an utterance unit with any turn-keeping or turn-yielding prosodic cues but no syntactic or semantic cues is encountered. Conversation remains in the same state as long as inputs of these types are encountered. If a long pause is encountered, conversation reaches the TRP state, i.e. the model predicts that the turn will be released. In the Keep-Turn state, if an utterance unit with syntactic completion and no other cues is observed, the syntactic completion state is reached and from this state only a short pause is needed to reach the TRP state. In a similar vein, if an utterance unit with syntactic completion and one or more turn-yielding prosodic cues (uu_+typ_syn) is encountered in the Keep-Turn state is observed, then the TRP state is reached immediately. Any form of groundings (uu_*typ_syn_grn), closings (uu_*typ_syn_thn or uu_*typ_syn_bye) or questions (uu_*typ_syn_ques_*tyse) also have the same effect. If an utterance unit tagged with hesitancy (uu_*p_hes) is observed, the conversation reaches the Hesitation state and a significantly longer pause is required for the conversation to reach the TRP state. This example clearly demonstrates how pauses are interpreted differently in different contexts.

4.7 Implementation

The finite state machine has been implemented in Quintus Prolog. The program takes as input the utterance units tagged with the various cues. Figure 3 gives a conversation as it actually happened (with respect to the turns in the conversation) and how the model predicted it would happen. The predictions of the model are given in boldface below the transcriptions of the actual conversation. **RT** represents a prediction by the model that the turn has been released, **TT** represents a prediction by the model that the turn has been taken, and a blank space represents a prediction by the model that the turn is kept.

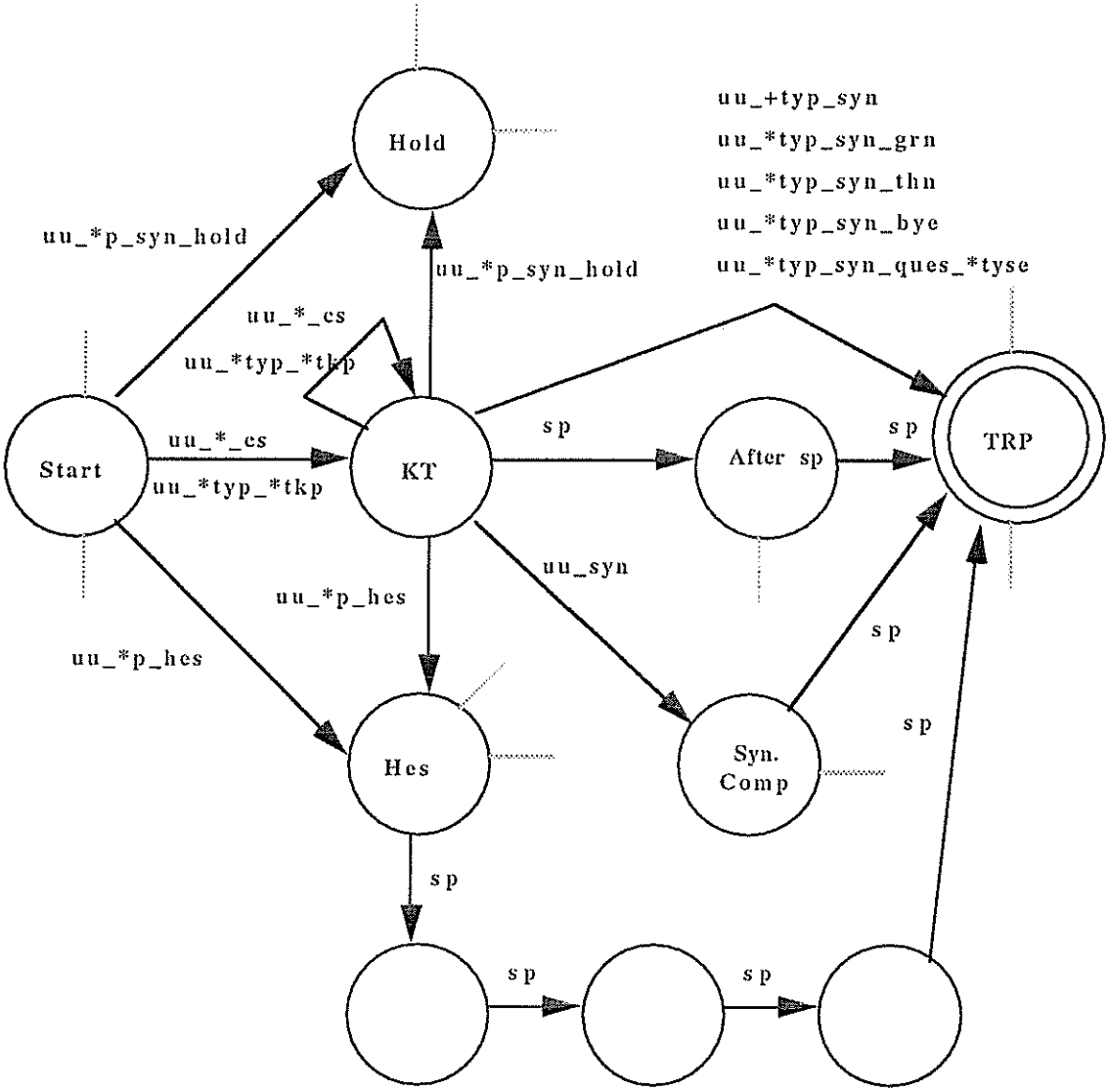

Figure 2. State Diagram of a Portion of the Finite State Machine

S: hhh hello this is southwestern bell's phone service can I help you {short}
RT

U: um good morning I need to make um hhh um I need to add someone to my {short}
um speed call d::irectory {short}
RT

S: hhh o::k {short} what would you like to >a::dd
RT **RT**

U: um his telephone number {short} is 7 2 7 hhh 7 5 0 7 {short}
RT

S: um hmm = 7 {short} 2 7 {short} 7 5 0 7 {short}
RT

U: r::ight
RT

S: o::k {short}
RT

U: hhh and the e::ntry um {short} will be for m::ike {short}
RT

S: for mike {short} ok {short} I've got that {short}
RT **RT** **RT**

U: o::k {short}
RT

S: that's all/ {short}
RT

U: that's i::t {short}
RT

S: thank you
RT

U: thank you = goodbye
RT

S: byebye
RT

Figure 3. Sample Output

5. Evaluation

5.1 Introduction

The turn-taking model, that was presented in the previous chapter, is built in such a way that it monitors a conversation between two participants and predicts when the turn will be switched, when it will not be switched (or kept) and when an interruption is successful (or when a turn has been taken). This model was built by studying the distributional pattern of the various cues at places where the turn is switched, kept and interrupted. Out of a total of forty-five dialogues, thirty-four of them (which we will call the *training set*) were used for the above purpose. The remaining eleven (which we will call the *testing set*) were used to evaluate the performance of the model. We evaluated the performance of the model on the training set also, and the results for both the sets are given in this chapter. The evaluation was done by comparing the predictions of the model with what actually happened in the conversation.

5.2 The Training Set

This set consisted of the thirty-four dialogues that were used to develop the turn-taking model. Table 3 gives the number of times a turn was actually kept, released or taken in each dialogue and the number of times the corresponding predictions of the model. Here is a summary of the evaluation results for the training set:

- Number of times that the turn was actually kept: 4906.
  Number of times (out of these) that the model correctly predicted: 4828
  Accuracy: 98.4%
- Number of times that the turn was actually released: 703
  Number of times (out of these) that the model correctly predicted: 692
  Accuracy: 98.4%
- Number of times that the turn was taken by an interruption: 10
  Number of times (out of these) that the model correctly predicted: 10
  Accuracy: 100%
- Total number of actual turn-related events: 5619
  Number of times (out of these) that the model correctly predicted: 5530
  Accuracy: 98.4%

Table 3. Evaluation Results for the Training Set

	Actual			Model		
	KT	RT	TT	KT	RT	TT
Dialogue 1a	61	12	0	58	12	0
Dialogue 1b	121	21	1	118	21	1
Dialogue 1c	183	32	1	181	32	1
Dialogue 1d	128	24	1	125	24	1
Dialogue 2b	162	22	0	155	21	0
Dialogue 2c	97	19	1	94	19	1
Dialogue 2d	70	13	0	69	13	0
Dialogue 3a	171	29	0	170	29	0
Dialogue 3c	67	12	0	66	10	0
Dialogue 3d	120	19	0	119	19	0
Dialogue 4a	183	24	0	181	24	0
Dialogue 4b	74	11	0	73	11	0
Dialogue 4c	173	19	0	171	19	0
Dialogue 4d	237	28	0	236	28	0
Dialogue 5a	68	10	0	66	10	0
Dialogue 5b	86	8	0	85	8	0
Dialogue 5c	302	31	0	302	31	0
Dialogue 5d	157	21	0	154	21	0
Dialogue 6a	130	18	1	129	18	1
Dialogue 6b	217	36	0	213	35	0
Dialogue 6c	111	16	0	103	16	0
Dialogue 6d	101	16	0	101	16	0
Dialogue 7a	196	35	1	194	33	1
Dialogue 7b	175	28	1	173	28	1
Dialogue 7c	173	26	1	173	25	1
Dialogue 7d	135	18	0	131	17	0
Dialogue 8a	111	24	0	105	24	0
Dialogue 8b	63	12	0	61	12	0
Dialogue 8c	99	12	1	99	12	1
Dialogue 8d	156	24	0	155	24	0
Dialogue 9a	95	13	0	95	13	0
Dialogue 9b	240	24	0	239	21	0
Dialogue 9c	244	23	0	240	23	0
Dialogue 9d	199	22	1	193	22	1
Total	4906	703	10	4828	692	10

5.3 The Testing Set

This set consisted of eleven dialogues that were not used at all in the development of the model. Table 4 gives the number of times a turn was actually kept, released or taken in each dialogue and the number of times the model predicted that the turn will be kept, released or had been taken. Here is a summary of the evaluation results for the testing set:

- Number of times that the turn was actually kept: 1550.
  Number of times (out of these) that the model correctly predicted: 1529
  Accuracy: 98.6%
- Number of times that the turn was actually released: 219
  Number of times (out of these) that the model correctly predicted: 214
  Accuracy: 97.7%
- Number of times that the turn was taken by an interruption: 1
  Number of times (out of these) that the model correctly predicted: 1
  Accuracy: 100%
- Total number of actual turn-related events: 1770
  Number of times (out of these) that the model correctly predicted: 1744
  Accuracy: 98.5%

5.4 Overall

The results for all the dialogues are:

- Number of times that the turn was actually kept: 6456.
  Number of times (out of these) that the model correctly predicted: 6357
  Accuracy: 98.4%
- Number of times that the turn was actually released: 922
  Number of times (out of these) that the model correctly predicted: 906
  Accuracy: 98.2%
- Number of times that the turn was taken by an interruption: 11
  Number of times (out of these) that the model correctly predicted: 11
  Accuracy: 100%
- Total number of actual turn-related events: 7389
  Number of times (out of these) that the model correctly predicted: 7274
  Accuracy: 98.4%

Table 4. Evaluation Results for the Testing Set

	Actual			Model		
	KT	RT	TT	KT	RT	TT
Dialogue 10a	121	16	1	118	15	1
Dialogue 10b	213	31	9	211	31	0
Dialogue 10c	132	21	0	128	21	0
Dialogue 10d	67	10	0	67	9	0
Dialogue 11a	200	28	0	197	27	0
Dialogue 11b	221	27	0	217	26	0
Dialogue 11c	53	13	0	53	12	0
Dialogue 11d	194	15	0	193	15	0
Dialogue 12a	163	33	0	163	33	0
Dialogue 12b	75	15	0	72	15	0
Dialogue 12c	111	10	0	110	10	0
Total	1550	219	1	1529	214	1

5.5 Overall Distributional Pattern for the Cues

Table 5 gives the frequency of usage of the cues in turn-medial, turn-final and turn-initial (in an interruption) positions in all the dialogues. The cues have been ranked in increasing order of ambiguity, i.e. their inability to be recognized as either turn-medial, turn-final or turn-initial (in an interruption).

5.6 Some Remarks about the Evaluation Process

In order to be as fair as possible, the predictions of the model were compared to what actually happened in the dialogue. If the prediction of the model does not agree with what actually happened, it has been termed as an error. However, such an evaluation process is not without its problems. For example, consider the following dialogue fragment:

(11) *C-12B*

U: um good morning I need to make um hhh a I need to add someone to my {short} um speed call d::irectory {short}

S: hhh o::k {short} what would you like to >a::dd

U: um his telephone number {short} is 7 2 7 hhh 7 5 0 7 {short}

In the utterance by S, the short pause is preceded by an utterance unit that marks syntactic completion, is lengthened and is a form of grounding. This combination of cues enables the model to predict that the turn will be switched here, but actually it is not. It is quite possible that S intended to relinquish the turn here but observing that U does not respond, S takes up the turn again. However, according to the evaluation system, this occurrence is counted as an error.

In a similar way, there are a few instances where the model predicts that the turn will be kept, but in the actual dialogue it was switched although it could have equally well been kept. Consider this dialogue fragment:

(12) *C-7A*

S: o::k {short}

U: and um {short} um {long}

S: what changes would you like to >make {short}

U: well on monday >I have a meeting um from 2 to 4 in the a::fternoon {short}

Table 5. Distributional Characteristics of the Groups of Cues for all the Dialogues

	KT	RT	TT
uu_*_cs, t < 2 sp	294	0	0
uu_*p_syn_hold	11	0	0
uu_rp_ra_*sy_*se, t < 2 sp	0	0	11
uu_*typ_syn_bye	0	55	0
uu_*typ_*tkp, t < 2 sp	5280	4	0
t >= 2 sp	0	3	0
uu_*p_hes, t < 4 sp	245	2	0
t >= 4 sp	0	2	0
uu_*typ_syn_ques_*tyse	1	94	0
uu_syn, t < sp	529	8	0
t >= sp	22	89	0
uu_*typ_syn_thn	3	81	0
uu_+typ_syn	15	275	0
uu_*typ_syn_grn	56	309	0

At the end of the first utterance by U, when the turn is switched after a long pause, the model predicts that the turn will be kept because there is an utterance unit that indicates hesitation preceding the pause. It is entirely possible that the user did not intend to relinquish the turn, but S, being knowledgeable about the domain, jumped in to make the interaction proceed faster by offering help. The point is that, conceivably S could have not taken the turn, and U would have started speaking again. Once again, such an occurrence is counted as an error by the evaluation process.

This is not to say that all the errors fall into one of the two categories above. Consider the following example where the model incorrectly predicts that a turn will be released:

(13) *C-1C*

S: hello this is southwestern bell's phone service can I help you {short}

U: hhh yes hhh I need to make {short} a few changes in my weekly call forwarding schedule {short}

S: certainly hhh what can I change for >you

In the second utterance by S, immediately after S says "certainly", the model predicts that the turn will be switched since it constitutes a form of grounding. However, the turn is actually not released and it would be uncooperative on the part of S to just say "certainly" after U makes an indirect request.

Consider the following dialogue fragment which demonstrates an error on the part of the model in predicting that a turn will be kept:

(14) *C-3C*

U: uh huh = and >um {short} when I {short} call and ask for mike >that'll be the number that I get is that correct/

S: that's correct

U: o::k {short} thank you = very much {short}

At the end of the utterance by S, the model predicts that the turn will not be released, but the model does say that if there is a short pause following the word "correct", then the turn will be released. However, if we look at the first utterance by U, we can see that it ends with a question that expects a yes or a no (or something to that effect) for an answer. Upon getting an expected answer from S (in the form of "that's correct"), U need not wait to be sure that S has relinquished the turn.

6. Conclusion

In this work, we developed a computational model for turn-taking in dyadic telephone conversations. We saw the empirical approach which was employed in the development of this model. The way the model works is that it monitors the conversation between two participants, on the telephone, and predicts when a turn will be released, when a turn will be kept and when a turn has been taken via an interruption. These predictions are made on the basis of rules that were empirically developed and which operate on cues from three dimensions - prosody, syntax and semantics. The working of the model was evaluated on actual human conversations and it was found to be very accurate in its predictions. Of all the turn-related events, the model was able to correctly predict close to 98% of them. We believe that the nature of the conversational setting (a task-oriented dialogue in a limited domain) has a lot to do with the high accuracy of the model. The subjects had to give a lot of information to the operator (in the form of telephone numbers, times, days, names etc.) and this required a lot of rapid turn switches for confirmation, which were easily predicted by the model. It would be interesting to see how it works in an unconstrained setting. It is also important to remember that the transcriptions were done by human transcribers, a process that is not very precise and open to some error. We saw that although the evaluation process is not foolproof, it is still a reliable benchmark to ascertain the applicability of the model.

The model is still preliminary and some more work needs to be done to reduce the errors and to make it more flexible. We saw in the previous chapter that some of the errors were due to limitations in the model. Lets consider dialogues (13) and (14) again:

(13) *C-1C*

S: hello this is southwestern bell's phone service can I help you {short}

U: hhh yes hhh I need to make {short} a few changes in my weekly call forwarding schedule {short}

S: certainly hhh what can I change for >you

This is an example of an incorrect prediction by the model that the turn will be exchanged. In the second utterance by S, immediately after S says "certainly", the model predicts that the turn will be switched, since "certainly" constitutes a form of grounding. However, the turn is actually not released. Closer observation reveals that in the previous utterance, U has made an indirect request and would expect the system to be cooperative about it, in the sense that U would expect the system to do more than just acknowledge the request. Thus, a knowledge of the user's beliefs would be required to know what the expectations are and to respond accordingly. So, the model should be able to distinguish between situations which call for acknowledgements only, and situations which call for more than just acknowledgements in a cooperative setting.

(14) *C-3C*

U: uh huh = and >um {short} when I {short} call and ask for mike >that'll be the number that I get is that correct/

S: that's correct

U: o::k {short} thank you = very much {short}

This is an example of an incorrect prediction by the model that the turn will be kept. At the end of the utterance by S, the model predicts that the turn will be kept by S, but it is actually exchanged. The model does say that if there is a short pause following the word "correct", in the utterance by S, then the turn will be released. However, if we look at the first utterance by U, we can see that it ends with a question that expects a yes or a no (or something equivalent) for an answer. Upon getting an expected answer from S (in the form of "that's correct", which is really another way of saying "yes"), U need not wait to be sure that S has relinquished the turn. Thus, a discourse context would be required in order to interpret certain responses as complete.

The model currently classifies utterances as back-channel responses (responses which are not intended to take the turn) if they have a form of grounding, if they overlap with the other participant's speech, and they do not have an interruptive form (a rising pitch and rising amplitude). The first utterance by S in the following dialogue fragment constitutes a back-channel response:

(15) *C-1C*

U: then on f::riday hhh in the m::orning {short} from 7 30 until n::oon
[ ]
S: ok
{short} I will be {short} at 3 5 6 {short} 4 9 3 0:: {short}

S: o::k {long} that's friday 7 30 to noon {short} and the number is 3 5 6 {short} 4 9 3 0 {short}

However, a similar utterance, if said in a non-overlapping manner, would not constitute a back-channel response and would be considered as part of a different turn. The problem with such a classification is that utterances which serve the same purpose (of grounding what has been said) are classified differently. This can be seen from the following dialogue fragment, where the utterance by S is similar to the one in the previous dialogue, but is interpreted differently:

(16) *C-5C*

U: hhh on friday I also won't be at work during the hours from 7 30 till 12 n::oon {short}

S: o::k {short}

U: hhh and I need my calls at that point to be forwarded to a number of 3 5 6:: {short} 4 9 {short} 3 0 {short}

A possible solution for a better classification scheme for back-channel responses could be to base the classification on the semantic content (apart from the other things mentioned before) of the preceding utterance. So, in dialogue (15), since the preceding utterance ("then on friday") does not have any semantic content as far as the domain for

the conversation is considered, the utterance by S ("ok") could be classified as a back-channel response. This is in contrast to dialogue (16), where the preceding utterance does have a significant semantic content.

The model has been developed as a passive entity, in the sense that it does not take part in the actual conversation. It would be fairly easy to modify the model so that it represents the turn-taking ability of a participant in a conversation. In doing so, the model could be tuned to be *aggressive*, on one extreme, or to be *compliant*, on the other. An aggressive model could be used where it would be required for the model to interrupt the users or to compete for the turn in general (it may be necessary for the system to interrupt a user when a misconception is detected by the system on the part of the user, especially if the interaction between the system and the user is of a critical nature). On the other hand, a compliant model could be used where the system is intended to let the users have more initiative in the dialogue with the system responding with mostly acknowledgements, clarifications, and requests for information. In such a scenario, any attempt to speak by the user could be interpreted as an intention to take the turn.

Knowledge of turn-taking in conversations is essential for any spoken natural language interface to be *natural* in the true sense of the word. Not only will turn-taking knowledge improve the naturalness (by making the interactions as close to human conversations as possible) of an interface, but it will also lead to increased efficiency. The interface and the users will not have to rely on pauses (or other crude turn-yielding cues) to signal turn-completions, thereby making the whole interaction less time consuming.

Appendix A

Transcription Conventions

The transcription system, for capturing the auditory details of conversation, that is reported here is based on the system designed by Gail Jefferson (reported in Sacks et al, 1974).

It makes use of basic English orthography (as opposed to a phonemic system) and so sections of transcribed data can be read in much the same way as the basic text. Such material is, however, as different from the rest of the text as the statistical tables found in many journal articles. Both comprehension and evaluation of such data require that the material be attended to in quite specific ways. We will outline only those distinctions that will be necessary for our analysis. (The system transcribes phenomena like the gaze of the conversants which is not relevant to our analysis as we are dealing with telephone conversations).

To facilitate understanding, symbols and key aspects of their meaning will appear in boldface. The following dialogue fragments include some of the conventions that are most important to our analysis:

(17) *C-5C*

U: hhh on friday I also won't be at work during the hours from 7 30 till 12 n::oon {short}
1

Two colons (1) within a word indicates that the articulation of the **sound for the word is noticeably lengthened.**

(18) *C-4A*

S: certainly {short} what changes would you like to >make {short}
2

A **greater than sign** (2) preceding a word indicates that the **word following it was spoken at a noticeably lower volume.**

(19) *C-4A*

S: and and that's all/ {short}
3

A **slash** (3) following a word indicates that the word was spoken with a **rising pitch.**

(20) *C-4B*

U: ok = thank you = very much
4 4

An **equals sign** (4) between two words indicates that **the words were spoken without any pause in between (contiguously).**

(21) *C-7C*

U: so I {short} um {short} how do I do that {short} do I::
[ ]
5 6

S: do// you want
7

us = do you want to go ahead now/ {short}

An **opening square bracket** (5) indicates **the beginning of overlapping talk.**

A **closing square bracket** (6) indicates **the end of overlapping talk.**

Two slashes following a word (7) indicate that **the word was spoken with a rising pitch and a rising volume.**

(22) *C-8A*

S: hhh o::k {short} I've made that change {long}
8 9

U: I'm sorry = just a moment {short} ok/

The word **short** within brackets (8) indicates a **short pause,** and the word **long** within brackets (9) indicates a **long pause,** roughly twice the duration of a short pause.

(23) *C-3D*

S: sure = ( ) um {short} all you need to do is ask for um to cancel
10

call >waiting {short}

A **blank within parentheses** (10) indicates that the transcriber was **not able to recover what was said.**

(ours)

12 is a hell of a >discussion

(this)

11

Words within parentheses (11) indicate a **possible hearing. Two sets of parentheses containing words** (12) show that **alternative hearings** are possible. The marking of multiple hearings might indicate either disagreement among cotranscribers, agreement to both possibilities by cotranscribers, or double hearings by a single transcriber.

In the entire transcriptions, the use some of text-based conventions like punctuation marks and capital letters have been avoided.

Appendix B

Transcriptions

C-1A

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| I would like to a::dd {short} a number |syn| a name |syn| to my speed call d::irectory |syn|
S: certainly |syn| |grn| what would you like t::o a::dd |syn| |ques| {short}
U: hhh t::he number is 7 2 7 {short} 7 5 0 7 |syn| {short}
S: 7 2 7 {short} 7 5 0 7 |syn| |grn|
U: y::es |syn| |grn| {short}
S: ok= |syn| |grn| =and the n::ame/ |syn| {short}
U: hhh mike |syn| {short} m i k e |syn| {short}
S: mike |syn| |grn| {short} ok= |syn| |grn| =I've made that change |syn| {short}
U: thank you |syn| |thn| {long}
S: byebye |syn| |bye|
U: bye |syn| |bye|
{hang up}

C-1B

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| hhh I want to make a long distance c::all |syn| {short} but I don't want to be interrupted |syn| {short} during the c::all |syn|
S: o::k |syn| |grn| {short}
U: and I need to know how t::o {short} blank out the other calls coming >in |syn| {short}
S: hhh ok= |syn| |grn| =you can do that |syn| using um cancel call w::aiting |syn|
U: {mumble} {short} pardon me/ |syn| {short}
S: you can do that |syn| using cancel call w::aiting |syn| {short} before that you dial the call that you don't want to be i::nterrupted |syn| {short}
U: um hmm |syn| |grn|
S: you can ask for call waiting |syn| to be cancelled |syn| {long}
U: who do I ask |syn| |ques|
S: just {short} you can ask me |syn| >we'll we'll do it |syn| for >you |syn|
U: oh alright |syn| |grn| {short}
S: so that's what you want |syn| >done |syn| {short}
U: y::es |syn|
S: ok= |syn| |grn| =that is >that's now done |syn| {long}
U: alright |syn| |grn| {short}
S: anything can I help you |syn| with anything else |syn| |ques|
U: [ ] no// |syn| that's it |syn| {short}
S: ok= |syn| |grn| =thank you= |syn| |thn| =for using the >service |syn| |thn|
U: y::es |syn| |grn|
S: byebye |syn| |bye|
{hang up}

C-1C

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| hhh I need to make {short} a few changes |syn| in my weekly call forwarding schedule |syn| {short}
S: certainly |syn| |grn| hhh what can I change |syn| for >you |syn| |ques|
U: on m::onday {short} I have a meeting |syn| from 2:: in the afternoon |syn| until 4::
[ ]
S: um hmm |syn| |grn|
|syn| {short}
S: o::k |syn| |grn|
U: and those numbers {short} or they should c::all 7 7 6 {short} 1 2 3 4::= |syn| =during that >t::ime |syn|
S: ok= |syn| |grn| =so that's 7 7 6 {short} 1 2 3 4 |syn| hhh between 2 and 4 |syn| on >monday |syn|
U: y::es |syn| |grn| {short}
S: fine |syn| |grn| {short}
U: then on f::riday hhh in the m::orning {short} from 7 30 until n::oon {short} I will
[ ]
S: ok |syn| |grn|
be {short} at 3 5 6 {short} 4 9 3 0:: |syn| {short}
S: o::k |syn| |grn| {long} that's friday 7 30 to noon |syn| {short} and the number is 3 5 6 {short} 4 9 3 0 |syn| {short}
U: that's right |syn| |grn| {short}
S: I have that |syn| |grn|
U: and then {short} in the a::fternoon {short} through the whole time period {short} that I have scheduled through 6 o'clock {short} I will be at 3 3 9 {short} 2 3 2 3 |syn| {short}
S: ok= |syn| |grn| =hhh so is that every a::fternoon/ |syn| |ques|
U: um just friday >afternoon |syn|
[ ]
S: just// friday |syn| |grn| ok= |syn| |grn| =hhh and that was from n::oon= |syn| =until 6 >o'clock |syn| {short}
U: hhh r::ight |syn| |grn| {short}
S: hhh and that's 3 3 9 {short} 2 3 2 3 |syn|
U: y::es |syn| |grn| {short}
S: o::k |syn| |grn| {short}
U: that's >it |syn| >then |syn|
[ ]
S: that's all the changes |syn| {short}
U: yes |syn| |grn|
S: ok= |syn| |grn| =( ) so we've updated the >schedule |syn| {short}
U: thank you |syn| |thn| {short}
S: byebye |syn| |bye|
U: um bye |syn| |bye|
{hang up}

C-1D

S: hello |syn| this is southwestern bell's >phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| I:: would like to make some changes |syn| on my speed calling directory |syn| {short}
S: o::k |syn| |grn| {short}
U: hhh u::nder {short} mrs p::opplestein {short} and that'll either be under be m or p:: |syn| {short}
S: ok= |syn| |grn| =>let me get that out |syn| |hold| {long} yes |syn| I have that |syn| {short}
U: hhh I want to {short} c::hange mrs popplestein |syn| to t::he {short} name p::oopsie= |syn| ={laughter} hhh and that's capital p o o p s i e:: |syn| {short}
S: o::k |syn| |grn| {long}
U: {mumble} hhh alright |syn| and then hhh I need to i::nclude or add on another
[ ]
S: yes |syn| |grn|
name |syn| {short} its j::oe's a::utobody |syn| {short}
S: joe's autobody |syn| |grn|
U: y::es |syn| |grn|
S: o::k |syn| |grn| {long} o::k |syn| |grn| {short}
U: and the number for that is 2 2 5:: {short} 1 3 2 >0:: |syn| {short}
S: 2 2 5 {short} 1 3 2 >0 |syn| |grn|
U: y::es |syn| |grn| {short}
S: o::k |syn| |grn| {short} I've made that
[ ]
U: and// then the third c::hange |syn| hhh I want the name j::ane
[ ]
S: yes |syn| |grn|
{short} d::eleted |syn| {short} j a n >e |syn| {short}
S: ok |syn| |grn| that's done |syn| {short}
U: yes= |syn| |grn| =that's it |syn| >then |syn| {short}
S: thank you= |syn| |thn| =for using using the >service |syn| |thn| {short}
U: thank you |syn| |thn|
{hang up}

C-2B

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| um {short} I have my um calls on a forwarding s::chedule/ |syn| {short}
S: y::es |syn| |grn|
U: hhh but I need to change some things |syn| that are on >it |syn| {short}
S: ok= |syn| |grn| =what would you like to >c::hange |syn| |ques| {short}
U: hhh o::k |syn| |grn| {short} um {short} on monday hhh I:: am going to be {short} from 2 to 4:: in the afternoon {short}
S: o::k |syn| |grn| {short}
U: at 7 7 6 {short} 1 2 3 >4 |syn| {short}
S: 7 7 6 {short} 1 2 3 4 |syn| |grn| {short}
U: um hmm |syn| |grn| {short}

S: o::k |syn| |grn| {short}
U: hhh on f::riday hhh I'm going to be at 3 5 6 {short} 4 9 3 0 |syn| hhh um {short} from 7 30 to n::oon |syn| {short}
S: o::k |syn| |grn| {short} that's 3 5 6 {short} 4 9 3 >0 |syn| |grn| {short}
U: um hmm= |syn| |grn| =hhh 7 30 to >noon= |syn| =and on {short} >let's see |syn| {long} hhh a::nd {short} I'll be at my grandma's |syn| let's see |syn| the number is 3 3 9 {short} 2 3 2 3 |syn| {short}
S: o::k |syn| |grn|
U: um till 8:: {short} >on friday |syn| {short} >night |syn| {short}
S: ok= |syn| |grn| =so >that's friday |syn| {short} um until 8 |syn| {short} when >do you want that |syn| to s::tart |syn| |ques|
U: from noon |syn| from noon |syn| until 8 |syn| {short}
S: from noon |syn| till 8 |syn| |grn| {short} o::k= |syn| |grn| =I have t::hose |syn| {short}
U: hhh ok |syn| |grn| very good |syn| thank you |syn| |thn| {short}
S: thank you |syn| |thn|
U: um hmm= |syn| |grn| =bye |syn| |bye|
{hang up}

C-2C

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| I need to make some changes |syn| in my speed d::irectory |syn| {short}
S: o::k |syn| |grn| {short}
U: hhh
[ ]
S: what// changes would you like |syn| to >make |syn| |ques|{short}
U: hhh well {short} there's an entry |syn| that's under p::opplestein/ |syn| {long}
S: yes= |syn| |grn| =I have that |syn| {short}
U: hhh I need to change that to p::oopsie |syn| {short}
S: o::k |syn| |grn| making that change now |syn| {short}
U: thank you |syn| |thn| {short}
S: o::k |syn| |grn| {long}
U: o::k |syn| |grn| um I also have my um {short} joe's autobody/ |syn| {short}
S: y::es |syn| |grn| {short}
U: hhh ok= |syn| |grn| =that number needs to b::e {short} 2 2 5 {short} 13 20 |syn| {short}
S: ok= |syn| |grn| =2 2 5 {short} 1 3 2 >0 |syn| |grn| {short}
U: um hmm= |syn| |grn| =hhh and I also need to delete um {short} the entry for j::ane |syn| {short}
S: o::k= |syn| |grn| =that's now deleted |syn| {short}
U: ok |syn| |grn| very good |syn| thank you= |syn| |thn| =very much |syn| |thn|
S: thank you |syn| |thn|
U: um hmm |syn| |grn|
{hang up}

C-2D

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| I need to make a change |syn| to my speed d::irectory/ |syn| {short}
S: o::k |syn| |grn| {short}
U: actually I need to make an >addition= |syn| =hhh um {short} um I need to put m::ike |syn| {short}
S: ok= |syn| |grn| =I have mike |syn| |grn| {short}
U: and it's 7 2 7 {short} 7 5 0 7 |syn| {short}
S: 7 2 7 {short} 7 5 0 >7 |syn| |grn| {short}
U: um hmm |syn| |grn|
S: ok= |syn| |grn| =that's now added |syn| {short}
U: hhh ok |syn| |grn| I also need to call him |syn| {short}
S: fine= |syn| |grn| =I'll go ahead and call him |syn| for >you |syn|
U: thank you |syn| |thn|
{ring}
U: thank you |syn| |thn|
{hang up}

C-3A

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques|
U: yes |syn| I'd like to have my calls forwarded |syn| from my home |syn| to my o::ffice/ |syn| {short}
S: certainly= |syn| |grn| =let me just get that |syn| for >you |syn| |hold| {long} when would you like >this |syn| |ques| {short}
U: hhh um monday through friday |syn| from 7 30 in the morning |syn| till 6 >p::m |syn| {short}
S: o::k |syn| |grn| {long}
U: at the other times I:: want to receive the calls at home |syn| ok/ |syn| {short}
S: ok =|syn| |grn| =I have that |syn| |grn|
U: ok= |syn| |grn| =on monday I um have a meeting |syn| from 2 |syn| to 4:: |syn| {long}
S: o::k |syn| |grn|
U: and I'd like to be reached at 7 7 6 {short} 1 2 3 4:: |syn| {short}
S: 7 7 6 {short} 1 2 3 4 |syn| |grn|
U: um hmm |syn| |grn| {short}
S: o::k |syn| |grn|
U: and on friday I'd like to be reached at 3 5 6:: {short} 4 9 3 0 |syn| {short}
S: 3 5 6 {short} 4 9 3 0 |syn| |grn| {short}
U: from 7 30 am |syn| until >n::oon |syn| {short}
S: 7 30= |syn| |grn| =till >noon |syn| |grn| {short} o::k |syn| |grn|
U: [ ] ok= |syn| |grn| =and um {long} hhh till 8 um from about noon till 8 pm on friday I'll be at this >number |syn| {short}
S: hhh o::k= |syn| |grn| =could you give me that number |syn| p::lease |syn| |ques|
U: 3 3 9 {short} 2 3 2 3 |syn| {short}
S: 3 3 9 {short} 2 3 2 3 |syn| |grn|
U: um hmm |syn| |grn| {short}
S: ok |syn| |grn| {long} were there any other c::hanges/ |syn| |ques|

U: no |syn| that'll be a::ll |syn|
S: ok |syn| |grn| fine |syn| |grn|
[ ]
U: ok//= |syn| |grn| =thank you |syn| |thn| {short}
S: thank you |syn| |thn|
{hang up}

C-3C

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques|
U: yes |syn| I'd like to make an um entry with my speed call directory |syn| p::lease |syn| {short}
S: o::k |syn| |grn| {short}
U: the name is m::ike |syn| {long}
S: got that |syn| |grn|
U: and the number is 7 2 7 {short} 7 5 0 >7 |syn| {short}
S: 7 2 7 {short} 7 5 0 7 |syn| |grn|
U: um hmm= |syn| |grn| =and >um {short} when I {short} call and ask for mike >that'll be the number that I get |syn| is that correct/ |syn| |ques|
S: that's correct |syn|
U: o::k |syn| |grn| {short} thank you= |syn| |thn| =very much |syn| |thn| {short}
S: byebye |syn| |bye|
U: bye |syn| |bye|
{hang up}

C-3D

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: hhh yes |syn| um {short} I would like to know if you could give me some information |syn| about um {short} how to hhh not have phone calls ringing |syn| while I'm making another phone c::all |syn| {short} like I have call waiting |syn|
[ ]
S: {mumble}
{mumble} {long}
S: hello/ |syn|
U: I'm s::orry |syn| {short}
S: ok |syn| |grn|
U: um {short}
S: do you want t::o {short} um {short} to be to be able to make a call |syn| without being interrupted/ |syn| |ques|
U: yes |syn| um hmm |syn| |grn|
S: ok |syn| |grn|
U: could you tell me how to go about doing that= |syn| |ques| =please |syn| |ques|
S: sure |syn| ( ) um {short} all you need to do is ask for um to cancel call >waiting |syn| {short}
U: ok |syn| |grn|
S: to ask me to cancel call waiting |syn| b::efore hhh you want to make the >phone >c::all |syn|
U: ok= |syn| |grn| =is that all/ |syn| |ques| {short}

S: y::es |syn| that's all you need |syn| to do |syn|
[ ]
U: ok |syn| |grn| {short} great |syn| {short} thank you= |syn| |thn| =very much |syn| |thn| {short}
S: thank you |syn| |thn|
U: um hmm= |syn| |grn| =bye |syn| |bye|
{hang up}

C-4A

S: hhh hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques| {short}
U: yes |syn| I'd like to um {short} make some changes |syn| on my speed calling |syn| d::irectory |syn| {short}
S: certainly |syn| |grn| {short} what changes would you like to >make |syn| |ques|{short}
U: um um on the first one {short} I would like to c::hange hhh um {short} p::opplestein |syn| {short}
S: o::k |syn| |grn|
U: to p::oopsie |syn| {short}
S: ok |syn| |grn| {short}
U: and would you like me to {mumble} sorry just a minute |syn| |hold| {long} >13 >20 ok= |syn| |grn| =hhh um ok= |syn| |grn| =on the first one I would like to change Popplestein |syn| to p::oopsie |syn| {short}
S: ok= |syn| |grn| =I have that |syn| |grn| {short}
U: hhh and um I would like to c::hange joe's a::utobody |syn| {short}
S: o::k= |syn| |grn| =you want to add {short} joe's autobody/ |syn| {long}
U: well I would like to change the >number |syn| {short}
S: ok= |syn| |grn| =what number do you want to change >that >to |syn| |ques|
U: um the number {short} I would like to c::hange would be 4 3 2 17 69 |syn| hhh I would like to change that to {short} 2 2 5 {short} 1 3 2 0:: |syn| {short}
S: o::k |syn| |grn| I'll put that >in= |syn| =2 2 5 {short} 1 3 2 >0 |syn| |grn|
U: right= |syn| |grn| =hhh a::nd um {long} hhh I would no long I would like to have the um entry j::ane um {short} erased |syn| from >my directory |syn| my speed call directory |syn| >please |syn|
S: o::k= |syn| |grn| =that I've deleted that |syn| {short}
U: a::lright |syn| |grn| {short}
S: and and that's all/ |syn| {short}
U: that is it |syn|
S: ok |syn| |grn|
U: thank you= |syn| |thn| =ma'am |syn| |thn| {short}
S: thank you |syn| |thn|
U: bye |syn| |bye|
{hang up}

C-4B

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques|
U: yes |syn| operator |syn| I would like t::o um {short} make an entry under m::ike |syn| {short}
S: o::k |syn| |grn|

U: and I would like to dial him |syn| by name |syn| {short}
S: that's mike/ |syn| |grn| {short}
U: yes= |syn| |grn| =hhh >and {short} mike's number is 7 2 7 hhh 7 5 0 7 |syn| {short}
S: 7 2 7 {short} 7 5 0 7 |syn| |grn|
U: right= |syn| |grn| =and I would like to be able to dial him |syn| by by mike |syn|
S: ok =|syn| |grn| =that is now set up |syn| for >you |syn|
U: ok =|syn| |grn| =thank you= |syn| |thn| =very much |syn| |thn|
S: thank you |syn| |thn|
{hang up}

C-4C

S: hello |syn| this is southwestern bell's phone service |syn| can I help you |syn| |ques|
U: yes |syn| operator |syn| hhh um {short} I would like um {short} hhh I need to call my mom long distance |syn| on a frequent b::asis |syn| {short} and I don't want to
[ ]
S: ( )
waste my time |syn| or money |syn| um hhh answering other calls |syn| while I'm talking to her |syn| and I'm unsure how to how to do t::his |syn| {short} um {long}
[ ]
S: ( )
yes she's here |syn| {short}
S: hello/ |syn| {short}
U: y::es |syn| {short}
S: hhh o::k= |syn| |grn| =um {short} s::o what is it= |syn| |ques| =that you want= |syn| |ques| =>p::lease |syn| |ques| {short}
U: hhh um I'm thinking |syn| I want to have um I'm {short} I'm thinking that I want to have my call waiting taken off |syn| while I'm talking |syn| to my mom |syn| long d::istance |syn|
S: o::k |syn| |grn| all that you need to do when you want to dial is ask us to cancel call waiting |syn| {short} before >you make the call |syn| and we'll do >that |syn| {short}
U: ok= |syn| |grn| =so we'll just {short} so all I'll have to do is just call y::ou |syn| {short}
S: that's
[ ]
U: hhh and and say >um for this um {short} for the time being |syn| just take it o::ff |syn| {short}
S: that's right |syn| |grn|
U: that's i::t/ |syn|{short}
S: that's it |syn| {short}
U: hhh alright= |syn| |grn| =can y::ou um do that for me= |syn| |ques| =now |syn| |ques|
S yeah |syn| {short} we'll turn it off |syn|
[ ]
U: thank you |syn| |thn| {short}
S: byebye |syn| |bye| {short}
U: bye |syn| |bye|
{hang up}