Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 25 (2011) 601–634
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Turn-taking cues in task-oriented dialogue☆

Agustín Gravano[a,b,*], Julia Hirschberg[c]

[a] *Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina*
[b] *Laboratorio de Investigaciones Sensoriales, Hospital de Clínicas, Universidad de Buenos Aires, Argentina*
[c] *Department of Computer Science, Columbia University, New York, NY, USA*

Received 1 April 2010; received in revised form 18 August 2010; accepted 4 October 2010
Available online 20 October 2010

Abstract

As interactive voice response systems become more prevalent and provide increasingly more complex functionality, it becomes clear that the challenges facing such systems are not solely in their synthesis and recognition capabilities. Issues such as the coordination of turn exchanges between system and user also play an important role in system usability. In particular, both systems and users have difficulty determining when the other is taking or relinquishing the turn. In this paper, we seek to identify turn-taking cues correlated with human–human turn exchanges which are automatically computable. We compare the presence of potential prosodic, acoustic, and lexico-syntactic turn-yielding cues in prosodic phrases preceding turn changes (SMOOTH SWITCHES) vs. turn retentions (HOLDS) vs. BACKCHANNELS in the Columbia Games Corpus, a large corpus of task-oriented dialogues, to determine which features reliably distinguish between these three. We identify seven turn-yielding cues, all of which can be extracted automatically, for future use in turn generation and recognition in interactive voice response (IVR) systems. Testing Duncan's (1972) hypothesis that these turn-yielding cues are linearly correlated with the occurrence of turn-taking attempts, we further demonstrate that, the greater the number of turn-yielding cues that are present, the greater the likelihood that a turn change will occur. We also identify six cues that precede backchannels, which will also be useful for IVR backchannel generation and recognition; these cues correlate with backchannel occurrence in a quadratic manner. We find similar results for overlapping and for non-overlapping speech.

Keywords: Dialogue; Turn-taking; IVR systems; Prosody

1. Introduction

Interactions with state-of-the-art interactive voice response (IVR) systems are often described by users as "confusing" and even "intimidating". As speech technology continues to improve, it is becoming clear that such negative judgments are not due solely to errors in the speech recognition and synthesis components. Rather, coordination problems in the exchange of speaking turns between system and user are a plausible explanation for part of the deficient user experience (Ward et al., 2005; Raux et al., 2006).

Currently the most common method for determining when the user is willing to yield the conversational floor consists in waiting for a silence longer than a prespecified threshold, typically ranging from 0.5 to 1 s (Ferrer et al., 2002). However, this strategy is rarely used by humans, who rely instead on cues from sources such as syntax, acoustics and

☆ This paper has been recommended for acceptance by Koichi Shinoda.

* Corresponding author at: Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina.
E-mail addresses: gravano@dc.uba.ar (A. Gravano), julia@cs.columbia.edu (J. Hirschberg).

0885-2308/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.10.003

prosody to anticipate turn transitions (Yngve, 1970). If such TURN-YIELDING CUES can be modeled and incorporated in IVR systems, it should be possible to make faster, more accurate turn-taking decisions, thus leading to a more fluent interaction. Additionally, a better understanding of the mechanics of turn-taking can be used to inform the output of IVR systems to produce turn-yielding cues at the end of the system's turn and to avoid producing such cues when the system intends to continue the turn.

Another source of difficulty for state-of-the-art IVR systems are backchannel responses uttered by the user. BACKCHANNELS are short expressions, such as *uh-huh* or *mm-hm*, uttered by listeners to convey that they are paying attention, and to encourage the speaker to continue (Duncan, 1972; Ward and Tsukahara, 2000). Particularly when users are expected to provide large amounts of information, such as lists or long descriptions, the ability for systems to produce backchannel responses should improve the coordination between the two parties, letting the user know that the system is still attending. To achieve this, a system must first be able to detect points in the user's input where it will be appropriate to produce backchannels. We hypothesize that such points may be indicated by the user's production of BACKCHANNEL-INVITING CUES, by which we mean simply indicators that a backchannel may felicitously be produced by the system. Conversely, when the user utters a backchannel during a system turn, current IVR systems typically interpret such input as a turn-taking attempt, or BARGE-IN, thus leading the system to stop and listen—the opposite of the user's intention. Therefore, knowing when to interpret user input as a backchannel should also be a valuable tool for IVR systems.

In this paper we provide new information on the mechanisms used in human–human conversation to signal the end of a turn and to identify situations in which a backchannel is appropriate. We believe that such information will be useful for designers of IVR systems, both for system output production and for user input recognition, in the following situations:

Q1. The system wants to keep the floor: how should it formulate its output to avoid an interruption from the user?
Q2. The system wants to keep the floor but to ensure that the user is paying attention: how should it produce output encouraging the user to utter a backchannel?
Q3. The system is ready to yield the floor: how should it convey this to the user?
Q4. The user is speaking but pauses: how can the system decide whether the user is giving up the turn?
Q5. The user is speaking: how does the system decide whether and when to produce a backchannel as positive feedback to the user?

In this paper, we examine potential turn-taking cues correlated with human–human turn exchanges which are automatically computable. We compare acoustic, prosodic and lexico-syntactic features of pause-separated phrases which precede turn changes, turn retentions, and backchannels from the interlocutor in overlapping and non-overlapping speech to see which features best distinguish among these different situations. We also test Duncan's (1972) often-cited but as yet unverified hypothesis that turn-yielding cues are linearly correlated with the occurrence of turn-taking attempts—i.e., the more turn-yielding cues that are present in a phrase, the more likely that phrase is to be followed by a turn change. We examine the same question for backchannels.

In Section 2 we discuss previous work on modeling turn-taking and on end-of-turn detection. In Section 3 we describe the corpus we use for our studies, the Columbia Games Corpus, a corpus of task-oriented human–human dialogues. We describe the turn-yielding cues and the backchannel-inviting cues in non-overlapping speech that we have identified as reliable cues in Sections 4 and 5. In Section 6 we extend this analysis to overlapping speech. We conclude in Section 7 and discuss future research.

We note that, in examining features of phrases that precede turn changes, retentions, and backchannels, we make no strong cognitive claims about speaker awareness that they are 'signalling' an end of turn or of listener awareness that they are being 'signalled'. Our goal is only to identify correlations between turn-taking behaviors and automatically extractable features of human–human conversation which can be used to inform production and recognition in IVR systems. However, we hope that our findings will also be of value in understanding some of the mechanisms of human–human conversation.

2. Previous research on turn-taking

In influential work in conversational analysis, Sacks et al. (1974) present a characterization of turn-taking in conversations between two or more persons. Based on their identification of fourteen "grossly apparent facts" about human

V: *and his knee was being worn/ okay/ wait/ it was bent/ that way/*

D: *I mean it's it's not like wine/ it doesn't taste like wine/ but it's*
W: *fermented/*
D: *white/ and milky/ but it's fermented/*

Fig. 1. Examples of syntactic completion points, indicated by slashes. Taken from Ford and Thompson (1996, p. 144).

conversation, such as "speaker change recurs" or "one party talks at a time", they propose a basic set of rules governing turn allocation: at every TRANSITION-RELEVANCE PLACE (TRP),

(a) if the current speaker (CS) selects a conversational partner as the next speaker, then such partner must speak next;
(b) if CS does not select the next speaker, then anyone may take the next turn;
(c) if no one else takes the next turn, then CS may take the next turn.

The authors do not provide a formal definition of TRPs, but conjecture that these tend to occur at syntactic "possible completion points", with intonation playing a decisive role.

More detailed discussion of the types of cues humans exploit for engaging in synchronized conversation has been addressed extensively in subsequent decades. In a descriptive study of a face-to-face dialogue in American English, Yngve (1970) observes that pausing in itself is *not* a turn-yielding signal, in clear opposition to the strategy used in most of today's IVR systems.

In a series of analyses of face-to-face American English conversations, Duncan (1972, 1973, 1974, 1975) and Duncan and Fiske (1977) conjecture that speakers display complex signals at turn endings, composed of one or more of six behavioral cues: (1) any phrase-final intonation other than a sustained, intermediate pitch level; (2) a drawl on the final syllable of a terminal clause; (3) the termination of any hand gesticulation; (4) a stereotyped expression like *you know*; (5) a drop in pitch and/or loudness in conjunction with such a stereotyped expression; (6) the completion of a grammatical clause. He further proposes that the likelihood of a turn-taking attempt by a listener increases linearly with the number of turn-yielding cues conjointly displayed. This work has been criticized for two reasons (Beattie, 1981; Cutler and Pearson, 1986). First, it lacks a formal, objective description of the cues observed; his data are merely his own subjective impressions. Second, the robustness of its statistical analysis is questionable. For example, while he reports a correlation of 0.96 ($p < 0.01$) between number of turn-yielding cues displayed and percentage of interlocutor turn-taking attempts, this computation is based on a very small sample size. As few as nine instances of the simultaneous display of five cues are reported; therefore, a small fluctuation in the data may change the results substantially. Nonetheless, Duncan is the first to posit the existence of complex turn-yielding signals combined such that, the more complex the signal, the higher the likelihood of a speaker change. This crucial finding has laid the groundwork for a number of subsequent investigations.

In one such study, Ford and Thompson (1996) examine two of Duncan's individual cues, GRAMMATICAL COMPLETION and intonation, and their correlation with speaker change in two naturally occurring conversations in American English. They define grammatical completion in terms of SYNTACTIC COMPLETION POINTS—those points at which an utterance could be interpreted as syntactically complete "so far" in the discourse context, independent of intonation or pause (see Fig. 1 for examples). For intonation, they consider a binary distinction between *final* (either rising or falling) or *non-final* (all other) pitch contours. They find that syntactic completion points operate together with a rising or falling final intonation as an important turn-yielding cue. Also, they show that, while almost all (98.8%) intonationally complete utterances are also syntactically complete,[1] only half (53.6%) of syntactically complete utterances are intonationally complete, thus highlighting the prominent role played by intonation in marking discourse and dialogue structure.

Wennerstrom and Siegel (2003) enrich Ford and Thompson's approach with a more precise definition of final intonation based on the system developed by Pierrehumbert (1980), a predecessor of the ToBI transcription framework (Pitrelli et al., 1994; Beckman and Hirschberg, 1994).[2] They use six phrase-final intonational categories: *high rise*

[1] Ford and Thompson use a perceptual definition of intonational unit by Du Bois et al. (1993): "a stretch of speech uttered under a single coherent intonation contour"; and rely on acoustic, prosodic and timing cues to manually identify unit boundaries, independently of syntax.

[2] See Section 3.3 for a full description of the ToBI framework.

(H–H% in the ToBI system), *low* (L–L%), *plateau* (H–L%), *low rise* (L–H%), *partial fall* (also L–L%),[3] and *no boundary*. They find *high rise* intonation to be a strong cue of turn finality, with 67% of its occurrences coinciding with turn shifts, followed by *low*, with 40%. The remaining four intonational categories strongly correlate with turn holds. Additionally, Wennerstrom and Siegel analyze the interaction between intonation and Ford and Thompson's syntactic completion, and report similar findings.

A potential problem of observational studies such as the ones presented above is that they only collect indirect evidence of turn-yielding cues, arising from the fact that conversational decisions are *optional*. A listener who intends to let the speaker continue to hold the floor may choose not to act on turn-yielding cues displayed by the speaker. Furthermore, when using corpora of spontaneous conversations, it is extremely difficult to obtain a balanced set of utterances controlling for the diverse features under study; e.g., utterance pairs from the same speaker, with the same syntactic and semantic meaning, but one half in turn-medial position and the other half in turn-final position. To address these issues, there have been several production and perception experiments aimed at replicating in the laboratory the turn-taking decisions made by speakers in normal conversation. In a typical production study, participants read or enact fabricated dialogues with controlled target utterances; in a typical perception study, subjects classify a set of utterances into turn-medial or turn-final according to the believed speaker's intentions. These settings give the experimenter a great amount of control over the experimental conditions.

For instance, Schaffer (1983) presents a perception study to compare non-visual turn-taking cues in face-to-face and non-face-to-face conversations in American English. She reports that syntactic and lexical information appears to be more useful to listeners than prosodic information in judging turn boundaries in both conditions. Also, listeners show a great amount of variability in their perception of intonation as a turn-yielding cue. In a production and perception study of turn-taking in British English, Cutler and Pearson (1986) obtain similar results: wide listener variability in perception of intonation as a turn-yielding cue. They also find a slight tendency to characterize a phrase-final "downstep in pitch" as a turn-yielding cue, and an "upstep in pitch" as a TURN-HOLDING CUE (that is, a cue that typically prevents turn-taking attempts from the listener). While this is contra Duncan's hypothesis for American English, it is not surprising to find differences between the two varieties in intonation.

In two perception experiments designed to study intonation and syntactic completion in British English turn-taking, Wichmann and Caspers (2001) find only mild support for Duncan's claim that both syntactic completion and anything but a high level tone work as turn-yielding cues. Again, it is important to note that it is reasonable to expect different dialects and cultures to have different turn-taking behaviors. Therefore, findings even for languages within the same group, like British vs. American English, may differ substantially.

Recent perception studies in Swedish by Hjalmarsson (2009) indicate the existence of an additive effect of turn-yielding and turn-holding cues on the judgment of turn finality by non-participating listeners: the higher the number of cues displayed on a speech stimulus, the higher the inter-subject agreement regarding whether a turn change has occurred after a given turn. Additionally, Hjalmarsson finds a similar effect on listeners' perceptions when turn-yielding and turn-holding cues are included in synthesized speech, a result with important implications for the development of IVR systems.

Backchannel-inviting cues – that is, events in the current speaker's speech that precede a backchannel response – have received less attention in the literature than turn-yielding cues, although they have been examined extensively for other purposes. They constitute a type of LINGUISTIC FEEDBACK in the conversational analysis literature and are also sometimes referred to as CONTINUERS, indicating that the current speaker should continue talking (Yngve, 1970; Duncan, 1972; Kendon, 1967; Schegloff, 1982; Jefferson, 1984). Novick and Sutton (1994) propose an alternative categorization of linguistic feedback in task-oriented dialogue, which is based on the structural context of exchanges, rather than on the characteristics of the preceding utterance, and include: (i) *other* → *ackn*, where an acknowledgment immediately follows a contribution by *other* speaker; (ii) *self* → *other* → *ackn*, where *self* initiates an exchange, *other* eventually completes it, and *self* utters an acknowledgment; and (iii) *self* + *ackn*, where *self* includes an acknowledgment in an utterance independently of *other*'s previous contribution. Mushin et al. (2003) study the prosody of acknowledgments in an Australian English Map Task corpus, finding that acknowledgments such as *okay* or *yeh* are often produced with a 'non-final' intonational contour and followed by speech by the same speaker continuing an intonational phrase. As

[3] The *partial fall* category is described as a "downward sloping pitch contour that subsided before reaching the bottom of the speaker's range" [p. 84], and corresponds to a special type of L–L% in the ToBI system called 'suspended fall' (Pierrehumbert, 1980).

part of a larger project on modeling discourse structure in American English, Jurafsky et al. (1998) examine utterances identified as backchannels in 1155 conversations in the Switchboard Corpus (Godfrey et al., 1992), finding that the lexical realization of the dialogue act is the strongest cue to its identity; e.g., backchannel is the preferred function for *uh-huh* and *mm-hm*. They also find that backchannels are shorter in duration, have lower pitch and intensity, and are more likely to end in a rising intonation than agreements. Two related studies on the automatic classification of dialogue act classification on a subset of the same corpus (Shriberg et al., 1998; Stolcke et al., 2000) also find that the disambiguation of backchannels and agreements can be improved by using duration, pause and intensity information as well as lexical identity. There is also considerable evidence that linguistic feedback does not take place at arbitrary locations in conversation; rather, it mostly occurs at or near TRPs (Sacks et al., 1974; Goodwin, 1981).

Ward and Tsukahara (2000) describe a region of low pitch lasting at least 110 ms as a backchannel-inviting cue. They show that, in a corpus of spontaneous non-face-to-face dyadic conversations in American English, 48% of backchannels follow a low-pitch region, while only 18% of such regions precede a backchannel response. In a corpus study of Japanese dialogues, Koiso et al. (1998) find that both syntax and prosody play a central role in predicting the occurrence of backchannels, and Cathcart et al. (2003) propose a method for automatically predicting the placement of backchannels in conversation, based on pause durations and part-of-speech tags, that outperforms a random baseline model.

Recent studies investigate ways of improving the turn-taking decisions made by IVR systems, by incorporating some of the features shown in previous studies to correlate with turn or utterance endings. Ferrer et al. (2002, 2003) present an approach for online detection of utterance boundaries in the ATIS corpus (Hemphill et al., 1990), combining decision trees trained with prosodic features (related mainly to pitch range, pitch slope and duration) and *n*-gram language models. They train different classifiers to detect utterance endings at different pausal durations (from 30 to 800 ms) and report that speaker-normalized pitch range and normalized syllable durations appear to be the strongest predictors of speaker change, although no detailed analysis of features is presented. Edlund et al. (2005) experiment with a hand-crafted rule for detecting utterance boundaries: if a long-enough pause follows a long-enough speech segment that does not end in a level pitch slope, then mark the pause as an utterance end. Speech and silence segments are determined using intensity-based voice activity detection; pitch slope levels are discretized from the pitch track into three categories: rising, falling and level tones. This simple end-of-utterance predictor significantly outperforms a silence-based baseline system in a corpus consisting of 5 min of read English speech produced by one German speaker.

Schlangen (2006) trains a set of machine learning classifiers to detect turn-medial and turn-final utterance boundaries at different pausal durations after the target word (0, 100, 250 and 500 ms) – an approach similar to Ferrer et al.'s – on a subset of English Switchboard. All four classifiers significantly outperform a simple majority-class baseline; furthermore, the performance increases monotonically with the pausal duration considered. Schlangen reports word-final pitch and intensity levels and *n*-gram based features as the most predictive ones. This work is continued by Atterer et al. (2008), who improve performance on classifying each word as utterance final or non-final, independent of subsequent pauses. They report lexico-syntactic information such as word and part-of-speech *n*-grams to be most powerful predictors, and intensity-based features as the most useful prosodic features.

Raux and Eskenazi (2008) present an algorithm to dynamically set the threshold used for determining that a silence follows a turn boundary, based on a number of features extracted from the preceding turns. These include discourse structure information, captured by the system's dialogue act immediately preceding the current user turn; semantic information, drawn from the interpretation of partial speech recognition results in the current dialogue context; prosodic features, such as pitch and intensity slope and mean computed over the final part of the current user turn; and timing features, such as the time elapsed since the beginning of the utterance and the number of pauses observed so far in the current user turn. Raux and Eskenazi report that each of these feature types provides information useful in predicting turn finality.

All of these studies attack the problem of automatically detecting TRPs by exploiting knowledge based on previous descriptive studies, such as those mentioned earlier in this section. They rely upon a handful of hypothesized turn-yielding cues, especially on those related to pitch and intensity levels, and on some notion of syntactic or semantic completion. However, the evidence that these hypothesized cues do indeed correlate reliably with turn shifts is still lacking, due to (a) the small size of the corpora considered in most previous descriptive studies, and (b) the reliance of such studies on subjective impressions rather than on objective measurements. For instance, even though Duncan's (1972) work hypothesizing a linear relation between the number of co-occurring turn-yielding cues and the likelihood of a turn-taking attempt has been frequently cited and discussed, no study has yet tested this hypothesis on large

corpora using objective measures of features and statistical techniques. Duncan and others' descriptions of turn-taking behavior often fail to define features precisely and objectively. Another deficiency in previous descriptive studies consists in failing to distinguish between three radically different turn-taking phenomena: turn switches, backchannels, and interruptions, all of which are usually collapsed into a single class of turn changes.

Our work addresses these open questions, investigating a larger and more varied set of potential turn-yielding and backchannel-inviting cues than previous descriptive studies, comparing phrases preceding turn changes, backchannels, and turn-retentions to see whether they can be reliably distinguished from one another by automatically extractable cues. We provide objective definitions of our features and detailed information about the predictive power of each cue type, to expand knowledge of human–human cues. Our study also tests Duncan's hypothesis that there is a linear correlation between turn-yielding and backchannel-inviting cues and the subsequent likelihood of a turn change or a backchannel. We also examine overlapping speech separately in these categories, comparing it to non-overlapping tokens. Our corpus is also larger than that of most previous studies, permitting more statistically robust results, and has been annotated to distinguish between the three main dialogical categories mentioned above: turn switches, backchannels and interruptions. It also involves conversational partners engaged in collaborative tasks with performance incentives to enhance participant engagement, naturalness and spontaneity.

3. The Columbia Games Corpus

The materials for all experiments in this study were taken from the Columbia Games Corpus, a collection of 12 spontaneous task-oriented dyadic conversations elicited from native speakers of Standard American English (SAE). The corpus was collected and annotated jointly by the Spoken Language Group at Columbia University and the Department of Linguistics at Northwestern University, as part of an ongoing study of prosodic variation in SAE.

In each of the 12 sessions, two subjects were paid to play a series of computer games requiring verbal communication to achieve joint goals of identifying and moving images on the screen. Each subject used a separate laptop computer and could not see the screen of the other subject. Subjects sat facing each other in a soundproof booth, with an opaque curtain hanging between them, so that all communication was verbal. The subjects' speech was not restricted in any way, and it was emphasized at the session beginning that the game was **not** timed. Subjects were told that their goal was to accumulate as many points as possible over the entire session, since they would be paid additional money for each point they earned.

3.1. Game tasks

Subjects were first asked to play three instances of a Cards game, where they were shown cards with one to four images on them. Images were of two sizes (small or large) and various colors, and were selected to contain primarily voiced consonants, which facilitates pitch track computation (e.g., *yellow lion*, *blue mermaid*), to improve pitch track computation. There were two parts to each Cards game, designed to vary genre from primarily monologue to dialogue.

In the *first part* of the Cards game, each player's screen displayed a stack of 9 or 10 cards (Fig. 2a). Player A was asked to describe the top card on her pile, while Player B was asked to search through *his* pile to find the same card, clicking a button when he found it. This process was repeated until all cards in Player A's deck were matched. In all cases, Player B's deck contained one additional card that had no match in Player A's deck, to ensure that she would need to describe all cards.

In the *second part* of the Cards game, each player saw a board of 12 cards on the screen (Fig. 2b), all initially face down. As the game began, the first card on one player's (the Describer's) board was automatically turned face up. The Describer was told to describe this card to the other player (the Searcher), who was to find a matching card from the cards on his board. If the Searcher could not find a card exactly matching the Describer's card, but *could* find a card depicting one or more of the objects on that card, the players could decide whether to declare a partial match and receive points proportional to the numbers of objects matched on the cards. At most three cards were visible to each player at any time, with cards seen earlier being automatically turned face down as the game progressed. Players switched roles after each card was described and the process continued until all cards had been described. The players were given additional opportunities to earn points, based on other characteristics of the matched cards, to make the game more interesting and to encourage discussion.

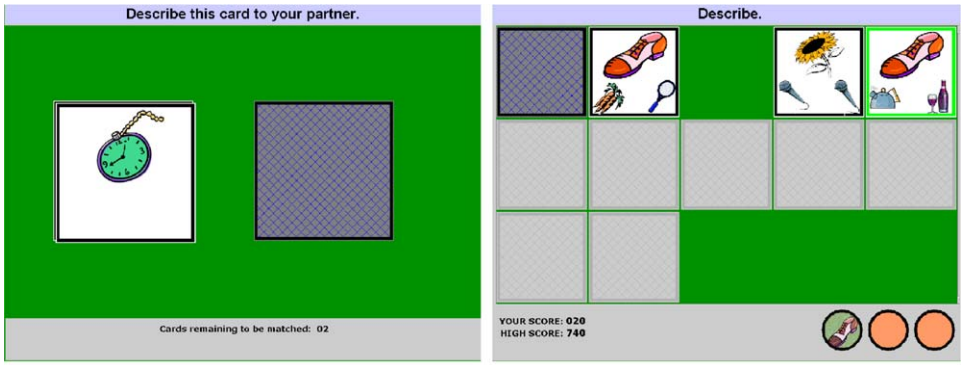

Fig. 2. Sample screens from the Cards games.

After completing all three instances of the Cards game, subjects were asked to play a final game, the OBJECTS GAME. As in the Cards game, all images were selected to have likely descriptions which were as sonorant as possible. In the Objects game, each player's laptop displayed a game board with 5–7 objects (Fig. 3). Both players saw the same set of objects at the same position on the screen, except for one (the TARGET). For the DESCRIBER, the target object appeared in a random location among other objects on the screen; for the FOLLOWER, the target object appeared at the bottom of the screen. The Describer was instructed to describe the position of the target object on her screen so that the Follower could move his representation to the same location on his own screen. After players negotiated what they believed to be their best location match, they were awarded 1–100 points based on how well the Follower's target location matched the Describer's.

The Objects game proceeded through 14 tasks. In the initial four tasks, one of the subjects always acted as the Describer, and the other one as the Follower. In the following four tasks their roles were inverted: the subject that played the Describer role in the initial four tasks was now the Follower, and vice versa. In the final six tasks, they alternated the roles with each new task.

3.2. *Subjects and sessions*

Thirteen subjects (six female, seven male) participated in the study, which took place in October 2004 in the Speech Lab at Columbia University. Eleven of the subjects participated in two sessions on different days, each time with a different partner. All subjects reported being native speakers of Standard American English and having no hearing impairments. Their ages ranged from 20 to 50 years (mean: 30.0; standard deviation: 10.9), and all subjects lived in

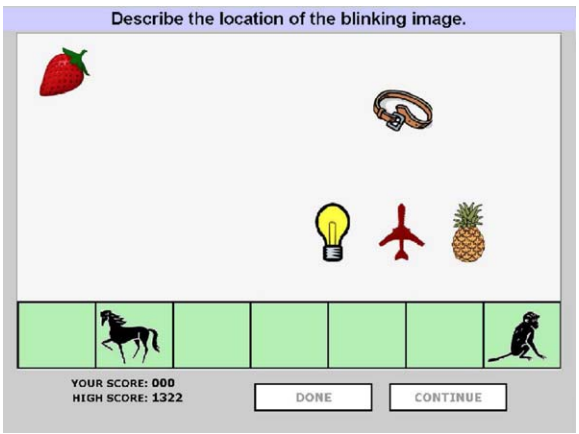

Fig. 3. Sample screen from the Objects games.

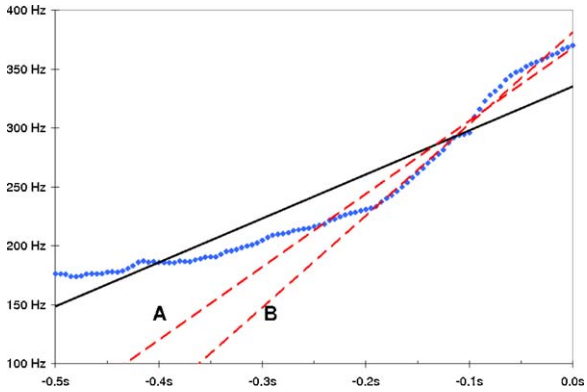

Fig. 4. Sample pitch track with three linear regressions: computed over the whole IPU (bold line), and over the final 300 ms (A) and 200 ms (B).

the New York City area at the time of the study. They were contacted through the classified advertisements website `craigslist.org`.

We recorded 12 sessions, each containing an average of 45 min of dialogue, totaling roughly 9 h of dialogue in the corpus. Of those, 70 min correspond to the first part of the Cards game, 207 min to the second part of the Cards game, and 258 min to the Objects game. On average, the first part of each Cards game took 1.9 min; the second part, 5.8 min; and the Objects game, 21.5 min. Each subject was recorded on a separate channel of a DAT recorder, at a sampling rate of 48 kHz with 16-bit precision, using a Crown head-mounted close-talking microphone. Each session was later downsampled to 16 kHz, and saved as one stereo wav file with one player per channel, and also as two separate mono wav files, one for each player.

Trained annotators orthographically transcribed the recordings of the Games Corpus and manually aligned the words to the speech signal, yielding a total of 70,259 words and 2037 unique words in the corpus. Additionally, self-repairs and certain non-word vocalizations, including laughs, coughs and breaths, were marked.

3.3. Feature extraction

We extracted a number of lexical, discourse, timing, acoustic and prosodic features from the speech, which we used in the experiments presented in the following sections. Part-of-speech tags were labeled automatically for the whole corpus using Ratnaparkhi's et al. (1996) maxent tagger trained on a subset of the Switchboard Corpus (Charniak and Johnson, 2001) in lower-case with all punctuation removed, to simulate spoken language transcripts. Each word had an associated POS tag from the full Penn Treebank tag set (Marcus et al., 1993), and one of the following simplified tags: noun, verb, adjective, adverb, contraction or other.

All acoustic features were extracted automatically for the whole corpus using the Praat toolkit (Boersma and Weenink, 2001). These include pitch, intensity, stylized pitch, ratio of voiced frames to total frames, jitter, shimmer, and noise-to-harmonics ratio. Pitch slopes were computed by fitting least-squares linear regression models to the F_0 data points extracted from given portions of the signal, such as a full word or its last 200 ms. This procedure is illustrated in Fig. 4, which shows the pitch track of a sample utterance (dotted line) with three linear regressions, computed over the whole utterance (solid black line), and over the final 300 and 200 ms ('A' and 'B' dashed lines, respectively).

We used a similar procedure to compute the slope of intensity and stylized pitch measurements. Stylized pitch curves were obtained using the algorithm provided in Praat: look up the pitch point p that is closest to the straight line L that connects its two neighboring points; if p is further than 4 semitones away from L, end; otherwise, remove p and start over. All normalizations were calculated using z-scores: $z = (x - \mu)/\sigma$, where x is a raw measurement to be normalized (e.g., the duration of a particular word), and μ and σ are the mean and standard deviation of a certain population (e.g., all instances of the same word by the same speaker in the whole conversation).

For the calculation of turn units, we define an INTER-PAUSAL UNIT (IPU) as a maximal sequence of words surrounded by silence longer than 50 ms. A TURN is a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor. Boundaries of IPUs and turns are computed automatically from the time-aligned transcriptions. A TASK in the Cards games corresponds to matching a card, and in the Objects games

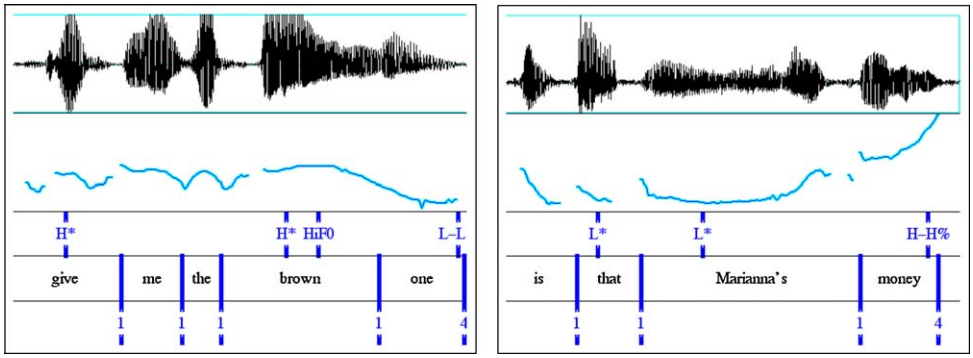

Fig. 5. A standard declarative contour (left), and a standard *yes–no* question contour. The top panes show the waveform and the fundamental frequency (F_0) track.

to placing an object in its correct position. Task boundaries were extracted from the logs collected automatically during the sessions, and subsequently checked by hand.

Intonational patterns and other aspects of the prosody were identified using the ToBI transcription framework. All of the Objects portion of the corpus (258 min of dialogue) and roughly one third of the Cards portion (90 min) were intonationally transcribed by trained annotators. The ToBI system consists of annotations at four time-linked levels of analysis: an ORTHOGRAPHIC TIER of time-aligned words; a TONAL TIER describing targets in the fundamental frequency (F_0) contour; a BREAK INDEX TIER indicating degrees of juncture between words; and a MISCELLANEOUS TIER, in which phenomena such as disfluencies may be optionally marked. The tonal tier describes events such as PITCH ACCENTS, which make words intonationally prominent and are realized by increased F_0 height, loudness, and duration of accented syllables. A given word may be accented or not and, if accented, may bear different tones, or different degrees of prominence, with respect to other words.

Five types of pitch accent are distinguished in the ToBI system for American English: two simple accents **H*** and **L***, and three complex ones, **L*+H**, **L+H***, and **H+!H***. An **L** indicates a low tone and an **H**, a high tone; the asterisk indicates which tone of the accent is aligned with the stressable syllable of the lexical item bearing the accent. Some pitch accents may be DOWNSTEPPED, such that the pitch range of the accent is compressed in comparison to a non-downstepped accent. Downsteps are indicated by the '!' diacritic (e.g., **!H***, **L+!H***). BREAK INDICES define two levels of phrasing: level 3 corresponds to Pierrehumbert's (1980) INTERMEDIATE PHRASE and level 4, to Pierrehumbert's INTONATIONAL PHRASE. Level 4 phrases consist of one or more level 3 phrases, plus a high or low BOUNDARY TONE (**H%** or **L%**) indicated in the tonal tier at the right edge of the phrase. Level 3 phrases consist of one or more pitch accents, aligned with the stressed syllable of lexical items, plus a PHRASE ACCENT, which also may be high (**H–**) or low (**L–**). For example, a standard declarative contour consists of a sequence of **H*** pitch accents ending in a low phrase accent and low boundary tone (**L–L%**); likewise, a standard *yes-no* question contour consists of a sequence of **L*** pitch accents ending in **H–H%**. These are illustrated in Fig. 5.

3.4. Turn-taking in the Games Corpus

The Games Corpus offers an excellent opportunity to study the turn-taking management mechanisms occurring in spontaneous conversation, and to provide answers to the research questions posited in Section 1. A superficial analysis of the corpus reveals it to be rich in all kinds of turn-taking phenomena, as all subjects became engaged in active conversation to achieve the highest possible performance in the various game tasks.

All conversations in the corpus are between two people collaborating on a common task, and take place with no visual contact between the participants. These conditions roughly replicate the typical settings of current telephone IVR systems, in which a person is assisted by a remote computer using natural speech over the telephone to perform relatively simple tasks, such as making travel reservations or requesting banking information.

When visual contact is permitted between the conversation participants, a whole new dimension of complexity is introduced to the analysis of turn-taking phenomena. For instance, eye gaze and hand gesticulation are known to be strong turn-taking cues (Duncan, 1972; Kendon, 1972; McNeill, 1992). When collecting the Games Corpus, visual

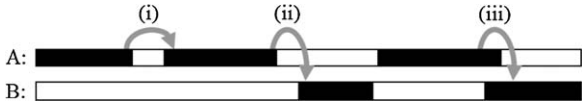

Fig. 6. Simple 3-way definition of turn exchanges. Black segments represent speech; white segments, silence. (i) Hold, (ii) change without overlap, and (iii) change with overlap.

contact was impeded by a curtain between the two participants, thus forcing all communicational to be verbal. The lack of visual contact allows us to effectively isolate audio-only cues, the central object of study in our experiments.

Finally, we take several steps to achieve results as general as possible—i.e., not true only for a specific set of speakers, but generalizable to a larger population. First, the corpus contains 12 conversations recorded from 13 different people. Second, the participants of each conversation had never met each other before the recording session. This allows us to avoid any potential communicational codes or behaviors arising from pre-existing acquaintances between the subjects, which are also beyond the scope of our study. Third, in the statistical studies presented in the following sections, we pay great attention to speaker variation. Specifically, for each result holding for all 13 speakers together, we report whether the same results holds for each individual speaker.

3.4.1. Labeling scheme

Our main research goal is to investigate the existence of acoustic, prosodic, lexical and syntactic turn-yielding and backchannel-inviting cues. That is, we search for events in the speech produced by the person holding the conversational floor that may cue the listener about an imminent turn yielding event, or that may instead invite the listener to utter a backchannel response. With this goal in mind, we need first to define various types of turn-taking phenomena in the corpus, which we will later analyze. For example, in our search for turn-yielding cues, we need to define and identify turn boundaries, to later compare turn-final utterances against turn-medial ones. In this section we consider a number of labeling systems used in previous work, and describe in detail the one we choose for our experiments.

In an approach adopted by several studies, all exchanges are collapsed into a single CHANGE category, defined as a transition from a turn by the participant currently holding the floor to a new turn by the other participant. Typically, this category is further subdivided into CHANGE WITH OVERLAP and CHANGE WITHOUT OVERLAP, depending on whether the two contributions have a non-empty temporal intersection, as shown in Fig. 6 (Koiso et al., 1998; Edlund et al., 2005, inter alia). The second main class in this approach is the HOLD category, defined as a transition between two adjacent IPUs within a turn by the same speaker. The change and hold categories are typically contrasted to look for turn-yielding cues, with the assumption that instances of the former are more likely to contain such cues than instances of the latter.

The main advantage of these simple binary and ternary distinctions is that they can be computed automatically from the speech signal: turn boundaries can be estimated using an energy-based silence detector, provided that each speaker has been recorded on a separate channel. In our case, this labeling system oversimplifies the problem, since we need to be able to differentiate phenomena such as backchannels and interruptions from regular turn changes. In other words, we need a finer grained categorization of speaker changes.

One such categorization is introduced by Ferguson (1977) for a study of behavioral psychology that investigates simultaneous speech and interruptions as measures of dominance in family interaction. Beattie (1982) revises Ferguson's system to a decision tree in a study of two political interviews comparing the turn-taking styles of former British Prime Ministers Jim Callaghan and Margaret Thatcher. Beattie reports an almost perfect inter-labeler agreement using this scheme, with a Cohen's κ score (Cohen, 1960) of 0.89.[4] We adopt a slightly modified version of Beattie's scheme, as depicted in Fig. 7. This system is better suited to our experiments on turn-yielding cues than those using only binary and ternary distinctions.[5]

[4] The κ measure of agreement above chance is interpreted as follows: 0 = none, 0–0.2 = small, 0.2–0.4 = fair, 0.4–0.6 = moderate, 0.6–0.8 = substantial, 0.8–1 = almost perfect.

[5] For example, it distinguishes two exchange types (SMOOTH SWITCH and OVERLAP) in which turn-yielding cues are likely to be present, given that a turn exchange occurs and the first speaker (i.e., the one originally holding the floor) manages to finish the utterance. The remaining three types (INTERRUPTION, PAUSE INTERRUPTION and BUTTING-IN) are less likely to contain turn-yielding cues, given that the first speaker is interrupted at arbitrary times.

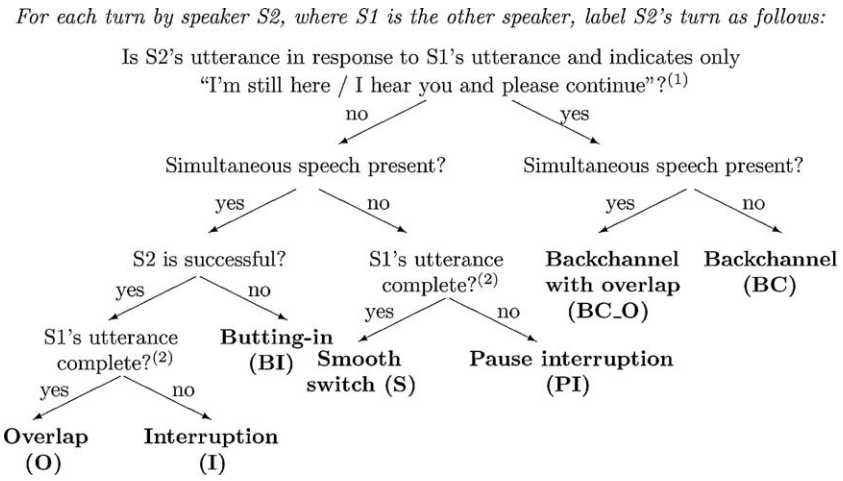

Fig. 7. Turn-taking labeling scheme used in the present study, with seven turn-taking categories.

Fig. 8. Simultaneous start: B_1 occurs in response to A_1, rather than A_2.

Backchannels play an important role in our research goals, but Beattie explicitly excludes them from his study. Therefore, we incorporate backchannels into our labeling scheme by adding the decision marked (1) at the root of the decision tree. All backchannels in the corpus were identified as part of a study of affirmative cue words (Gravano et al., 2007); thus, we use these labels, and annotators of turn-taking are not asked to make this decision. For the decision marked (2) in Fig. 7, we use Beattie's informal definition of utterance completeness: "Completeness [is] judged intuitively, taking into account the intonation, syntax, and meaning of the utterance" [p. 100]. The decision "Simultaneous speech present?" is placed high up in the tree, as it is pre-computed automatically based on the manual orthographic transcripts of the conversations. Additionally, we identify three cases that do not correspond to actual turn exchanges, and thus receive special labels:

- *Task beginning*: Turns beginning a new game task are labeled **X1**.
- *Continuation after a backchannel*: If a turn t is a continuation after a **BC** or **BC_O** from the other speaker, it is labeled **X2_O** if t overlaps the backchannel, or **X2** if not.
- *Simultaneous start*: Fry (1975) reports that humans require at least 210 ms to react verbally to a verbal stimulus. Thus, if two turns begin within 210 ms of each other, they are most probably connected to preceding events than to one another. In Fig. 8, A_1, A_2 and B_1 represent turns from speakers A and B. Most likely, A_2 is simply a continuation from A_1, and B_1 occurs in response to A_1. Thus, B_1 is labeled with respect to A_1 (not A_2), and A_2 is labeled **X3**.

Finally, all continuations from one IPU to the next within the same turn are labeled automatically with the special label **H**, for 'hold'.

Needless to say, the categories defined in this taxonomy are too broad to accommodate the wide spectrum of variation in human conversation. However, they are well suited for our turn-taking experiments, as they allow us to look for turn-yielding cues by contrasting the places where such cues are likely to occur (e.g., before smooth switches) against the places where they are not likely to occur (e.g., before holds or interruptions). Furthermore, more fine-grained distinctions, albeit closer to representing the full diversity of turn-taking events present in spontaneous dialogue, would have the cost of data sparsity, thus compromising the statistical significance of the results.

Table 1
Distribution of turn-taking labels in the Games Corpus.

Label	Count	Percentage
BC	553	6.8
BC_O	202	2.5
BI	104	1.3
I	158	1.9
O	1067	13.1
PI	275	3.4
S	3247	39.9
X1	1393	17.1
X2	449	5.5
X2_O	59	0.7
X3	590	7.3
?	37	0.5
Total	8134	100.0

Two trained annotators labeled the whole Objects portion of the corpus separately, with a Cohen's κ score (Cohen, 1960) of 0.913 corresponding to 'almost perfect' agreement.[6] Subsequently, we performed the following steps to correct potential labeling errors. Cases with dissimilar judgments were marked for revision and given back to one of the annotators (*ANN1*), without specifying the labels assigned by the other annotator (*ANN2*). *ANN1* corrected what he considered were errors in his labels, and the process was repeated for *ANN2*, who revised the remaining differences, again blind to *ANN1*'s choices. At the end of this process, the κ score improved to 0.9895. Given the high inter-labeler agreement obtained in the Objects portion of the corpus, the Cards portion was labeled by just one trained annotator. Table 1 shows the distribution of turn-taking labels in the entire corpus. Additionally, there are 8123 instances of 'hold' transitions (**H**) in the Games Corpus, as defined above.[7]

4. Turn-yielding cues

We begin our study of turn-taking in the Columbia Games Corpus by investigating turn-yielding cues—events from acoustic, prosodic or syntactic sources, inter alia, produced by the speaker when approaching the potential end of a conversational turn, that may be used by the listener to detect, or even anticipate, an opportunity to take the floor. We adopt the assumption proposed by Duncan (1972) that individually identifiable cues may be combined together to form a complex turn-yielding signal. As discussed in the previous sections, a number of non-visual turn-yielding cues have been hypothesized in the literature: any final intonation other than a sustained pitch level; a drawl on the final syllable of a terminal clause; a drop in intensity and pitch levels; stereotyped expressions such as *you know* or *I think*; and the completion of a grammatical clause. In this section we examine these cues in our corpus, and present results introducing two turn-yielding cues mentioned only rarely in the literature, related to voice quality (Ogden, 2002) and IPU duration (Cutler and Pearson, 1986). After considering individual cues, we describe how they combine to form a complex signal, and show the manner in which the likelihood of a turn switch increases with the number of cues present in such a signal.

[6] Note that this κ score does not include the identification of backchannels, performed by different annotators as described in Gravano et al. (2007).

[7] An analysis of the distribution of gap durations in smooth switches shows that 30% of such gaps are 200 ms or shorter, 63% are 200 ms to 2 s long, and the remaining 7% are longer than 2 s. In our study we considered using only a subset of the data, discarding smooth switches with too short or too long gaps. However, in both cases it is unclear what thresholds should be used. For too short gaps, even though Fry (1975) and other studies show that humans need roughly 200 ms to react to the occurrence of a verbal stimulus, they do not say anything about the opposite phenomenon: the reaction time to the termination of an ongoing stimulus, such as a conversational turn. Thus, the choice of such a threshold would be arbitrary. For very long gaps, it is even harder to set a threshold that will effectively separate speaker-shifts elicited by cues in the previous utterance from those that are not, since it would be difficult to be certain that, e.g., the speaker has not just taken a few seconds to think about his/her contribution. In fact, hesitations are common in our corpus, due to the nature of the game tasks. In consequence, we decided to take the simplest option of including all instances of smooth switches in our analysis.

Table 2
ToBI phrase accent and boundary tone for IPUs preceding **S** and **H**.

	S		**H**	
H–H%	484	(22.1%)	513	(9.1%)
[!]H–L%	289	(13.2%)	1680	(29.9%)
L–H%	309	(14.1%)	646	(11.5%)
L–L%	1032	(47.2%)	1387	(24.7%)
No boundary tone	16	(0.7%)	1261	(22.4%)
Other	56	(2.6%)	136	(2.4%)
Total	2186	(100%)	5623	(100%)

Our general approach consists in contrasting IPUs immediately preceding smooth switches (**S**) with those immediately preceding holds (**H**). We hypothesize that turn-yielding cues are more likely to occur before **S** than before **H**. It is important to emphasize the optionality of all turn-taking phenomena and decisions: for **H**, turn-yielding cues – whatever their nature – may still be present; and for **S**, they may be sometimes absent. However, we hypothesize that their likelihood of occurrence should be much higher before **S**. Finally, as mentioned above, we make no claims regarding whether speakers intend to produce turn-yielding cues, or whether listeners consciously perceive and/or use them to aid their turn-taking decisions.

4.1. Intonation

The literature contains frequent mention of the propensity of speaking turns to end in any intonation contour *other than* a plateau (a sustained pitch level, neither rising nor falling). We first analyze the categorical prosodic labels in the portion of the Columbia Games Corpus annotated using the ToBI annotations.

We tabulate the phrase accent and boundary tone labels assigned to the end of each IPU, and compare their distribution for the **S** and **H** turn exchange types, as shown in Table 2. A chi-square test indicates that there is a significant departure from a random distribution ($\chi^2 = 1102.5$, $d.f. = 5$, $p \approx 0$). Only 13.2% of all IPUs immediately preceding a smooth switch (**S**) – where turn-yielding cues are most likely present – end in a plateau ([!]H–L%); most of the remaining IPUs end in either a falling pitch (L–L%) or a high rise (H–H%). For IPUs preceding a hold (**H**) the counts approximate a uniform distribution, with the plateau contours being the most common; this supports the hypothesis that this contour functions as a TURN-HOLDING CUE (that is, a cue that typically prevents turn-taking attempts from the listener). The high counts for the falling contour preceding a hold (24.7%) may be explained by the fact that, as discussed above, taking the turn is optional for the listener, who may choose not to act despite hearing some turn-yielding cues. It is not entirely clear what the role is of the low-rising contour (L–H%), as it occurs in similar proportions before **S** and before **H**. Finally, we note that the absence of a boundary tone works as a strong indication that the speaker has not finished speaking, since nearly all (98%) IPUs without a boundary tone precede a hold transition.

Next, we examine four objective acoustic approximations of this perceptual feature: the absolute value of the speaker-normalized F_0 slope, both raw and stylized, computed over the final 200 and 300 ms of each IPU. The case of a plateau corresponds to a value of F_0 slope close to zero; the other case, of either a rising or a falling pitch, corresponds to a high absolute value of F_0 slope. As shown in Fig. 9, we find that the final slope before **S** is significantly higher than before **H** in all four cases. These findings provide additional support for the hypothesis that turns tend to end in falling and high-rising final intonations, and provide automatically identifiable indicators of this turn-yielding cue.

4.2. Speaking rate and IPU duration

Duncan (1972) hypothesizes a "drawl on the final syllable or on the stressed syllable of a terminal clause" [p. 287] as a turn-yielding cue, which would probably correspond to a noticeable decrease in speaking rate. We examine this hypothesis in our corpus using two common definitions of speaking rate: syllables per second and phonemes per second. Syllable and phoneme counts were estimated from dictionary lookup, and word durations were extracted from the manual orthographic alignments.

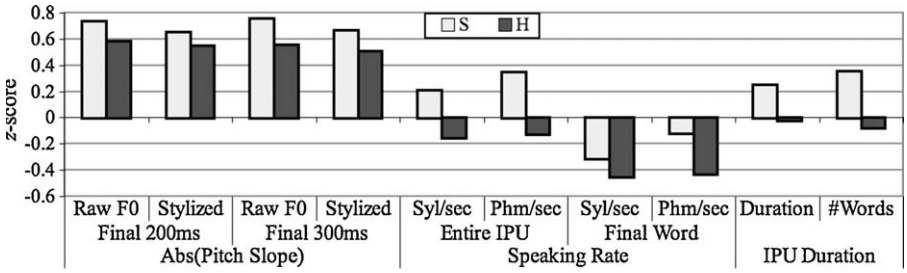

Fig. 9. Individual turn-yielding cues: intonation, speaking rate and IPU duration. In all cases, the difference between the two groups is significant (one-way ANOVA, $p < 0.01$).

Fig. 9 shows that the speaking rate is significantly slower over the final word than over the whole IPU both before **S** and before **H**. This result is in line with phonological theories that predict a segmental lengthening near prosodic phrase boundaries (Wightman et al., 1992). This finding may indeed correspond to the drawl or lengthening described by Duncan before turn boundaries. However, it seems to be the case – at least for our corpus – that the final lengthening tends to occur at all phrase final positions, not just at turn endings. In fact, Fig. 9 shows that both measures, computed over either the whole IPU or its final word, are significantly higher before **S** than before **H**, which indicates an *increase* in speaking rate before turn boundaries. In other words, these results indicate that the final lengthening is more prominent in turn-medial IPUs than in turn-final ones.

We also find that turn-final IPUs tend to be significantly longer than turn-medial ones, both when measured in seconds and in number of words (Fig. 9). This suggests that IPU duration could function as a turn-yielding cue, supporting similar findings in perceptual experiments by Cutler and Pearson (1986).

4.3. Acoustic cues

In the Columbia Games Corpus, we find that IPUs followed by **S** have a mean intensity significantly lower than IPUs followed by **H**, where intensity is computed over the IPU-final 500 and 1000 ms (see Fig. 10). Also, the differences increase toward the end of the IPU. This suggests that speakers tend to lower their voices when approaching potential turn boundaries, whereas they reach turn-internal pauses with a higher intensity.

Phonological theories conjecture a declination in the pitch level, which tends to decrease gradually within utterances and across utterances within the same discourse segment as a consequence of a gradual compression of the pitch range (Pierrehumbert and Hirschberg, 1990). For conversational turns, then, we would expect to find that speakers tend to lower their pitch level as they reach potential turn boundaries. This hypothesis is verified by the dialogues in our corpus, where we find that IPUs preceding **S** have a significantly lower mean pitch than those preceding **H** (Fig. 10). In consequence, pitch level may also work as a turn-yielding cue.

Next we examine three acoustic features, jitter, shimmer and noise-to-harmonics ratio (NHR), which have been associated with the perception of voice quality (Eskenazi et al., 1990; Kitch et al., 1996; Bhuta et al., 2004). Jitter and shimmer correspond to variability in the frequency and amplitude of vocal-fold vibration, respectively; NHR is the energy ratio of noise to harmonic components in the voiced speech signal. We compute jitter and shimmer only over voiced frames for improved robustness. Fig. 10 summarizes the results for these features, computed over the IPU-final

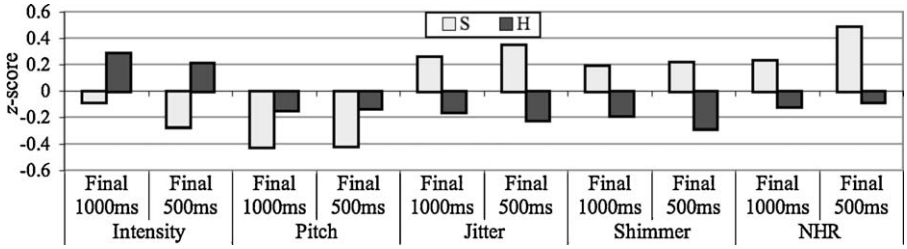

Fig. 10. Individual turn-yielding cues: intensity, pitch and voice quality. In all cases, the difference between the two groups is significant (one-way ANOVA, $p < 0.01$).

Table 3
Twenty-five most frequent final bigrams preceding each turn-taking type.

	S	Count	Perc.	**H**	Count	Perc.
1	*okay*	241	7.4	*okay*	402	4.9
2	*yeah*	167	5.1	*on top*	172	2.1
3	*lower right*	85	2.6	*um*	136	1.7
4	*bottom right*	74	2.3	*the top*	117	1.4
5	*the right*	59	1.8	*of the*	67	0.8
6	*hand corner*	52	1.6	*blue lion*	57	0.7
7	*lower left*	43	1.3	*bottom left*	56	0.7
8	*the iron*	37	1.1	*with the*	54	0.7
9	*the onion*	33	1.0	*the um*	54	0.7
10	*bottom left*	31	1.0	*yeah*	53	0.7
11	*the ruler*	30	0.9	*the left*	48	0.6
12	*mm-hm*	30	0.9	*and*	48	0.6
13	*right*	28	0.9	*lower left*	46	0.6
14	*right corner*	27	0.8	*uh*	45	0.6
15	*the bottom*	26	0.8	*oh*	45	0.6
16	*the left*	24	0.7	*and a*	45	0.6
17	*crescent moon*	23	0.7	*alright*	44	0.5
18	*the lemon*	22	0.7	*okay um*	43	0.5
19	*the moon*	20	0.6	*the uh*	42	0.5
20	*tennis racket*	20	0.6	*the right*	41	0.5
21	*blue lion*	19	0.6	*the bottom*	39	0.5
22	*the whale*	18	0.6	*I have*	39	0.5
23	*the crescent*	18	0.6	*yellow lion*	37	0.5
24	*the middle*	17	0.5	*the middle*	37	0.5
25	*of it*	17	0.5	*I've got*	34	0.4

500 and 1000 ms. For all three features, the mean value for IPUs preceding **S** is significantly higher than for IPUs preceding **H**, with the difference increasing towards the end of the IPU. Therefore, voice quality seems to play a clear role as a turn-yielding cue.

4.4. Lexical cues

Stereotyped expressions such as *you know* or *I think* have been proposed in the literature as lexical turn-yielding cues. However, in the Games Corpus we find that none of the most frequent IPU-final unigrams and bigrams, preceding either **S** or **H**, correspond to such expressions (Table 3 lists the 25 most frequent IPU-final bigrams). Instead, most of such unigrams and bigrams are specific to the computer games in which the subjects participated. For example, the game objects tended to be spontaneously described by subjects from top to bottom and from left to right, as shown in the following excerpt (pauses are indicated with #):

A: *I have a blue lion on top # with a lemon in the bottom left # and a yellow crescent moon in- # i- # in the bottom right*
B: *oh okay* [...]

In consequence, bigrams such as *lower right* and *bottom right* are common before **S**, while *on top* or *bottom left* are common before **H**. These are all task-specific lexical constructions and do not constitute stereotyped expressions in the traditional sense. Also very common among the most frequent IPU-final expressions are AFFIRMATIVE CUE WORDS—heavily overloaded words, such as *okay* or *yeah*, that are used both to initiate and to end discourse segments, among other functions (Gravano et al., 2007). The occurrence of these words does not constitute a turn-yielding or turn-holding cue *per se*; rather, additional contextual, acoustic and prosodic information is needed to disambiguate their meaning.

Affirmative cue words and game-specific expressions cover the totality of the 25 most frequent IPU-final bigrams listed in Table 3. Further down in the list, we find some rare uses of stereotyped expressions preceding smooth switches, all with only marginal counts: *I guess* (6 instances, or 0.18% of the total), *I think* (4), and *you know* (2). Notably, there were more instances of each of these expressions before holds: 6, 5 and 21, respectively, challenging the idea that the mere occurrence of these expressions works as a strong turn-yielding cue. As with affirmative cue words, more information from other sources seems to be necessary to disambiguate the meaning of these expressions.

While we do not find clear examples of lexical turn-yielding cues in our task-oriented corpus, we do find two lexical turn-holding cues: word fragments (e.g., *incompl-*) and filled pauses (e.g., *uh*, *um*). Of the 8123 IPUs preceding **H**, 6.7% end in a word fragment, and 9.4% in a filled pause. By contrast, only 0.3% of the 3246 IPUs preceding **S** end in a word fragment, and 1% in a filled pause. These differences are significant ($\chi^2 = 491.6$, $d.f. = 2$, $p \approx 0$) and suggest that, after either a word fragment or a filled pause, the speaker is much more likely to intend to continue holding the floor. This notion of disfluencies functioning as a turn-taking cue has been studied by Goodwin (1981), who shows that they may be used to secure the listener's attention at turn beginnings.

4.5. Textual completion

Several authors claim that some sort of completion independent of intonation and interactional import functions as a turn-yielding cue (Duncan, 1972; Sacks et al., 1974; Ford and Thompson, 1996; Wennerstrom and Siegel, 2003). Although some call this *syntactic completion*, all authors acknowledge the need for semantic and discourse information in judging utterance completion. Therefore, we choose the more neutral term TEXTUAL COMPLETION for this phenomenon. We manually annotated a portion of our corpus with respect to textual completion and trained a machine learning (ML) classifier to automatically label the whole corpus. From these annotations we then examined how textual completion labels relate to turn-taking categories in the corpus.

4.5.1. Manual labeling

In conversation, listeners judge textual completion incrementally and without access to later material. To simulate these conditions in the labeling task, annotators were asked to judge the textual completion of a turn up to a target pause from the written transcript alone, without listening to the speech. They were allowed to read the transcript of the full previous turn by the other speaker (if any), but they were not given access to anything after the target pause. These are a few sample tokens:

A: *the lion's left paw our front*
B: *yeah and it's th- right so the*

A: *and then a tea kettle and then the wine*
B: *okay well I have the big shoe and the wine*

A: —
B: *okay there is a belt in the lower right a microphone in the lower left*

We selected 400 tokens at random from the Games Corpus; the target pauses were also chosen at random. To obtain good coverage of the variation present in the corpus, tokens were selected in such a way that 100 of them were followed by a hold transition (**H**), 100 by a backchannel (BC), 100 by a smooth switch (**S**), and 100 by a pause interruption (PI). Three annotators labeled each token independently as either complete or incomplete according to these guidelines: *"Determine whether you believe what speaker B has said up to this point could constitute a complete response to what speaker A has said in the previous turn/segment. Note: if there are no words by A, then B is beginning a new task, such as describing a card or the location of an object."* To avoid biasing the results, annotators were not given the turn-taking labels of the tokens. Inter-annotator reliability is measured by Fleiss' κ at 0.814, which corresponds to the 'almost perfect' agreement category. The mean pairwise agreement between the three subjects is 90.8%. For the cases

Table 4
Mean accuracy of each classifier for the textual completion labeling task, using 10-fold cross validation on the training data.

Classifier	Accuracy
Majority-class ('complete')	55.2%
C4.5	55.2%
Ripper	68.2%
Bayesian networks	75.7%
SVM, RBF kernel	78.2%
SVM, linear kernel	80.0%
Human labelers (mean agreement)	90.8%

in which there is disagreement between the three annotators, we adopt the MAJORITY LABEL as our gold standard; that is, the label chosen by two annotators.

4.5.2. Automatic classification

Next, we trained a machine learning model using the 400 manually annotated tokens as training data to automatically classify all IPUs in the corpus as either complete or incomplete. For each IPU *u* we extracted a number of lexical and syntactic features from the beginning of the turn containing *u* up to *u* itself: the lexical identity of the IPU-final word (*w*); POS tags and simplified POS tags (N, V, Adj, Adv, other) of *w* and of the IPU-final bigram; number of words in the IPU; a binary flag indicating if *w* is a word fragment; size and type of the biggest (*bp*) and smallest (*sp*) phrase that end in *w*; binary flags indicating if each of *bp* and *sp* is a major phrase (NP, VP, PP, ADJP, ADVP); binary flags indicating if *w* is the head of each of *bp* and *sp*.

We chose these features in order to capture as much lexical and syntactic information as possible from the transcripts. The motivation for lexical identity and part-of-speech features is that complete utterances are unlikely to end in expressions such as *the* or *but there*, and more likely to finish in nouns, for example. Since fragments indicate almost by definition that the utterance is incomplete, we also included a flag indicating if the final word is a fragment. As for the syntactic features, our intuition is that the boundaries of textually complete utterances tend to occur between large syntactic phrases; a similar approach is used by Koehn et al. (2000) for predicting intonational phrase boundaries in raw text. Our syntactic features were computed using two different parsers: Collins, a high-performance statistical parser (Collins, 2003); and CASS, a partial parser especially designed for use with noisy text (Abney, 1996).

We experimented with several learners, including the propositional rule learner RIPPER (Cohen, 1995), the decision tree learner C4.5 (Quinlan, 1993), Bayesian networks (Heckerman et al., 1995; Jensen, 1996) and support vector machines (SVM) (Vapnik, 1995; Cortes and Vapnik, 1995). We used the implementation of these algorithms provided in the WEKA machine learning toolkit (Witten and Frank, 2000). Table 4 shows the accuracy of the majority-class baseline and of each classifier, using 10-fold cross validation on the 400 training data points, and the mean pairwise agreement by the three human labelers. The linear-kernel SVM classifier achieved the highest accuracy, significantly outperforming the majority-class baseline (Wilcoxon signed rank sum test, $p < 0.001$), and approaching the mean agreement of human labelers. However, there is still room for further improvement. New approaches could include features capturing information from the previous turn by the other speaker, which was available to the human labelers but not to the ML classifiers. Also, the sequential nature of this classification task might be better exploited by more advanced graphical learning algorithms, such as Hidden Markov Models (Rabiner, 1989) and conditional random fields (Lafferty et al., 2001).

4.5.3. Results

First we examine the 400 tokens that were manually labeled by three human annotators, considering the majority label as the gold standard. Of the 100 tokens followed by a smooth switch, 91 were labeled textually complete, a significantly higher proportion than the 42% followed by **H** that were labeled complete ($\chi^2 = 51.7$, $df = 1$, $p \approx 0$). Next, we used our highest performing classifier, the linear-kernel SVM, to automatically label all IPUs in the corpus. Of the 3246 IPUs preceding a smooth switch, 2649 (81.6%) were labeled textually complete, and about half of all IPUs preceding a hold (4272/8123, or 52.6%) were labeled complete. The difference is also significant ($\chi^2 = 818.7$,

Table 5
Features used to estimate the presence of individual turn-yielding cues. All features were speaker normalized using z-scores.

Individual cues	Features
Intonation	Absolute value of the F_0 slope over the IPU-final 200 ms Absolute value of the F_0 slope over the IPU-final 300 ms
Speaking rate	Syllables per second over the whole IPU Phonemes per second over the whole IPU
Intensity level	Mean intensity over the IPU-final 500 ms Mean intensity over the IPU-final 1000 ms
Pitch level	Mean pitch over the IPU-final 500 ms Mean pitch over the IPU-final 1000 ms
IPU duration	IPU duration in ms Number of words in the IPU
Voice quality	Jitter over the IPU-final 500 ms Shimmer over the IPU-final 500 ms Noise to harmonics ratio over the IPU-final 500 ms

$df = 1, p \approx 0$). These results suggest that textual completion as defined above constitutes a necessary, but not sufficient, turn-yielding cue.

4.6. Combining turn-yielding cues

So far, we have shown strong evidence supporting the existence of individual acoustic, prosodic and textual turn-yielding cues. Now we shift our attention to the manner in which they combine together to form more complex turn-yielding signals. For each individual cue type, we choose two or three features shown to correlate strongly with smooth switches, as shown in Table 5 (e.g., the speaking rate cue is represented by two automatic features: syllables and phonemes per second over the whole IPU).

We consider a cue c to be PRESENT on IPU u if, for any feature f modeling c, the value of f on u is closer to f_S than to f_H, where f_S and f_H are the mean values of f across all IPUs preceding **S** and **H**, respectively. Otherwise, we say c is ABSENT on u. For the IPUs in the corpus automatically annotated for textual completion, IPUs classified as complete are considered to contain the textual completion turn-yielding cue.

Table 6
Top 10 frequencies of complex turn-yielding cues for IPUs preceding **S**, **H**, **PI** and **BC**. For each of the seven cues, a digit indicates presence, and a dot, absence. 1: intonation; 2: speaking rate; 3: intensity level; 4: pitch level; 5: IPU duration; 6: voice quality; 7: textual completion.

S		**H**		**PI**		**BC**	
Cues	Count	Cues	Count	Cues	Count	Cues	Count
1234567	267	...4...	392	.23456.	17	.2..5.7	53
.234567	226	7	247	...4...	13	.2....7	29
1234.67	138	....	223	...45..	12	12..5.7	23
.234.67	109	...4..7	218	....	9	.2.45.7	23
.23..67	98	...45..	178	123..6.	7	12..567	21
..34567	94	.2....7	166	.234.6.	7	.2..5..	21
123..67	93	1234.67	163	.2.4.6.	7	12.4567	18
.2.4567	73	.2..5.7	157	..3456.	7	.2.4567	17
.2.45.7	73	123..67	133	..34.6.	7	1234567	16
12.4.67	70	1234567	130	...4..7	7	12....7	16
	...		...		...		...
Total	3246	Total	8123	Total	274	Total	553

Table 7
Distribution of number of turn-yielding cues displayed in IPUs preceding **S**, **H**, **PI** and **BC**.

# Cues	**S**		**H**		**PI**		**BC**	
0	4	(0.1%)	223	(2.7%)	9	(3.3%)	1	(0.2%)
1	52	(1.6%)	970	(11.9%)	33	(12.0%)	15	(2.7%)
2	241	(7.4%)	1552	(19.1%)	59	(21.5%)	82	(14.8%)
3	518	(16.0%)	1829	(22.5%)	59	(21.5%)	140	(25.3%)
4	740	(22.8%)	1666	(20.5%)	53	(19.3%)	137	(24.8%)
5	830	(25.6%)	1142	(14.1%)	46	(16.8%)	113	(20.4%)
6	594	(18.3%)	611	(7.5%)	12	(4.4%)	49	(8.9%)
7	267	(8.2%)	130	(1.6%)	3	(1.1%)	16	(2.9%)
Total	3246	(100%)	8123	(100%)	274	(100.0%)	553	(100.0%)

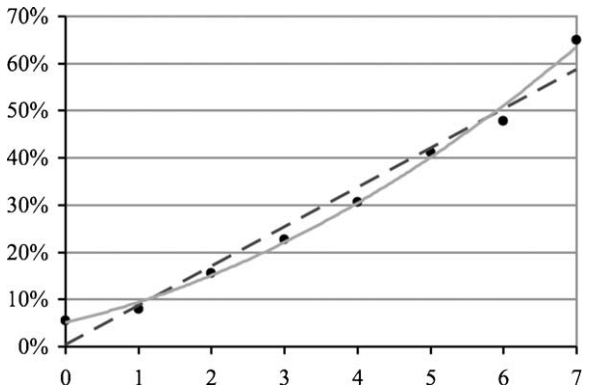

Fig. 11. Percentage of turn-taking attempts (either **S** or **PI**) following IPUs with 0–7 turn-yielding cues.

We first analyze the frequency of occurrence of conjoined individual turn-yielding cues. Table 6 shows the top 10 frequencies for IPUs immediately before smooth switches (**S**), holds (**H**), pause-interruptions (**PI**) and backchannels (**BC**). For IPUs preceding a smooth switch (**S**), the most frequent cases correspond to all, or almost all, cues present at once. For IPUs preceding a hold (**H**), the opposite is true: those with no cues, or with just one or two, represent the most frequent cases. Two different conditions seem to occur before pause interruptions (**PI**): some of the IPUs exhibit four or even five conjoined cues; others show evidence of almost none, as before **H**. This is consistent with two plausible explanations for a **PI** to occur in the first place: (1) that the speaker displays – possibly involuntarily – one or more turn-yielding cues, thus leading the listener to believe that a turn boundary has been reached; or (2) that the listener chooses to break in, regardless of any turn-yielding cues. Finally, the distribution of cues before a **BC** does not show a clear pattern, suggesting that backchannel-inviting cues do indeed differ from turn-yielding cues. Backchannel-inviting cues are discussed in detail in Section 5.

Table 7 shows similar results, now grouping together all IPUs with the same *number* of cues, independently of the cue types. Again, we observe that larger proportions of IPUs preceding **S** present more conjoined cues than IPUs preceding **H**, **PI** and **BC**.

Next we look at how the likelihood of a turn-taking attempt varies with respect to the number of individual cues displayed by the speaker, a relation hypothesized to be linear by Duncan (1972). Fig. 11 shows the proportion of IPUs with 0–7 cues present that are followed by a turn-taking attempt from the interlocutor. The proportion of turn-taking attempts is computed for each cue count as the number of **S** and **PI** divided by the number of **S**, **PI**, **H** and **BC**, according to our labeling scheme.[8] The dashed line in Fig. 11 corresponds to a linear model fitted to the data (Pearson's correlation test: $r^2 = 0.969$), and the continuous line, to a quadratic model ($r^2 = 0.995$). The high correlation coefficient

[8] In this analysis we only consider non-overlapping exchanges, thus leaving out **O**, **I**, **BI** and **BC_O**; overlapping exchanges are addressed in Section 6. Also, note that backchannels are not considered turn-taking attempts.

Table 8
Multiple logistic regression model fit to the data in our corpus: coefficient estimates, standard errors, and t and p values.

X_i	$\hat{\beta}_i$	SE	t	p
Textual completion	0.988	0.047	20.819	≈0
Voice quality	0.698	0.049	14.350	≈0
Speaking rate	0.531	0.047	11.251	≈0
Intensity level	0.470	0.048	10.496	≈0
Pitch level	0.378	0.045	8.357	≈0
IPU duration	0.249	0.044	5.668	≈0
Intonation	-0.044	0.046	-0.967	0.333

of the linear model supports Duncan's hypothesis of a linear relation. An ANOVA test reveals that the quadratic model fits the data significantly better than the linear model ($F(1, 5) = 23.014$; $p = 0.005$), even though the curvature of the quadratic model is only moderate, as can be observed in the figure. We may conclude that, in our corpus, the observed likelihood of a turn-taking attempt by the interlocutor increases in a nearly linear fashion with respect to the number of cues displayed by the speaker.

Lastly, we fit a multiple logistic regression model to the data in our corpus to assess the relative importance of each of the seven turn-yielding cues. The model can be expressed as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7$, where $Y \in \{0, 1\}$ represents whether a turn-taking attempt took place after an IPU, $X_i \in \{0, 1\}$ captures the presence of the ith cue on the IPU, and β_i is the corresponding regression coefficient. Table 8 lists the seven turn-yielding cues sorted by their regression coefficients. According to this model, the textual completion cue ranks first in importance, followed by voice quality, speaking rate, intensity level, pitch level and IPU duration. Notably, for intonation we find no evidence that its coefficient differs significantly from 0 ($p = 0.333$); i.e., the intonation cue plays no role in the model. When we fit a simple logistic model $Y = \beta_0 + \beta_1 X_1$ to the data, with X_1 corresponding to the intonation cue, we obtain $\hat{\beta}_1 = 0.349$ ($SE = 0.041$, $t = 8.48$, $p \approx 0$). This indicates that the intonation cue has a positive and significant correlation with the likelihood of a turn-taking attempt; however, in a multiple logistic regression model this cue seems to be redundant with the information contained in the other six cues.

4.7. Speaker variation

To investigate possible speaker dependence in our turn-yielding cues, we examine evidence for each cue for each of our 13 speakers. Table 9 summarizes these data. For each speaker, a check (√) indicates that there is significant evidence of the speaker producing the corresponding individual turn-yielding cue (at $p < 0.05$, using the same statistical tests described in the previous sections). Five speakers show evidence of all seven cues, while the remaining eight speakers show either five or six cues. Pitch level is the least reliable cue, present only for seven subjects. Notably, the cues related to speaking rate, textual completion, voice quality, and IPU duration are present for all thirteen speakers.

Table 9
Summary of results for each individual speaker.

Speaker	101	102	103	104	105	106	107	108	109	110	111	112	113
Intonation	√	√	√	√	√	√	√	√	√		√	√	
Spk. rate	√	√	√	√	√	√	√	√	√	√	√	√	√
Intensity		√	√	√	√	√	√	√	√	√	√	√	√
Pitch	√	√			√		√		√	√		√	
Completion	√	√	√	√	√	√	√	√	√	√	√	√	√
Voice quality	√	√	√	√	√	√	√	√	√	√	√	√	√
IPU duration	√	√	√	√	√	√	√	√	√	√	√	√	√
LM r^2	.92	.93	.82	.88	.97	.96	.95	.95	.97	.91	.95	.97	.89
QM r^2	.98	.95	.95	.92	.98	.98	.96	.95	.99	.94	.98	.99	.90
LM vs. QM	⋆		⋆			·			⋆		·	⋆	

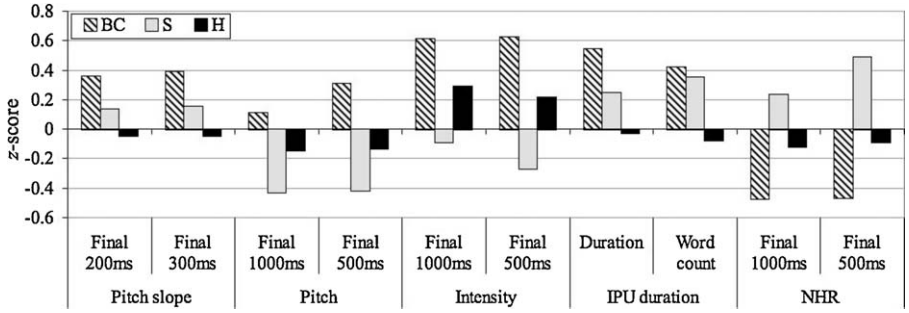

Fig. 12. Individual backchannel-inviting cues: intonation, pitch, intensity, IPU duration and voice quality. In all cases, the difference between the **BC** and **H** groups is significant (ANOVA, $p < 0.01$).

The bottom part of Table 9 shows the correlation coefficients (r^2) of linear and quadratic regressions performed on the data from each speaker. In all cases, the coefficients are very high, indicating that the models explain most of the variation in the data. The fit is significantly better for the quadratic model than for the linear model for four speakers (marked with a star on the final row: ANOVA, $p < 0.05$), and this difference approaches significance for two other speakers (marked with a dot: ANOVA, $p < 0.1$). For the remaining seven speakers, both models provide statistically indistinguishable explanations of the data. This further supports the hypothesis that the likelihood of a turn-taking attempt increases almost linearly with the number of displayed cues.

5. Backchannel-inviting cues

We continue our study of turn-taking phenomena by focusing on a second set of cues produced by the speaker that may induce a particular behavior from the listener, which we term BACKCHANNEL-INVITING CUES. Backchannels are short expressions, such as *uh-huh* or *mm-hm*, uttered by the listener to convey that they are paying attention, and to encourage the speaker to continue. Normally, they are neither disruptive nor acknowledged by the speaker holding the conversational floor. Hypothetically, speakers produce a set of cues marking specific moments within speaking turns at which listeners are welcome to produce backchannel responses.

Finding out whether such cues exist and being able to model them could help answer two of the empirical questions discussed in Section 1: Q2. The system wants to keep the floor but to ensure that the user is paying attention: how should it produce output encouraging the user to utter a backchannel? Q5. The user is speaking: how does the system decide whether and when to produce a backchannel as positive feedback to the user?

In this section we investigate the existence of lexical, acoustic and prosodic backchannel-inviting cues. Using the turn-taking categories available in our corpus, we compare IPUs preceding a backchannel (**BC**) to IPUs preceding a hold (**H**), making the assumption that such cues, if they exist, are more likely to occur in the former group. Additionally, we contrast IPUs before a **BC** with those before a smooth switch (**S**), to study how backchannel-inviting cues differ from turn-yielding cues.

5.1. Individual cues

We repeat the procedures described in Section 4, now looking for individual backchannel-inviting cues instead of turn-yielding cues. We find significant differences between IPUs preceding **BC** and **H** for final intonation, pitch and intensity levels, IPU duration, and voice quality. These results are summarized in Fig. 12.

IPUs immediately preceding backchannels show a clear tendency towards a final rising intonation. All pitch slope measures (raw and stylized, over the IPU-final 200 and 300 ms) are significantly higher before **BC** than before **S** or **H**. Categorical ToBI labels support this finding, as seen in Table 10. More than half of the IPUs preceding a backchannel end in a high-rise contour (H–H%), and about a quarter in a low-rise contour (L–H%). Together, these two contours account for more than 81% of all IPUs before **BC**, but only 36.2% and 20.6% of those before **S** and **H**, respectively. Thus, final intonation presents significantly different patterns in IPUs preceding these three turn-taking categories:

Table 10
ToBI phrase accent and boundary tone for IPUs preceding **BC**, **S** and **H**.

	BC		**S**		**H**	
H–H%	257	55.7%	484	(22.1%)	513	(9.1%)
[!]H–L%	27	5.9%	289	(13.2%)	1680	(29.9%)
L–H%	119	25.8%	309	(14.1%)	646	(11.5%)
L–L%	52	11.3%	1032	(47.2%)	1387	(24.7%)
No boundary tone	4	0.9%	16	(0.7%)	1261	(22.4%)
Other	2	0.4%	56	(2.6%)	136	(2.4%)
Total	461	100.0%	2186	(100.0%)	5623	(100.0%)

either high-rising or low-rising before backchannels, either falling or high-rising before smooth switches, and plateau before holds (chi-square test, $\chi^2 = 1903$, $d.f. = 10$, $p \approx 0$).

Mean pitch and mean intensity levels are significantly higher for IPUs before **BC** than before the other two categories. This suggests that backchannel-inviting cues related to these two features function in a manner opposite to turn-yielding cues. We also find that IPUs followed by backchannels are significantly longer than IPUs followed by either smooth switches or holds, both when measured in seconds and in number of words. Thus, IPU duration works not only as a potential turn-yielding cue, as we saw in previous sections, but also as a backchannel-inviting cue.

For voice quality, we find differences for just one of the three features under consideration. Noise-to-harmonics ratio (NHR) is significantly lower in IPUs preceding **BC** than in those preceding **H**. Again, this backchannel-inviting cue is the opposite of the related turn-yielding cue, which corresponds to a high level of NHR. For the other two voice quality features, jitter and shimmer, the two groups are statistically indistinguishable.

Next we look at lexical backchannel-inviting cues. We examine the distribution of part-of-speech tags in IPU-final phrases, and find that as many as 72.5% of all IPUs preceding backchannels end in either 'DT NN', 'JJ NN', or 'NN NN' (see Table 11)—that is, 'determiner noun' (e.g., *the lion*), 'adjective noun' (*blue mermaid*), or 'noun noun' (*top point*). In comparison, the same three final POS bigrams account for only 31.1% and 21.3% of IPUs preceding **S** and **H**, respectively. Furthermore, the three most frequent final POS bigrams before **S** and **H** add up to just 43.7% and 29.0%, showing more spread distributions, and suggesting that the part-of-speech variability for IPUs before a **BC** is relatively very low. These results strongly suggest the existence of a backchannel-inviting cue related to the part-of-speech tags of the IPU-final words.

Table 11
Count and cumulative percentage of the 10 most frequent IPU-final POS bigrams preceding **BC**, **S** and **H**.

BC			**S**			**H**		
POS	#	%	POS	#	%	POS	#	%
DT NN	**234**	**42.3%**	DT NN	600	18.5%	DT NN	1093	13.5%
JJ NN	**100**	**60.4%**	UH	578	36.3%	UH	832	23.7%
NN NN	**67**	**72.5%**	JJ NN	242	43.7%	JJ NN	430	29.0%
IN NN	12	74.7%	NN NN	168	48.9%	IN DT	374	33.6%
DT JJ	12	76.9%	DT JJ	111	52.3%	UH UH	243	36.6%
IN PRP	9	78.5%	NN UH	96	55.3%	DT JJ	225	39.4%
NN RB	7	79.7%	IN PRP	90	58.1%	IN NN	214	42.0%
DT NNP	7	81.0%	UH UH	83	60.6%	NN NN	211	44.6%
VBZ VBG	6	82.1%	JJR NN	83	63.2%	DT UH	154	46.5%
NNS NN	5	83.0%	IN DT	67	65.2%	NN IN	112	47.9%
	...			...			...	
Total	553	100%	Total	3246	100%	Total	8123	100%

Table 12
Acoustic features used to estimate the presence of individual backchannel-inviting cues. All features were speaker normalized using z-scores.

Individual cues	Features
Intonation	F_0 slope over the IPU-final 200 ms F_0 slope over the IPU-final 300 ms
Intensity level	Mean intensity over the IPU-final 500 ms Mean intensity over the IPU-final 1000 ms
Pitch level	Mean pitch over the IPU-final 500 ms Mean pitch over the IPU-final 1000 ms
IPU duration	IPU duration in ms Number of words in the IPU
Voice quality	Noise to harmonics ratio over the IPU-final 500 ms Noise to harmonics ratio over the IPU-final 1000 ms

5.2. Combining cues

After finding evidence of the existence of individual acoustic, prosodic and textual backchannel-inviting cues, we replicate the procedures described in previous sections to investigate how such cues combine together to form complex signals. For each individual cue, we choose two features shown to strongly correlate with IPUs preceding backchannels, as seen above. These features are shown in Table 12. For example, the individual cue related to IPU-final intonation is represented by two objective measures of the F_0 slope, computed over the final 200 and 300 ms of the IPU.

Next, we estimate the presence or absence in a given IPU of each of the individual cues in the left column of Table 12 using the same procedure described in the Section 4.6. Additionally, we annotate automatically all IPUs in the corpus according to whether they end in one of the three POS bigrams found to strongly correlate with IPUs preceding a backchannel: 'DT NN', 'JJ NN' and 'NN NN'. IPUs ending in any such POS bigram are considered to contain the 'POS bigram' backchannel-inviting cue. Since this feature is essentially binary, no further processing is necessary.

We first analyze the frequency of occurrence of conjoined individual cues before each turn-taking category. Table 13 shows the top ten frequencies for IPUs immediately before a backchannel (**BC**), a smooth switch (**S**), and a hold (**H**). For IPUs preceding **BC**, the most frequent cases correspond to all, or almost all, cues present at once. Very different are IPUs preceding **H**, which show few to no cues. For IPUs preceding **S**, those with no cues, or just one or two, represent the most frequent cases. This suggests that signals produced by speakers to yield the turn differ considerably from signals that invite the interlocutor to utter a backchannel response. Table 14 shows similar results, now grouping together all

Table 13
Top 10 frequencies of complex backchannel-inviting cues for IPUs preceding **BC**, **S** and **H**. For each of the six cues, a digit indicates presence, and a dot, absence. 1: intonation; 2: intensity level; 3: pitch level; 4: IPU duration; 5: voice quality; 6: final POS bigram.

BC		**S**		**H**	
Cues	Count	Cues	Count	Cues	Count
123456	83	.....	243	.2..5.	865
12.456	49	...4..	195	.23.5.	533
123.56	47	..3...	172	.....	513
.23456	27	1.....	153	..3...	414
12345.	24	1..4..	123	....5.	368
123.5.	19	1.3...	113	.2.45.	344
12.45.	16	...4.6	111	.2....	330
12..56	16	1..4.6	108	1.....	256
1.3456	14	...45.	107	...45.	237
.2.456	14	.2....	94	...4..	218
	...		...		...
Total	553	Total	3246	Total	8123

Table 14
Distribution of number of backchannel-inviting cues displayed in IPUs preceding **BC**, **S** and **H**.

# Cues	**BC**		**S**		**H**	
0	4	(0.7%)	243	(7.5%)	513	(6.3%)
1	17	(3.1%)	746	(23.0%)	1634	(20.1%)
2	57	(10.3%)	912	(28.1%)	2364	(29.1%)
3	90	(16.3%)	723	(22.3%)	1960	(24.1%)
4	139	(25.1%)	379	(11.7%)	1010	(12.4%)
5	163	(29.5%)	192	(5.9%)	501	(6.2%)
6	83	(15.0%)	51	(1.6%)	141	(1.7%)
Total	553	(100%)	3246	(100%)	8123	(100%)

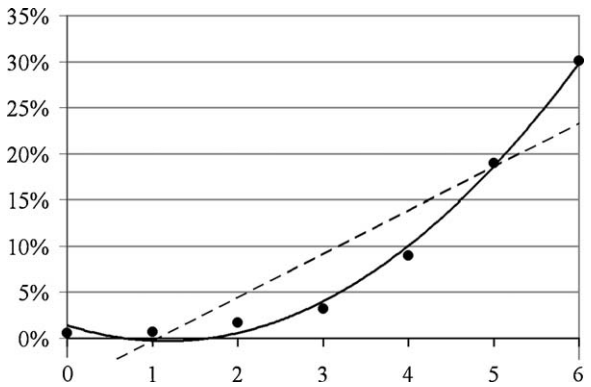

Fig. 13. Percentage of backchannels following IPUs with 0–6 backchannel-inviting cues.

IPUs with the same *number* of cues, independently of the cue types. Again, we observe that larger proportions of IPUs preceding **BC** show more conjoined cues than IPUs preceding **S** and **H**.

Next we look at how the likelihood of occurrence of a backchannel varies with respect to the number of individual cues conjointly displayed by the speaker. Fig. 13 shows the proportion of IPUs with 0–6 cues present that are followed by a backchannel from the interlocutor—namely, the number of **BC** divided by the number of **S**, **PI**, **H** and **BC**, for each cue count.[9] The dashed line in the plot corresponds to a linear model fitted to the data ($r^2 = 0.812$); the continuous line, to a quadratic model ($r^2 = 0.993$). The fit of the quadratic model is significantly better than that of the linear model (ANOVA, $F(1, 4) = 110.0$, $p < 0.001$). In this case, the fit of the linear model is not as good as in the case of turn-yielding cues. The quadratic model, on the other hand, achieves an almost perfect fit and shows a marked curvature, confirming that a quadratic model provides a plausible explanation for the relation between number of backchannel-inviting cues and occurrence of a backchannel.

Lastly, we fit a multiple logistic regression model to the data in our corpus to assess the relative importance of each of the six backchannel-inviting cues. The model can be expressed as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6$, where $Y \in \{0, 1\}$ represents whether a backchannel was uttered after an IPU, $X_i \in \{0, 1\}$ captures the presence of the ith cue on the IPU, and β_i is the corresponding regression coefficient. Table 15 lists the six backchannel-inviting cues sorted by their regression coefficients. According to this model, the POS bigram cue ranks first in importance, followed by intonation, intensity level, IPU duration, voice quality and pitch level. Note that in this case all six cues are included in the model, as opposed to the case of turn-yielding cues, for which intonation was excluded—in fact, in this case the intonation cue ranks second in relevance.

[9] Again, we only consider non-overlapping exchanges, thus leaving out **O**, **I**, **BI** and **BC_O**.

Table 15
Multiple logistic regression model fit to the data in our corpus: coefficient estimates, standard errors, and t and p values.

X_i	$\hat{\beta}_i$	SE	t	p
POS bigram	1.499	0.100	14.972	≈0
Intonation	1.153	0.099	11.667	≈0
Intensity level	0.649	0.109	5.946	≈0
IPU duration	0.626	0.097	6.479	≈0
Voice quality	0.618	0.121	5.104	≈0
Pitch level	0.366	0.095	3.840	0.0001

Table 16
Summary of results for individual speakers.

Speaker	102	103	105	106	108	110	111	112	113
Intonation	√	√	√	√		√		√	√
Pitch level							√	√	√
Intensity level		√		√	√		√	√	√
IPU duration		√	√	√	√	√	√	√	√
Voice quality		√	√	√		√	√	√	√
POS bigram	√	√	√	√	√	√	√	√	√
LM r^2	0.63	0.88	0.72	0.80	0.63	0.70	0.84	0.80	0.85
QM r^2	0.70	0.96	0.95	0.80	0.87	0.95	0.93	0.99	0.99
LM vs. QM		⋆	⋆		·	⋆	·	⋆	⋆

5.3. Speaker variation

We investigate the existence of the hypothesized backchannel-inviting cues for each individual speaker. Four subjects have fewer than 20 instances of IPUs preceding **BC**, a count too low for statistical tests, and are thus excluded from the analysis. Table 16 summarizes the evidence found for the remaining nine speakers. For each speaker, a check (√) means there is significant evidence of the existence of the corresponding cue.

Differences in intonation, duration and voice quality are significant for the great majority of speakers, and a smaller proportion of speakers display differences in pitch and intensity. Also, all nine speakers show a marked preference for at least two of the three final POS bigrams mentioned above before backchannels. Notably, no single acoustic/prosodic cue is used by all speakers; rather, each seems to use their own combination of cues. For example, speaker 102 varies only intonation, while speaker 108 varies only intensity level and IPU duration. The bottom rows in Table 16 show the correlation coefficient (r^2) of the linear and quadratic regressions performed separately on the data from each speaker. The fit of the linear models ranges from moderate at 0.625 to high at 0.884. In seven out of nine cases, the fit of the quadratic models is significantly better, ranging from 0.702 to 0.990 (ANOVA, star: $p < 0.05$, dot: $p < 0.1$). Thus, even though speaker variation in the production of backchannel-inviting cues is not insignificant, a quadratic model seems to successfully explain the relation between the number of backchannel-inviting cues conjointly displayed, and the likelihood of occurrence of a backchannel.

6. Overlapping speech

Often in conversation speakers take the turn just before the end of their interlocutors' contribution, without interrupting the conversational flow (Sacks et al., 1974). There is evidence of the occurrence of these events in multiple languages, including Arabic, English, German, Japanese, Mandarin and Spanish (Yuan et al., 2007), and previous studies also report situational and genre differences. For example, non-face-to-face dialogues have significantly fewer speech overlaps than face-to-face ones (ten Bosch et al., 2005); people make fewer overlaps when talking with strangers (Yuan et al., 2007); and speakers tend to make fewer overlaps and longer pauses when performing difficult tasks (Bull and Aylett, 1998). The existence of this phenomenon suggests that listeners are capable of anticipating possible turn endings, and poses the question of how they manage to do this. One possible explanation could be the occurrence of

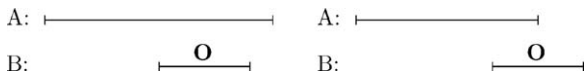

Fig. 14. Full and partial overlap types.

early turn-taking cues in the speaker's turn. Such cues might be perceived by listeners, allowing them to anticipate upcoming places for either taking the turn or producing a backchannel response.

Simultaneous speech poses serious difficulties for the ASR components of IVR systems (Shriberg et al., 2001). It is possible that knowledge about the nature of previous turn-taking cues could enable IVR systems to reduce the likelihood of occurrence of simultaneous speech by avoiding the production of utterances containing speech events that could be interpreted by the user as early turn-yielding or backchannel-inviting cues. Hypothetically, this strategy could prevent the user from taking the turn or backchanneling before the system has completely finished producing its current utterance, although clearly it could be difficult to separate early from late cues in a short utterance.

In this section we look for evidence of the occurrence of early turn-taking cues in conversation. For this, we first review the different patterns of overlapping speech defined by our labeling scheme, and examine their frequencies in the Games Corpus. Second, we compare IPUs preceding transitions with and without overlapping speech, exploring whether both groups present similar occurrence patterns of turn-yielding and backchannel-inviting cues. Third, we study the durational distribution of overlapping speech segments, aiming at identifying plausible locations to search for early turn-taking cues. Finally, we look for evidence of early turn-taking cues in turn-medial speech segments.

6.1. Types of overlapping speech in the Games Corpus

The turn-taking labeling scheme presented in Section 3.4 includes four categories of turn exchanges with simultaneous speech present: overlap (**O**), backchannel with overlap (**BC_O**), interruption (**I**) and butting-in (**BI**). In this study we consider only the first two classes (**O** and **BC_O**), and ignore the last two, since they correspond to disruptions of the conversational flow at arbitrary points during the speaker's turn, rather than unobtrusive overlap during fluent exchanges. Note that the existence of overlapping speech is the only difference between **O** and smooth switches (**S**), and between **BC_O** and backchannels (**BC**).

Instances of **O** can be divided into two cases: FULL OVERLAPS, which take place completely within the interlocutor's turn (as depicted in the left part of Fig. 14); and PARTIAL OVERLAPS, which begin during the interlocutor's turn but extend further after its end (right part of Fig. 14). Fully and partially overlapping backchannels are defined analogously. In this study we consider only instances of partial **O** and **BC_O**, which are clear cases of utterance endings overlapped by new utterances from the interlocutor. For fully overlapping instances, we have as yet no firm hypothesis of where in the prior speech we should look for cues that might trigger the fully overlapping contribution. Furthermore, full overlaps correspond to complex events in which the current speaker talks – without pausing – before, during and after a complete utterance from the interlocutor. In such occasions, one might say that the two speakers briefly *share* the conversational floor, an interesting phenomenon we will address specifically in future research.

In the Games Corpus, 767 of the 1067 instances of **O**, as well as 104 of the 202 tokens of **BC_O**, are partially overlapping. We use only these data in the present study. For clarity, we hereafter refer to partially overlapping **O** and **BC_O** simply as **O** and **BC_O**. To illustrate the procedures we describe below for investigating potential cues to turn changes and backchannel productions in speech preceding overlaps, we provide a schematic in Fig. 15.

6.2. Existence of cues in IPUs overlapped by *O* or *BC_O*

In Section 4 we presented a procedure to estimate the existence of seven turn-yielding cues before smooth switches (**S**). We begin our study of overlapping speech by searching for evidence of the same cues in the IPUs that are themselves

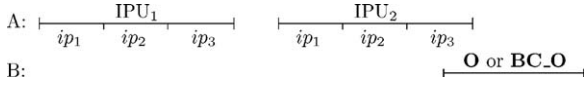

Fig. 15. Two inter-pausal units (IPU_1, IPU_2), each with three intermediate phrases (ip_1, ip_2, ip_3); ip_3 of IPU_2 is overlapped by speech from the interlocutor.

Table 17
Top: top 10 frequencies of complex turn-yielding cues for IPUs overlapped by **O** (*cf.* Table 6). Bottom: distribution of number of turn-yielding cues in IPUs overlapped by **O** (*cf.* Table 7).

Cues	Count
1234567	61
.234567	50
.234.67	26
.23456.	24
..34567	24
1234.67	22
..3..67	22
123456.	21
.2.4567	20
..34.67	20
	...

# Cues	O	
0	1	(0.1%)
1	15	(2.0%)
2	55	(7.2%)
3	111	(14.5%)
4	163	(21.3%)
5	213	(27.8%)
6	148	(19.3%)
7	61	(8.0%)
Total	767	(100%)

overlapped (corresponding to IPU_2 in Fig. 15) (**O**)—to see whether upcoming cues are present very close to the point of overlap. Results are summarized in Table 17. The table at the top lists the top ten frequencies of complex cues (1: intonation; 2: speaking rate; 3: intensity level; 4: pitch level; 5: IPU duration; 6: voice quality; 7: textual completion). Similarly to what we observe for IPUs followed by **S** (see Table 6), the most frequent cases correspond to all, or almost all, cues present at once. The bottom part of Table 17 shows the same results, now grouping together all IPUs with the same number of cues, independent of cue types (see Table 7). Again, we observe a marked tendency of IPUs preceding **O** to present a high number of conjoined turn-yielding cues. These results indicate that entire IPUs immediately preceding smooth switches (**S**) and overlaps (**O**) show a similar behavior in terms of the occurrence of our posited turn-yielding cues.

We repeat the same analysis to study the presence of backchannel-inviting cues in IPUs overlapped by backchannels (**BC_O**; again, corresponding to IPU_2 in Fig. 15). The results are summarized in Table 18, and are comparable to the results obtained for backchannels without overlap (**BC**), shown in Tables 13 and 14. In both cases, we observe that IPUs preceding **BC** or **BC_O** tend to have a high number of conjointly displayed cues. These results indicate that IPUs immediately preceding backchannels (**BC**) and IPUs overlapped by backchannels (**BC_O**) also show considerably similar patterns of occurrence of backchannel-inviting cues. We now turn to the question of where in the overlapped turn such cues may be found.

6.3. Early turn-yielding cues

In Section 6.2 we compared the occurrence of turn-yielding cues in (entire) IPUs preceding **S** and (entire) IPUs overlapped by **O**, and also compared the occurrence of backchannel-inviting cues in (entire) IPUs preceding **BC** and (entire) IPUs overlapped by **BC_O**. In this section we investigate the location of cues within smaller portions of overlapped IPUs. First, we examine the durational distribution of the overlapped portion of overlapped/overlapping IPUs in our corpus, looking for reasonable places to search for such cues. We then identify possible early turn-yielding cues by comparing features in different portions of the preceding IPUs. Given the low number of backchannels with overlap (**BC_O**) in the corpus, we restrict this study to turn-shifting overlaps (**O**).

Table 18
Top: top 10 frequencies of complex backchannel-inviting cues for IPUs overlapped by backchannels **BC_O** (*cf.* Table 13). Bottom: distribution of number of backchannel-inviting cues in IPUs preceding **BC_O** (*cf.* Table 14).

Cues	Count
123456	14
12.456	9
.23456	8
12345.	6
123.56	6
1..456	5
123.5.	4
12.45.	4
.2.456	3
.2.45.	3
	...

# Cues	BC_O	
0	1	(1.0%)
1	3	(2.9%)
2	8	(7.7%)
3	20	(19.2%)
4	28	(26.9%)
5	30	(28.8%)
6	14	(13.5%)
Total	104	(100%)

6.3.1. *Onset of overlaps*

The annotation of turn-taking phenomena in the Games Corpus specifies only the presence or absence of overlapping speech (e.g., **O** vs. **S**). However, it does not provide information about the duration of the overlapping segments, knowledge important for locating cues that could be perceived by listeners early enough to anticipate turn endings. We first examine the distribution of overlapped segments that occur in the corpus.

Fig. 16 shows the cumulative frequency distribution of the duration of overlapping speech segments in overlaps (**O**). Approximately 60% of **O**s have 200 ms or less of simultaneous speech, and 10% have 500 ms or more, although only a small number have more than one second. If we look at lexical rather than temporal units, we find that 613 (80%) of all instances begin during the last word in the previous turn; 100 (13%), during the penultimate word; and the remaining 54 (7%), still earlier. The mean duration of the final word before overlaps is 384 ms (stdev = 180 ms);

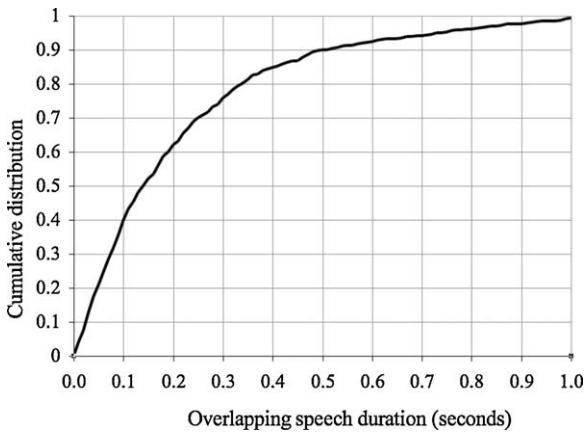

Fig. 16. Cumulative frequency distribution of the duration of overlapping speech segments in turn exchanges labeled **O**.

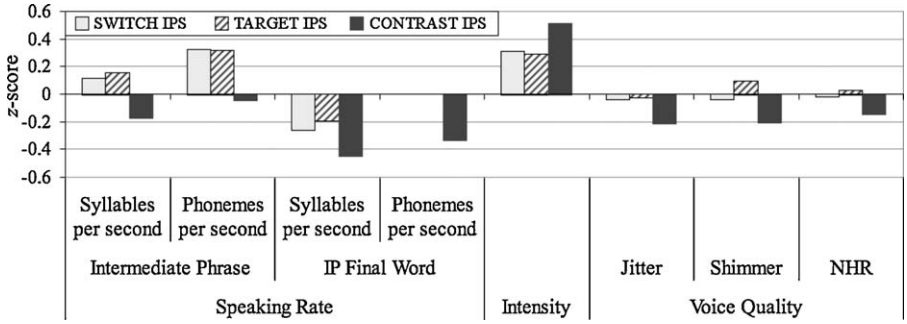

Fig. 17. Individual turn-yielding cues in SWITCH IPS, TARGET IPS, and in CONTRAST IPS: speaker-normalized speaking rate, intensity and voice quality features. In all cases, the difference between CONTRAST IPS and the other two groups is significant (ANOVA, $p < 0.05$).

and of the penultimate word, 376 ms (stdev = 170 ms). Finally, in terms of prosodic units, we find that over 95% of overlaps begin during the turn-final intermediate phrase (*ip*) of the IPU; we use our hand-annotated ToBI labels in this calculation.[10] The mean duration of the final *ip* before overlaps is 747 ms (stdev = 418 ms).

These results indicate that, while in most cases the overlapping turn begins just before the end of the previous turn, in some cases the overlapping speech spans up to several words. Nonetheless, since nearly all overlaps occur during the turn-final *ip*, the penultimate *ip*– occurring just before the overlapped *ip*– appears to be a plausible place to search for early turn-yielding cues. These *ip*s represent linguistically motivated units, which are better suited to the calculation of our F_0 features in particular.

6.3.2. Cues in penultimate intermediate phrases

We search for early turn-yielding cues in penultimate *ip*s preceding overlaps (**O**), using a slightly modified version of the procedure described in the previous sections. This intermediate phrase corresponds to ip_2 of IPU_2 in Fig. 15, and will in every case precede the overlapped speech in our corpus; we term these *ip*s TARGET IPS. To identify possible turn-yielding cues as such, we contrast these TARGET IPS with all *ip*s occurring *before* penultimate position for IPUs of any turn type; i.e., in Fig. 15, this would mean that ip_2 of IPU_2 would be contrasted with ip_1 from IPU_2 and ip_1, ip_2, ip_3 from IPU_1. For ease of reference, we call this set of *ip*s CONTRAST IPS. (Note that, in cases when the overlapping IPU contains only a single *ip*, we locate the TARGET IP as the last *ip* in the *previous* IPU within the same turn. If such instances have no preceding IPU within the same turn, they are excluded from consideration.) The question is, can we determine if the *ip* just before the onset of overlapped speech (the TARGET IP) does itself contain characteristics that are distinct from *ip*s occurring either earlier in the overlapped IPU itself or in even earlier speech in the turn (the CONTRAST IPS)? We propose that any significant difference found between the two groups in acoustic, prosodic, lexical and syntactic features might indicate the existence of a potential turn-yielding cue before **O**.

Additionally, we examine *ip*s in the same penultimate position before smooth switches (**S**), which we term SWITCH IPS to determine whether any such cues tend to occur in all turn endings, or whether they constitute a distinct phenomenon that typically occurs only before overlaps. In cases where the final IPU preceding **S** contains only a single *ip*, we follow the same procedure as described above and locate the SWITCH IP in the final *ip* of the preceding IPU in the same speaker turn; when there is only a single IPU with a single *ip*, we exclude this item from our calculations. Note that 58% of IPUs preceding **S** and 48% of IPUs preceding **O** contain exactly one *ip*.

We find significant differences in speaking rate, measured in number of syllables and phonemes per second, over the whole TARGET IP and over its final word, as shown in Fig. 17. The speaking rate of TARGET IPS is significantly faster than that of CONTRAST IPS. We also find that TARGET IPS are produced with significantly lower intensity, and with higher values of three voice quality features—jitter, shimmer and NHR. Additionally, TARGET IPS and SWITCH IPS show no significant differences with respect to all these features. In fact, these TARGET IPS do indeed appear to exhibit similar cues to those we find over the entire IPU preceding **S**, as described in Section 4. Recall that these IPUs preceding **S** tend to be produced with faster speaking rate, lower intensity, and higher jitter, shimmer and NHR than IPUs preceding **H**.

[10] This computation, as well as the subsequent analysis of early turn-yielding cues, considers only the portion of the Games Corpus that is annotated using the ToBI framework, which includes 538 instances of partially-overlapping **O**.

In sum, it does appear that there are cues in the *ip*s *preceding* **O** that are similar to those characterizing entire IPUs preceding **S** —and that SWITCH IPS of IPUs preceding **S** closely reflect the characteristics of entire IPUs preceding **S**. That is, there appear to be local turn-yielding signals in penultimate *ip*s of IPUs preceding **S** and in penultimate *ip*s of IPUs preceding **O**.

7. Discussion

In this study we have examined a number of turn-taking phenomena in human–human conversation, with the ultimate goal of improving the performance and naturalness of practical applications such as IVR systems. To discover possible cues to upcoming turn shifts and backchannels, we have investigated a larger and more varied set of acoustic, prosodic, and lexico-syntactic cues than previous descriptive studies, providing objective definitions of our features and detailed information about the predictive power of each cue type. Our corpus is larger than that of most previous studies, permitting more statistically robust results. Furthermore, in our corpus we have distinguished three importantly distinct phenomena: turn switches, backchannels and interruptions to provide a clearer picture of the different types of speech that precede them. We have examined these in both overlapping and non-overlapping speech.

We have presented evidence of the existence of seven measurable events that take place with a significantly higher frequency on IPUs preceding smooth switches (when the current speaker completes an utterance and the interlocutor takes the turn after a short pause) than on IPUs preceding holds (when the current speaker continues speaking after a short pause). These events may be summarized as follows: a falling or high-rising intonation at the end of the IPU; a reduced lengthening of IPU-final words; a lower intensity level; a lower pitch level; a point of textual completion; a higher value of three voice quality features: jitter, shimmer, and NHR; and a longer IPU duration. These seven events represent potential turn-yielding cues, such that when several cues occur simultaneously, the likelihood of a subsequent turn-taking attempt by the interlocutor increases in a linear manner. Our findings represent the first support for Duncan's (1972) general hypothesis from a large corpus of spontaneous dialogue—although the features we find correlated with turn shifts are somewhat different from those proposed by Duncan and other researchers. In the Games Corpus, the percentage of IPUs followed by a turn-taking attempt ranges from 5% when no turn-yielding cues are present, to 65% when all seven cues are present.

We believe that these findings could be used to improve the turn-taking decisions of state-of-the-art IVR systems. In particular, our model of turn-taking provides answers to three of the questions posed in Section 1:

Q1. The system wants to keep the floor: how should it formulate its output to avoid an interruption from the user?

According to our model, including as few as possible of the described turn-yielding cues in the system's output will decrease the likelihood that the user will take the turn. Therefore, when the system intends to continue holding the floor, it should end its IPUs in plateau intonation, with high intensity and pitch levels, leaving utterances textually incomplete (e.g., preceding pauses with expressions such as *and* or *also*), and so on.

Q3. The system is ready to yield the floor: how should it convey this to the user?

This situation represents the opposite of Q1. If the system includes in its output as many of the described turn-yielding cues as it can produce given its generation component, a turn-taking attempt by the user will be more likely to take place. Thus, if the system intends to cede the floor to the user, it should end its final IPU in either falling or high-rising intonation (depending on whether the system's message is a statement or a direct question), with low intensity and pitch levels, and so on.

Q4. The user is speaking but pauses: how can the system decide whether the user is giving up the turn?

Most current systems simply wait for a pause from the user above a system-defined threshold before attempting to take the turn. It should be possible to improve upon this simple and often unnatural technique using the findings in our study. Although the difficulty of estimating each turn-yielding cue will vary according to the implementation, we may draft a high-level description of a possible turn-taking decision procedure: at every pause longer than 50 ms, the system estimates the presence of as many of the cues we have described as possible over the user's preceding IPU. If the number of detected cues is high, relative to a predefined threshold, it may choose to conduct a turn-taking attempt immediately; otherwise, it may continue to wait, either for a longer pause (defaulting to traditional procedures) or for an IPU with a higher number of perceived cues.

We have also presented evidence of the existence of six backchannel-inviting cues—six measurable events present with a significantly higher frequency in IPUs preceding backchannels than in IPUs preceding holds or smooth switches. These events may be summarized as follows: a rising intonation at the end of the IPU, a higher intensity level, a higher pitch level, a final POS bigram equal to 'DT NN', 'JJ NN' or 'NN NN'; a lower value of noise-to-harmonics ratio (NHR); and a longer IPU duration. We have also shown that, when several backchannel-inviting cues occur simultaneously, the likelihood of occurrence of a backchannel from the interlocutor increases in a quadratic fashion, ranging from only 0% of IPUs followed by a backchannel when no cues are present, to more than 30% when all six cues are present. This latter low proportion of backchannels occurring even when all six cues are present may be explained by the *optionality* of backchannels in SAE as in other conversational speech: speakers do not backchannel at every opportunity and some speakers backchannel less than others; and it is even possible that an entire successful conversation can be completed without the production of any backchannels at all. However, our findings can still be of use in IVR systems in several ways.

We believe these results can help answer two of the IVR-related questions we posed initially:

> Q2. The system wants to keep the floor but to ensure that the user is paying attention: how should it produce output encouraging the user to utter a backchannel?

According to our model, if the system includes in its output as many of the described cues as possible, the likelihood of a backchannel from the user will increase. Thus, if the system intends to elicit a backchannel response from the user, it should end the final IPU in one of the listed part-of-speech bigrams, with rising intonation (preferably high-rising), high pitch and intensity levels, and so on.

> Q5. The user is speaking: how does the system decide whether and when to produce a backchannel as positive feedback to the user?

The decision about when the system should produce a backchannel might be coupled with the procedure described above for detecting turn endings based on turn-yielding cues. Every time the system estimates the presence of turn-yielding cues over the user's final IPU, it can also estimate the presence of backchannel-inviting cues. (Note that some features may be reused, as they belong to both cue sets.) If the number of detected backchannel-inviting cues is high enough, then the system may utter a backchannel; otherwise, it may keep silent. Since at least three backchannel-inviting cues differ in valence from corresponding turn-yielding cues (intensity, pitch and NHR) there may be less risk that the system will confuse turn-taking opportunities with backchannel-producing opportunities.

In an examination of overlapping speech in conversation, we have shown that IPUs preceding overlaps and those preceding smooth switches show comparable patterns of turn-yielding cues. Similarly, IPUs preceding backchannels with and without overlap show comparable patterns of backchannel-inviting cues. Additionally, we observe that some of the turn-yielding cues described in Section 4 appear to occur earlier in the turn, in the penultimate intermediate phrase (TARGET IP) of the overlapped IPU. Thus it even appears possible for future IVR systems to imitate common human propensity to overlap speech in conversation by detecting turn-yielding intentions before the user has completed their turn.

We believe that these findings can benefit future IVR systems, identifying cues that will be useful in improving turn-taking behaviors in them. Such improvements should improve the naturalness and usability of such systems by offering users a turn-taking experience that more closely resembles normal interaction in human–human conversation.

7.1. Directions for future research

Our studies have shown very low speaker variability in turn-yielding cues, with almost all speakers producing all seven cues, but a considerably higher speaker variation for backchannel-inviting cues—in fact, each speaker seems to use their own combination of such cues. Still, some characteristics are true across all speakers for utterances preceding backchannels: all tend to display at least two cues, and all share the POS bigram cue. In the future, we are interested in investigating how, when, and why speakers choose to use a particular set of backchannel-inviting cues.

Another topic of our future research is to investigate additional turn-taking cues related to voice quality. Features such as relative average perturbation (RAP), soft phonation index (SPI), and amplitude perturbation quotient (APQ), all of which have been shown to capture different aspects of voice quality, should be studied. Furthermore, we have

chosen to collapse jitter, shimmer and NHR into one simple voice quality cue, but these features could instead be used individually as finer grained cues.

Also, there is room for improvement in our automatic classification of textual completion. Our best performing classifier, based on support vector machines, achieves an accuracy of 80%, while the agreement for humans is 90.8%. New approaches could incorporate features capturing information from the previous turn by the other speaker, which was available to the human labelers but not to the machine learning classifiers. In addition, the sequential nature of this classification task might be better exploited by more advanced graphical learning algorithms, such as Hidden Markov Models and Conditional Random Fields.

Users of IVR systems sometimes engage in an uninterrupted flow of speech which the system might want to interrupt, either because it has already collected the information needed for the task at hand, or simply because it has lost track of what the user is saying and needs to start over. In such occasions, it is crucial for the system to interrupt in an acceptable manner. Modeling the way people interrupt in spontaneous, collaborative conversations should aid IVR systems in this aspect of turn-taking. Another direction for our future research is to examine the three types of interruptions we have annotated in our corpus (simple **I**, pause **PI**, and barge-in **BI** interruptions) to identify places where these are more likely to occur, and to describe the acoustic and prosodic properties of the interrupter's speech.

An additional future research direction will involve further examination of the location of early turn-yielding and backchannel-inviting cues. Since *ip*s are difficult to identify automatically, it will be useful to investigate different possible segmentations of speech preceding overlaps to see how early cues can most reliably be identified. We also want to investigate whether we can see patterns in cues across segmentations of IPUs preceding overlaps, to test the hypothesis that cues become more pronounced across the course of the IPU.

Finally, in future research we plan to use the information obtained in these studies to experiment with the automatic classification of turn types such as smooth switches and backchannels using information from the speech that precedes such phenomena. By predicting where a speaker's turn is likely to end or where a backchannel is likely to be produced and which of the features we have found to occur in such speech are most effective in predicting different types of turn-taking behavior, we believe we will develop even more practical methods for future IVR systems to employ in managing system-user dialogue.

Acknowledgments

This work was funded in part by NSF IIS-0307905 and IIS-0803148. We thank Štefan Beňuš, Héctor Chávez, Frank Enos, Michel Galley, Enrique Henestroza, Hanae Koiso, Jackson Liscombe, Michael Mulley, Andrew Rosenberg, Elisa Sneed German, and Gregory Ward, for valuable discussions and for their help in collecting, labeling and processing the data. We also thank our anonymous reviewers for valuable suggestions.

References

Abney, S., 1996. Partial parsing via finite-state cascades. Journal of Natural Language Engineering 2 (4), 337–344.

Atterer, M., Baumann, T., Schlangen, D., 2008. Towards incremental end-of-utterance detection in dialogue systems. In: Coling, Manchester, UK, pp. 11–14.

Beattie, G.W., 1981. The regulation of speaker turns in face-to-face conversation; some implications for conversation in soundonly communication channels. Semiotica 34, 55–70.

Beattie, G.W., 1982. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. Semiotica 39, 93–114.

Beckman, M.E., Hirschberg, J., 1994. The ToBI Annotation Conventions. Ohio State Univ.

Bhuta, T., Patrick, L., Garnett, J., 2004. Perceptual evaluation of voice quality and its correlation with acoustic measurements. Journal of Voice 18, 299–304.

Boersma, P., Weenink, D., 2001. Praat: Doing phonetics by computer, http://www.praat.org.

Bull, M., Aylett, M., 1998. An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In: ICSLP.

Cathcart, N., Carletta, J., Klein, E., 2003. A shallow model of backchannel continuers in spoken dialogue. In: EACL, pp. 51–58.

Charniak, E., Johnson, M., 2001. Edit detection and parsing for transcribed speech. In: Proceedings of NAACL.

Collins, M., 2003. Head-driven statistical models for natural language parsing. Computational Linguistics 29, 589–637.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Cohen, W.C., 1995. Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning.

Cortes, C., Vapnik, V., 1995. Support vector networks. Machine Learning, 273–297.

Cutler, E.A., Pearson, M., 1986. On the analysis of prosodic turn-taking cues. In: Johns-Lewis, C. (Ed.), Intonation in Discourse. College-Hill, San Diego, CA, pp. 139–156.

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

Duncan, S., 1973. Toward a grammar for dyadic conversation. Semiotica 9, 29–46.

Duncan, S., 1974. On the structure of speaker-auditor interaction during speaking turns. Language in Society 3, 161–180.

Duncan, S., 1975. Interaction units during speaking turns in dyadic, face-to-face conversations. In: Organization of Behavior in Face-to-Face Interaction. Mouton Publishers, Den Hague, pp. 199–213.

Duncan, S., Fiske, D., 1977. Face-To-Face Interaction: Research, Methods, and Theory. Lawrence Erlbaum Associates.

Du Bois, J., Schuetze-Coburn, S., Cumming, S., Paolino, D., pp. 45–89 1993. Outline of discourse transcription. In: Talking Data: Transcription and Coding in Discourse Research.

Edlund, J., Heldner, M., Gustafson, J., 2005. Utterance segmentation and turn-taking in spoken dialogue systems. Sprachtechnologie mobile Kommunikation und linguistische Ressourcen, 576–587.

Eskenazi, L., Childers, D., Hicks, D., 1990. Acoustic correlates of vocal quality. Journal of Speech, Language and Hearing Research 33, 298–306.

Ferguson, N., 1977. Simultaneous speech, interruptions and dominance. British Journal of Social and Clinical Psychology 16, 295–302.

Ferrer, L., Shriberg, E., Stolcke, A., 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In: Proceedings of ICASSP.

Ferrer, L., Shriberg, E., Stolcke, A., 2002. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody. In: Proceedings of the ICSLP, pp. 2061–2064.

Ford, C., Thompson, S., 1996. Interactional units in conversation: syntactic intonational and pragmatic resources for the management of turns. In: Ochs, E., Schegloff, E., Thompson, S. (Eds.), Interaction and Grammar. Cambridge University Press, pp. 134–184.

Fry, D., 1975. Simple reaction-times to speech and non-speech stimuli. Cortex 11, 355–360.

Godfrey, J., Holliman, E., McDaniel, J., 1992. Switchboard: telephone speech corpus for research and development. In: IEEE International Conference on Acoustics, Speech, and Signal Processing.

Goodwin, C., 1981. Conversational Organization: Interaction Between Speakers and Hearers. Academic Press.

Gravano, A., Benus, S., Hirschberg, J., Mitchell, S., Vovsha, I., 2007. Classification of discourse functions of affirmative words in spoken dialogue. In: Proceedings of Interspeech.

Heckerman, D., Geiger, D., Chickering, D., 1995. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning 20, 197–243.

Hemphill, C., Godfrey, J., Doddington, G., 1990. The ATIS spoken language systems pilot corpus. In: Proceedings of the Workshop on Speech and Natural Language, pp. 96–101.

Hjalmarsson, A., 2009. On cue—additive effects of turn-regulating phenomena in dialogue. In: Diaholmia.

Jefferson, G., 1984. Notes on a systematic deployment of the acknowledgement tokens "yeah"; and "mm hm". Research on Language & Social Interaction 17, 197–216.

Jensen, F., 1996. Introduction to Bayesian Networks. Springer-Verlag, New York.

Jurafsky, D., Shriberg, E., Fox, B., Curl, T., 1998. Lexical, prosodic and syntactic cues for dialog acts. In: Proceedings of ACL/COLING, Workshop on Discourse Relations and Discourse Markers, pp. 114–120.

Kendon, A., 1967. Some functions of gaze-direction in social interaction. Acta Psychologica 26, 22–63.

Kendon, A., 1972. Some relationships between body motion and speech. In: Siegman, A.W., Pope, B. (Eds.), Studies in Dyadic Communication. Pergamon Press, Elmsford, NY, pp. 177–210.

Kitch, J., Oates, J., Greenwood, K., 1996. Performance effects on the voices of 10 choral tenors: acoustic and perceptual findings. Journal of Voice 10, 217–227.

Koehn, P., Abney, S., Hirschberg, J., Collins, M., 2000. Improving intonational phrasing with syntactic information. In: Proceedings of ICASSP, vol. 3, pp. 1289–1290.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y., 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. Language and Speech 41, 295–321, Special issue on prosody and conversation.

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: 18th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 282–289.

Marcus, M., Marcinkiewicz, M., Santorini, B., 1993. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19, 313–330.

McNeill, D., 1992. Hand and Mind: What Gestures Reveal About Thought. University of Chicago Press.

Mushin, I., Stirling, L., Fletcher, J., Wales, R., 2003. Discourse structure, grounding, and prosody in task-oriented dialogue. Discourse Processes 35, 1–31.

Novick, D., Sutton, S., 1994. An empirical model of acknowledgment for spoken-language systems. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp. 96–101.

Ogden, R., 2002. Creaky voice and turn-taking in Finnish. In: Colloquium of the British Association of Audiological Physicians.

Pierrehumbert, J., 1980. The phonology and phonetics of English intonation. Ph.D. Thesis. Massachusetts Institute of Technology.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P.R., Morgan, J., Pollack, M.E. (Eds.), Intentions in Communication. MIT Press, Cambridge, MA, pp. 271–311.

Pitrelli, J.F., Beckman, M.E., Hirschberg, J., 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proceedings of ICSLP, pp. 123–126.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

Rabiner, L., 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proceedings of the IEEE 77, pp. 257–286.

Ratnaparkhi, A., Brill, E., Church, K., 1996. A maximum entropy model for part-of-speech tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 133–142.

Raux, A., Bohus, D., Langner, B., Black, A.W., Eskenazi, M., 2006. Doing research on a deployed spoken dialogue system: one year of Let's Go! experience. In: Proceedings of Interspeech.

Raux, A., Eskenazi, M., 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In: SIGdial, Columbus, OH.

Schegloff, E., pp. 71–93 1982. Discourse as an interactional achievement: some uses of uh huhand other things that come between sentences. In: Analyzing Discourse: Text and Talk.

Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Schaffer, D., 1983. The role of intonation as a cue to turn taking in conversation. Journal of Phonetics 11, 243–257.

Schlangen, D., 2006. From reaction to prediction: Experiments with computational models of turn-taking. In: Proceedings of Interspeech.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C., 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? Language and Speech 41 (3–4), 443–492.

Shriberg, E., Stolcke, A., Baron, D., 2001. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In: Eurospeech, pp. 1359–1362.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C., Meteer, M., 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26, 339–373.

ten Bosch, L., Oostdijk, N., Boves, L., 2005. On temporal aspects of turn taking in conversational dialogues. Speech Communication 47, 80–86.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Ward, N., Tsukahara, W., 2000. Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32, 1177–1207.

Ward, N., Rivera, A., Ward, K., Novick, D., 2005. Root causes of lost time and user stress in a simple dialog system. In: Interspeech.

Wennerstrom, A., Siegel, A.F., 2003. Keeping the floor in multi-party conversations: intonation, syntax, and pause. Discourse Processes 36, 77–107.

Wichmann, A., Caspers, J., 2001. Melodic cues to turn-taking in English: evidence from perception. In: Proceedings of the Second SIGdial Workshop on Discourse and Dialogue.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., Price, P., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. The Journal of the Acoustical Society of America 91, 1707–1717.

Witten, I., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

Yngve, V., 1970. On getting a word in edgewise. In: Proceedings of the Sixth Regional Meeting of the Chicago Linguistic Society, vol. 6, pp. 657–677.

Yuan, J., Liberman, M., Cieri, C., 2007. Towards an integrated understanding of speech overlaps in conversation. In: ICPhS XVI, Saarbrücken, Germany.