Turn-Yielding Cues in Task-Oriented Dialogue

Agustín Gravano
Department of Computer Science
Columbia University
New York, NY, USA
agus@cs.columbia.edu

Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY, USA
julia@cs.columbia.edu

Abstract

We examine a number of objective, automatically computable TURN-YIELDING CUES — distinct prosodic, acoustic and syntactic events in a speaker's speech that tend to precede a smooth turn exchange — in the Columbia Games Corpus, a large corpus of task-oriented dialogues. We show that the likelihood of occurrence of a turn-taking attempt from the interlocutor increases linearly with the number of cues conjointly displayed by the speaker. Our results are important for improving the coordination of speaking turns in interactive voice-response systems, so that systems can correctly estimate when the user is willing to yield the conversational floor, and so that they can produce their own turn-yielding cues appropriately.

1 Introduction and Previous Research

Users of state-of-the-art interactive voice response (IVR) systems often find interactions with these systems to be unsatisfactory. Part of this reaction is due to deficiencies in speech recognition and synthesis technologies, but some can also be traced to coordination problems in the exchange of speaking turns between system and user (Ward et al., 2005; Raux et al., 2006). Users are not sure when the system is ready to end its turn, and systems are not sure when users are ready to relinquish theirs. Currently, the standard method for determining when a user is willing to yield the conversational floor is to wait for a silence longer than a prespecified threshold, typically ranging from 0.5 to 1 second (Ferrer et al., 2003). However, this strategy is rarely used by humans, who rely instead on cues from sources such as syntax, acoustics and prosody to anticipate turn transitions (Yngve, 1970). If such TURN-YIELDING CUES could 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 could be used to vary the speech output of IVR systems to (i) produce turn-yielding cues when the system is finished speaking and the user is expected to speak next, and (ii) avoid producing such cues when the system has more things to say. In this paper we examine the existence of turn-yielding cues in a large corpus of task-oriented dialogues in Standard American English (SAE).

The question of what types of cues humans exploit for engaging in synchronized conversation has been addressed by several studies. Duncan (1972, inter alia) conjectures that speakers display complex signals at turn endings, composed of one or more discrete turn-yielding cues, such as the completion of a grammatical clause, or any phrase-final intonation other than a plateau. Duncan also hypothesizes that the likelihood of a turn-taking attempt by the listener increases linearly with the number of such cues conjointly displayed by the speaker. Subsequent studies have investigated some of these hypotheses (Ford and Thompson, 1996; Wennerstrom and Siegel, 2003). More recent studies have investigated how to improve IVR system's the turn-taking decisions by incorporating some of the features found to correlate with turn endings (Ferrer et al., 2003; Atterer et al., 2008; Raux and Eskenazi, 2008). All of these models are shown to improve over silence-based techniques for predicting turn endings, motivating further research. In this paper we present results

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 253–261,
Queen Mary University of London, September 2009.

of a large, corpus-based study of turn-yielding cues in the Columbia Games Corpus which verifies some of Duncan's hypotheses and adds additional cues to turn-taking behavior.

2 Materials and Method

The materials for our study are taken from the Columbia Games Corpus (Gravano, 2009), a collection of 12 spontaneous task-oriented dyadic conversations elicited from 13 native speakers of SAE. In each session, 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, while seated in a soundproof booth divided by a curtain to ensure that all communication was verbal. The subjects' speech was not restricted in any way, and the games were not timed. The corpus contains 9 hours of dialogue, which were orthographically transcribed; words were time-aligned to the source by hand. Around 5.4 hours have also been intonationally transcribed using the ToBI framework (Beckman and Hirschberg, 1994).

We automatically extracted a number of acoustic features from the corpus using the Praat toolkit (Boersma and Weenink, 2001), including pitch, intensity and voice quality features. Pitch slopes were computed by fitting least-squares linear regression models to the F_0 track extracted from given portions of the signal. Part-of-speech (POS) tags were labeled automatically using Ratnaparkhi's maxent tagger trained on a subset of the Switchboard corpus in lower-case with all punctuation removed, to simulate spoken language transcripts. All speaker normalizations were calculated using z-scores: $z = (x - \mu)/\sigma$, where x is a raw measurement, and μ and σ are the mean and standard deviation for a speaker.

For our turn-taking studies, we define an INTER-PAUSAL UNIT (IPU) as a maximal sequence of words surrounded by silence longer than 50 ms.[1] A TURN then is defined as 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. Two trained annotators classified each turn transition in the corpus using a labeling scheme adapted from Beattie (1982) that identifies, inter alia, SMOOTH SWITCHES — transitions from speaker A to speaker B such that (i) A manages to complete her utterance, and (ii) no overlapping speech occurs between the two conversational turns. Additionally, all continuations from one IPU to the next within the same turn were labeled automatically as HOLD transitions. The complete labeling scheme is shown in the Appendix.

Our general approach consists in contrasting IPUs immediately preceding smooth switches (**S**) with IPUs immediately preceding holds (**H**). (Note that in this paper we consider only non-overlapping exchanges.) 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 sometimes be absent. However, we hypothesize that their likelihood of occurrence should be much higher before **S**. Finally, note that we do **not** make claims regarding whether speakers consciously produce turn-yielding cues, or whether listeners consciously perceive and/or use them to aid their turn-taking decisions.

3 Individual Turn-Yielding Cues

Figures 1 and 2 show the speaker-normalized mean of a number of objective, automatically computed variables for IPUs preceding **S** and **H**. In all cases, one-way ANOVA and Kruskal-Wallis tests reveal significant differences (at $p < 0.001$) between the two groups. We we discuss these results in detail below.

3.1 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

	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%

Table 1: ToBI phrase accents and boundary tones.

[1] 50 ms was identified empirically to avoid stopgaps.

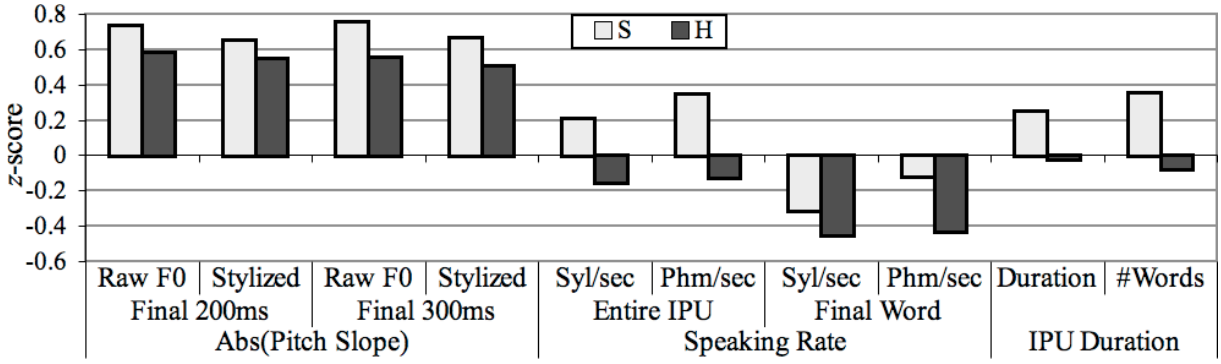

Figure 1: Individual turn-yielding cues: intonation, speaking rate and IPU duration.

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 1. A chi-square test indicates that there is a significant departure from a random distribution ($\chi^2 = 1102.5$, $df = 5$, $p \approx 0$). 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 ones 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, supporting 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 Figure 1, we find that the final slope before **S** is significantly higher than before **H** in all four cases. These findings provide additional support to the hypothesis that turns tend to end in falling and high-rising final intonations, and provide automatically identifiable indicators of this turn-yielding cue.

3.2 Speaking rate

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. Figure 1 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 rather than Duncan's hypothesized drawl.

Furthermore, the speaking rate is, in both cases (before **S** and before **H**), significantly slower on the final word than over the whole IPU, a finding that 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, our results indicate that the final lengthening is more prominent in turn-medial IPUs than in turn-final ones.

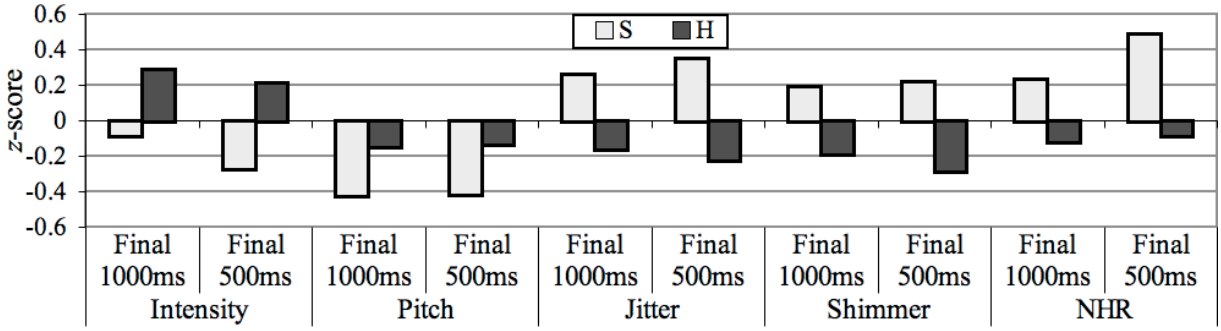

Figure 2: Individual turn-yielding cues: intensity, pitch and voice quality.

3.3 IPU duration and acoustic cues

In the Columbia Games Corpus, we find that turn-final IPUs tend to be significantly longer than turn-medial ones, both when measured in seconds and in number of words (Figure 1). This suggests that IPU duration could function as a turn-yielding cue, supporting similar findings in perceptual experiments by Cutler and Pearson (1986).

We also find that IPUs followed by **S** have a mean intensity significantly lower than those followed by **H** (computed over the IPU-final 500 and 1000 ms, see Figure 2). Also, the differences increase when moving towards 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** (Figure 2). In consequence, pitch level may also work as a turn-yielding cue.

Next we examine three acoustic features associated with the perception of voice quality: jitter, shimmer and noise-to-harmonics ratio (NHR) (Bhuta et al., 2004), computed over the IPU-final 500 and 1000 ms (Figure 2). We compute jitter and shimmer only over voiced frames for improved robustness. 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.

3.4 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, both preceding **S** and **H**, correspond to such expressions (see Table A.1 in the Appendix). Instead, 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.

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 constrast, only 0.3% of the 3246 IPUs preceding **S** end in a word fragment, and 1% in a filled pause. These differences 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.

3.5 Textual completion

Several authors (Duncan, 1972; Ford and Thompson, 1996; Wennerstrom and Siegel, 2003) claim that some form of syntactic or semantic completion, independent of intonation and interactional import, functions as a turn-yielding cue. Although some call this *syntactic completion*, since all authors acknowledge the need for semantic and discourse information in judging it, we choose the more neutral term TEXTUAL COMPLETION for this phenomenon. We 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.

3.5.1. 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 two 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*

We selected 400 tokens at random from the Games Corpus; the target pauses were also chosen at random. 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 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.

3.5.2. Automatic classification: Next, we trained a ML 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 we extracted a number of lexical and syntactic features from the current turn up to the IPU itself: 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 syntactic features were computed using two different parsers: the Collins statistical parser (Collins, 2003) and CASS, a partial parser especially designed for use with noisy text (Abney, 1996). We experimented with the learners listed in Table 2, using the implementations provided in the WEKA ML toolkit (Witten and Frank, 2000). Table 2 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 support-vector-machine (SVM) classifier achieves the highest accuracy, significantly outperforming the baseline, and approaching the mean agreement of human labelers.

Classifier	Accuracy
Majority-class ('complete')	55.2%
C4.5 (decision trees)	55.2%
Ripper (propositional rules)	68.2%
Bayesian networks	75.7%
SVM, RBF kernel ($c = 1, \varepsilon = 10^{-12}$)	78.2%
SVM, linear kernel ($c = 1, \varepsilon = 10^{-12}$)	80.0%
Human labelers (mean agreement)	90.8%

Table 2: Textual completion: ML results.

3.5.3. Results: First we examine the tokens that were manually labeled by the human annotators. Of the 100 tokens followed by **S**, 91 were labeled textually complete, a significantly higher proportion than the 42% followed by **H** that were labeled complete (χ^2=51.7, df=1, $p\approx$0). Next, we used our highest performing classifier, the linear-kernel SVM, to automatically label all IPUs in the corpus. Of the 3246 IPUs preceding **S**, 2649 (81.6%) were labeled textually complete, and about half of all IPUs preceding **H** (4272/8123, or 52.6%) were labeled complete. The difference is also significant ($\chi^2 = 818.7$, $df = 1$, $p \approx 0$). These results suggest that textual completion as defined above constitutes a necessary, but not sufficient, turn-yielding cue.

4 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 3 (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. Also, we automatically annotate all IPUs in the corpus for textual completion using the linear-kernel SVM classifier described in Section 3.5. IPUs classified as complete are considered to bear the textual completion turn-yielding cue.

We first analyze the frequency of occurrence of conjoined individual turn-yielding cues. Table 4 shows the top frequencies of complex turn-yielding cues for IPUs immediately before smooth switches (**S**) and holds (**H**). The most frequent cases before **S** 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.

Individual cues	**Automatic features**
Intonation	Abs(F_0 slope) over IPU-final 200 ms Abs(F_0 slope) over IPU-final 300 ms
Speaking rate	Syllables per second over whole IPU Phonemes per second over whole IPU
Intensity level	Mean intensity over IPU-final 500 ms Mean intensity over IPU-final 1000 ms
Pitch level	Mean pitch over IPU-final 500 ms Mean pitch over IPU-final 1000 ms
IPU duration	IPU duration in ms Number of words in IPU
Voice quality	Jitter over IPU-final 500 ms Shimmer over IPU-final 500 ms NHR over IPU-final 500 ms

Table 3: Features used to estimate the presence of individual turn-yielding cues.

S		**H**	
Cues	Count	Cues	Count
1234567	267	...4...	392
.234567	226	7	247
1234.67	138	.....	223
.234.67	109	...4..7	218
.23..67	98	...45..	178
..34567	94	.2....7	166
123..67	93	1234.67	163
.2.4567	73	.2..5.7	157
	...		...
Total	3246	Total	8123

Table 4: Top frequencies of complex turn-yielding cues for IPUs preceding **S** and **H**. A digit indicates the presence of a specific cue; a dot, its absence. 1: Intonation; 2: Speaking rate; 3: Intensity level; 4: Pitch level; 5: IPU duration; 6: Voice quality; 7: Textual completion.

Table 5 shows the same 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**.

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). Figure 3 shows the proportion of IPUs with 0-7 cues present that are followed by a turn-taking attempt from the interlocutor.[2] The dashed line cor-

[2] 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.

Cue count	S		H	
0	4	0.1%	223	2.7%
1	52	1.6%	970	11.9%
2	241	7.4%	1552	19.1%
3	518	16.0%	1829	22.5%
4	740	22.8%	1666	20.5%
5	830	25.6%	1142	14.1%
6	594	18.3%	611	7.5%
7	267	8.2%	130	1.6%
Total	3246	100%	8123	100%

Table 5: Distribution of the number of turn-yielding cues displayed in IPUs preceding smooth switches (**S**) and hold transitions (**H**).

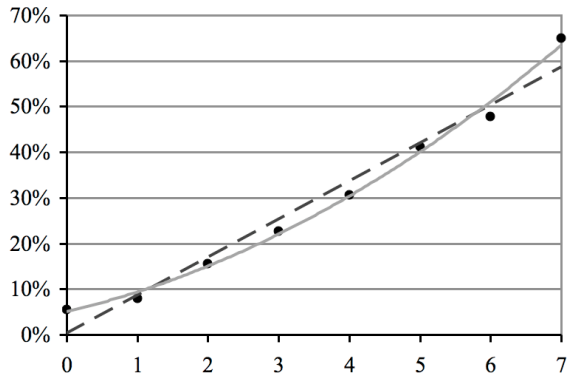

Figure 3: Percentage of turn-taking attempts from the listener (either **S** or **PI**) following IPUs containing 0-7 turn-yielding cues.

responds 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 of the linear model supports Duncan's hypothesis, that the likelihood of a turn-taking attempt by the interlocutor increases linearly with the number of individual cues displayed by the speaker. However, an ANOVA test reveals that the quadratic model fits the data significantly better than the linear model ($F(1, 5) = 23.01$; $p = 0.005$), even though the curvature of the quadratic model is only moderate, as can be observed in the figure.

5 Speaker Variation

To investigate possible speaker dependence in our turn-yielding cues, we examine evidence for each cue for each of our thirteen speakers. Table 6 summarizes this data. For each speaker, a check ($\surd$) 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,

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

Table 6: Summary of results for each individual speaker.

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.

The two bottom rows in Table 6 show 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. The fit of the quadratic model is significantly better for six speakers (shown in bold typeface); for the remaining seven speakers, both models provide statistically indistinguishable explanations of the data.

6 Discussion

We have examined seven turn-yielding cues — i.e., seven measurable events that take place with a significantly higher frequency on IPUs preceding smooth turn switches than on IPUs preceding hold transitions. These events may be summarized as follows: (i) a falling or high-rising intonation at the end of the IPU; (ii) an increased speaking rate; (iii) a lower intensity level; (iv) a lower pitch level; (v) a longer IPU duration; (vi) a higher value of three voice quality features: jitter, shimmer, and NHR; and (vii) a point of textual completion. We have also shown that, when several turn-yielding cues occur simultaneously, the likelihood of a subsequent turn-taking attempt by the interlocutor increases in an almost linear fashion.

We propose that these findings can be used to improve some turn-taking decisions of state-of-the-art IVR systems. For example, if a system wishes to yield the floor to a user, it should include in its output as many of the described cues as possible. Conversely, when the user is speaking, the system may detect appropriate moments to take the turn by estimating the presence of turn-

yielding cues at every silence. If the number of detected cues is high enough, then the system should take the turn; otherwise, it should remain silent.

Two assumptions of our study are that turn-yielding cues are binary and all contribute equally to the overall "count". In future research we will explore alternative methods of combining and weighting the different features — by means of multiple linear regression, for example — in order to experiment with more sophisticated models of turn-yielding behavior. We also plan to examine new turn-yielding cues, paying special attention to additional voice quality features, given the promising results obtained for jitter, shimmer and noise-to-harmonics ratio.

7 Acknowledgements

This work was funded in part by NSF IIS-0307905. We thank Stefan Benus, Enrique Henestroza, Elisa Sneed and Gregory Ward, for valuable discussion and for their help in collecting and labeling the data, and the anonymous reviewers for helpful comments and suggestions.

References

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

M. Atterer, T. Baumann, and D. Schlangen. 2008. Towards incremental end-of-utterance detection in dialogue systems. In *Proceedings of Coling*, Manchester, UK.

G. W. Beattie. 1982. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica*, 39(1/2):93–114.

M. E. Beckman and J. Hirschberg. 1994. The ToBI annotation conventions. *Ohio State University*.

T. Bhuta, L. Patrick, and J. D. Garnett. 2004. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice*, 18(3):299–304.

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

M. J. Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

E. A. Cutler and M. Pearson. 1986. On the analysis of prosodic turn-taking cues. In C. Johns-Lewis, Ed., *Intonation in Discourse*, pp. 139–156. College-Hill.

S. Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

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

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

C. Goodwin. 1981. *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press.

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

A. Gravano. 2009. *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. Ph.D. thesis, Columbia University, New York.

J. Pierrehumbert and J. Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, Eds., *Intentions in Communication*, pp. 271–311. MIT Pr.

A. Raux and M. Eskenazi. 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of SIGdial*.

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

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

A. Wennerstrom and A. F. Siegel. 2003. Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, 36(2):77–107.

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

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

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

For each turn by speaker S2, where S1 is the other speaker, label S2's turn as follows:

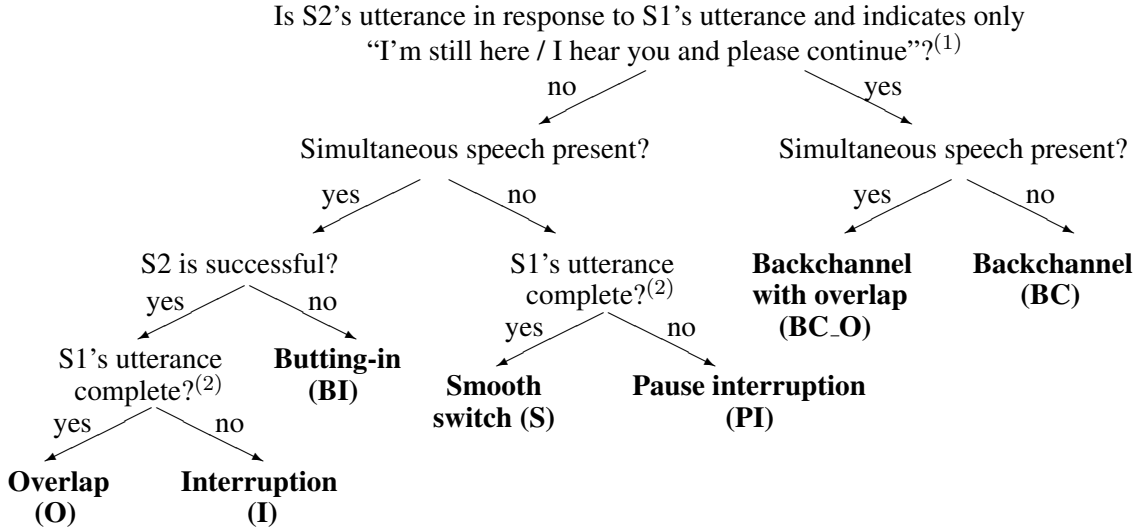

Figure A.1: Turn-taking labeling scheme.

Appendix: Turn-Taking Labeling Scheme

We adopt a slightly modified version of Beattie's (1982) labeling scheme, depicted in Figure A.1. We incorporate backchannels (excluded from Beattie's study) by adding the decision marked (1) at the root of the decision tree, for which we use the annotations described in Gravano et al. (2007). For the decision marked (2), 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]. All continuations from one IPU to the next within the same turn are labeled automatically **H**, for 'hold'. Also, we identify three special cases that do not correspond to actual turn exchanges:

Task beginnings: Turns beginning a new game task are labeled **X1**.

Continuations after BC or BC_O: If a turn t is a continuation after a backchannel b from the other speaker, it is labeled **X2_O** if t and b overlap, or **X2** if not.

Simultaneous starts: Fry (1975) reports that humans require at least 210 ms to react verbally to a verbal stimulus.[3] Thus, if two turns begin within 210 ms of each other, they are most probably connected to preceding events than to one another. In Figure A.2, 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**.

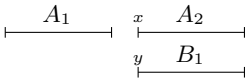

Figure A.2: Simultaneous start ($|y-x| < 210$ms).

S	Count	H	Count
okay	241	*okay*	402
yeah	167	*on top*	172
lower right	85	*um*	136
bottom right	74	*the top*	117
the right	59	*of the*	67
hand corner	52	*blue lion*	57
lower left	43	*bottom left*	56
the iron	37	*with the*	54
the onion	33	*the um*	54
bottom left	31	*yeah*	53
the ruler	30	*the left*	48
mm-hm	30	*and*	48
right	28	*lower left*	46
right corner	27	*uh*	45
the bottom	26	*oh*	45
the left	24	*and a*	45
crescent moon	23	*alright*	44
the lemon	22	*okay um*	43
the moon	20	*the uh*	42
tennis racket	20	*the right*	41
blue lion	19	*the bottom*	39
the whale	18	*I have*	39
the crescent	18	*yellow lion*	37
the middle	17	*the middle*	37
of it	17	*I've got*	34
	...		...
Total	3246	Total	8123

Table A.1: 25 most frequent final bigrams preceding smooth turn switches (**S**) and hold transitions (**H**). (See Section 3.4.)

[3] D. B. Fry. 1975. Simple reaction-times to speech and non-speech stimuli. *Cortex*, 11(4):355-60.