A Study of Turn-Yielding Cues in Human-Computer Dialogue

Agustín Gravano[1,2] and Claudia A. Jul Vidal[1]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Abstract. Previous research has made significant advances in understanding how humans manage to engage in smooth, well-coordinated conversation, and have unveiled the existence of several turn-yielding cues — lexico-syntactic, prosodic and acoustic events that may serve as predictors of conversational turn finality. These results have subsequently aided the refinement of turn-taking proficiency of spoken dialogue systems. In this study, we find empirical evidence in a corpus of human-computer dialogues that human users produce the same kinds of turn-yielding cues that have been observed in human-human interactions. We also show that a linear relation holds between the number of individual cues conjointly displayed and the likelihood of a turn switch.

1 Introduction

As spoken dialogue systems continue to increase in complexity and reliability, it is becoming more and more clear that a crucial aspect of their usability is the coordination between the user and the system. In particular, the timing of turn exchanges has been identified as a source of unnaturalness in human-computer interactions: inaccurate endpointing causes the system either to interrupt the user, or to make awkwardly long pauses before taking the floor [1]. As a consequence, in the past years we have witnessed growing research efforts to understand how humans manage to engage in smooth, well-coordinated conversations, which have unveiled a number of TURN-YIELDING CUES — lexico-syntactic, prosodic and acoustic events that may serve as predictors of conversational turn finality [2–4, inter alia]. These findings have subsequently been employed to build computational models of turn-taking intended to improve the coordination of spoken dialogue systems [5–7].

This study examines a corpus of human-computer task-oriented interactions, to empirically address the question of whether human users produce the same kinds of turn-yielding cues that have been observed in human-human dialogues. In other words, we test the hypothesis that humans behave in comparable ways when interacting with another human or with a computer. This question was indirectly addressed by Raux and Eskenazi [6] and Meena, Skantze and Gustafson [7], who trained their machine learning models of turn-taking on human-computer dialogues, thus providing evidence of the existence of different types of turn-yielding cues in system-directed human speech. In this study, we address this question more explicitly, aiming at identifying specific categories of turn-yielding cues and the manner in which they combine together to form complex turn-yielding signals.

2 Materials

The data for this study were extracted from interactions in Standard American English between users and the Let's Go! Bus Information System, a spoken dialogue system developed at Carnegie Mellon University that has been in service since 2005, providing bus schedule and route information to the Pittsburgh population over the telephone [8, 1]. We randomly selected 233 conversations collected in May 2007. In that period of time, Let's Go used a GMM-based voice activity detector trained on previously transcribed dialogues, and endpointing decisions were based on a fixed 700 ms threshold on the duration of the detected silences [9]. For the selected conversations, we manually checked and corrected all transcripts and time alignments generated by the automatic speech recognition component of the Let's Go system.

Our unit of analysis is the inter-pausal unit (IPU), defined as a maximal sequence of words from the same speaker surrounded by silence longer than 50 ms. A total of 490 IPUs produced by the (human) users were automatically extracted from the time-aligned transcripts of the selected conversations. We only extracted IPUs produced by users, since our goal is to analyze the existence of turn-yielding cues in system-directed human speech. IPUs in this dataset have a mean duration of 1.25 seconds (SD=0.86) and a mean word count of 3.48 words (SD=2.66). Of these, 278 IPUs (57%) were produced by female speakers. There were 145 unique speakers, 64 of which produced 1 or 2 IPUs in our dataset, 55 produced 3 or 4 IPUs, and 26 produced between 5 and 12 IPUs.

Next, each IPU was manually classified into one of the following turn-taking categories (see Figure 1): Hold (**H**), when it is followed after a silence by another IPU from the same speaker (the user); Switch (**S**), when it is followed after a silence by an IPU from the other speaker (the system); Other, when the IPU ends abruptly (*e.g.*, when the phone call was lost), or there is an overlap between the user and the system. Note that this labeling procedure is deterministic and unambiguous, leaving no room for interpretation from the labeler. All IPUs were labeled by one author, and subsequently checked for errors by the other. Of the 490 IPUs, 261 were classified as **H**, 214 as **S**, and 15 as Other. IPUs labeled

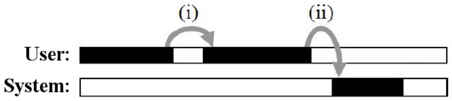


Fig. 1: Turn-taking categories: Black segments represent speech; white segments, silence. (i) Hold transition (**H**); (ii) Switch transition (**S**).

Other were excluded from the present study. In summary, we built a dataset balanced for gender and turn type, produced by a high number of speakers.

We automatically extracted a number of acoustic features from the speech using the Praat toolkit [10]. These include pitch, intensity, jitter, shimmer and noise-to-harmonics ratio (NHR). 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 the IPU final 200 ms. We also extracted several timing features: IPU duration (measured in seconds, number of syllables and number of words), and speaking rate (syllables per second; with syllable counts estimated from dictionary lookup).

We ruled out speaker normalization of features, because of the low number of IPUs per speaker: as mentioned above, nearly half of the speakers in our dataset contributed with just one or two IPUs. In consequence, we worked in this study with raw values of intensity, jitter, shimmer, NHR, IPU duration and speaking rate. Features related to pitch values, such as mean pitch or final pitch slope, are not comparable across genders because of the different pitch ranges of female and male speakers — roughly 75-500 kHz and 50-300 kHz, respectively. Therefore, before computing those features we applied a linear transformation to the pitch track values, thus making the pitch range of speakers of both genders approximately equivalent. We refer to this process as GENDER NORMALIZATION.

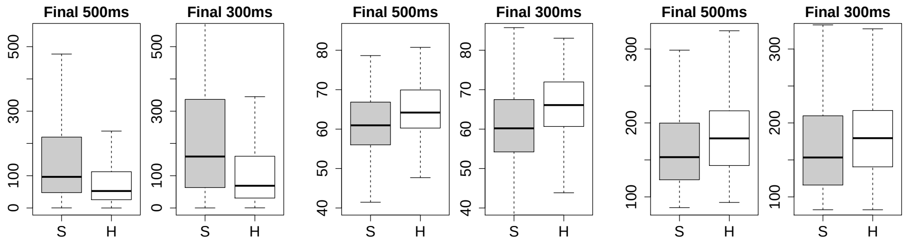
3 Methodology and Results

We begin our study of turn-taking in the Let's Go dialogues by investigating whether users produce the same individual turn-yielding cues that have been identified in human-human interactions. These cues 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 [4].

Our general approach consists in contrasting IPUs immediately preceding a switch transition (**S**) with those immediately preceding a hold transition (**H**). We hypothesize that turn-yielding cues are more likely to occur before **S** than before **H**. This methodology replicates the one described in [4]. For each feature f, we compare the **S** and **H** groups as follows. We compute for each speaker their mean value of f for IPUs of types **S** and **H**. Subsequently, we perform a paired Wilcoxon signed-rank test between the paired groups of values. Wilcoxon is a non-parametric alternative to the paired Student's t-test; we choose it because almost none of the features under study present a near-normal distribution.

Note however that this approach excludes from the analysis a high number of subjects from whom we have IPUs of just one type. In other words, using this method we must drop 34 of the 145 subjects in our corpus. Therefore, when we obtain a p-value higher than 0.05 we conduct a second statistical test that takes advantage of our data in a different way. In this approach, we have two bags of IPUs: one bag for **S** and the other for **H**. We randomly choose exactly one IPU from each speaker, and put it in the corresponding bag. We later run a one-way Kruskal-Wallis test (a non-parametric alternative for Student's t-test) to compare the two IPU groups. Given that the random selection of IPUs may bias the results, we repeat this procedure 1000 times, and report the average p-value obtained over the 1000 iterations. We call this approach 'KW1000'.

Intonation. The first individual cue that we analyze is final intonation. According to the literature, turns typically end in either a falling pitch (i.e., L-L% following the ToBI transcription framework [11]) or a high rise (H-H%). Hold transitions, in contrast, normally end in a *plateau* — a sustained pitch, neither falling nor rising (H-L%). We use a numeric estimate of this perceptual feature: the absolute value of the gender-normalized F_0 slope, computed over the final 500 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. Figure 2a shows that the absolute value of the final F_0 slope is significantly higher before **S** than before **H** (Wilcoxon $p < 0.001$). This supports the hypothesis that turns in system-directed speech tend to end in falling or high-rising final intonations, while a plateau is typically used as a turn-*holding* cue.



(a) Intonation (Hz/sec) (b) Intensity (decibels) (c) Pitch (Hertz)

Fig. 2: Individual turn-yielding cues: intonation (approximated with the absolute value of F_0 slope), intensity and pitch levels (each box depicts the median, quartiles, maximum and minimum).

Acoustic cues. Next we study a number of acoustic features that have been described as good predictors of turn finality: intensity, pitch, jitter, shimmer and NHR. We computed all of these features over the final 500 and 300 ms of each IPU with the intention of examining not only the existence of acoustic cues, but also their progression toward the phrase end.

Figure 2b shows that IPUs followed by **S** have a mean intensity significantly lower than IPUs followed by **H** (Wilcoxon, $p \approx 0$). Also, this difference increases 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. For pitch level, we find that IPUs preceding **S** have a significantly lower mean pitch than those preceding **H** (Figure 2c) (Wilcoxon,

$p < 0.0005$). This is consistent with phonological theories which predict 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 [12].

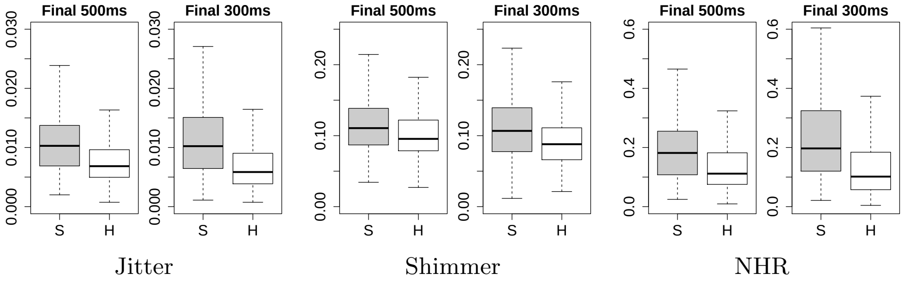
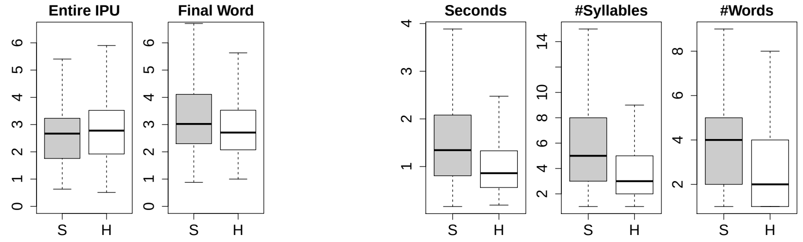


Fig. 3: Individual turn-yielding cues: jitter, shimmer, and noise-to-harmonics ratio.

Jitter, shimmer and NHR are three acoustic features that have been associated with the perception of voice quality [13]. 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. Figure 3 summarizes the results: For all three features, the mean value for IPUs preceding **S** is significantly higher than for IPUs preceding **H** (Wilcoxon, $p \approx 0$ for jitter 500ms and 300ms; $p < 0.01$ for shimmer 500ms; $p < 0.05$ for shimmer 300ms; $p \approx 0$ for NHR 500ms and 300ms). Also, these differences become more pronounced toward the end of the IPU. We conclude that voice quality seems to play a clear role as a turn-yielding cue in our corpus.

Speaking rate and IPU duration. We next examine two durational turn-yielding cues described in the literature: speaking rate and IPU duration. Duncan [14] hypothesizes a "drawl on the final syllable or on the stressed syllable of a terminal clause" [p. 287] as a turn-yielding cue, corresponding to a noticeable decrease in speaking rate. In contrast, Gravano and Hirschberg [4] present evidence contradicting Duncan's hypothesis: first, the final lengthening tends to occur at all phrase-final positions, not just at turn endings; second, the final lengthening is more prominent in turn-medial IPUs than in turn-final ones. In other words, the segmental lengthening near prosodic phrase boundaries predicted by phonological theories [15] seems to be *reduced* at the end of turns.

We examine this hypothesis in our corpus of human-computer interactions using a common definition of speaking rate: syllables per second. Figure 4a shows that, before **S**, the speaking rate is significantly faster over the final word than over the entire IPU (KW1000, $p < 0.02$). The difference is not significant before **H**: the speaking rate seems to remain constant toward the phrase end before hold transitions. This means that, in our corpus, we find no evidence of the final lengthening at all phrase-final positions — in fact, the speaking rate actually



(a) Speaking Rate (syl/sec)　(b) IPU Duration (sec, #syls, #words)

Fig. 4: Individual turn-yielding cues: speaking rate and IPU duration.

appears to *increase* toward the end of a conversational turn. Still, the two groups of IPUs preceding **H** and **S** present two types of significant differences in our corpus. The entire IPUs are typically produced with a faster speaking rate before **H** than before **S** (Wilcoxon, $p < 0.001$). Additionally, this difference is reverted toward the final word, which is produced significantly faster before a switch ($p < 0.005$). These findings are in clear contradiction with Duncan's claims, but are not far from Gravano and Hirschberg's, the main difference being that we find no evidence in our data that a lengthening occurs at all phrase-final positions. It is not clear what the cause might be for such a difference, and further research is needed. In any case, the significant differences found for speaking rate in our data may still be used as predictors of turn finality by spoken dialogue systems.

The second durational turn-yielding cue we examine is IPU duration. Figure 4b shows that turn-final IPUs have a longer duration than turn-medial ones, when measured in seconds, in number of syllables and in number of words. For these three variables, the differences are not significant according to the paired Wilcoxon tests ($p > 0.05$), but significant when we follow our KW1000 approach ($p < 0.01$, $p < 0.005$ and $p < 0.05$, respectively). This suggests that IPU duration could also be used as a predictor of turn finality in human-computer dialogues.

Textual completion. Several authors claim that *syntactic completion*, together with necessary semantic and discourse information, functions as a turn-yielding cue [14, 16–18]. Following [4], we use the more neutral term TEXTUAL COMPLETION for this phenomenon. We manually annotated all IPUs in our corpus with respect to textual completion, and examined how textual completion labels relate to turn-taking categories in our corpus.

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. In this way, we approximate the labeling task to the conditions of a conversation, in which listeners judge textual completion incrementally and without access to later material. Annotators were allowed to read the transcript of the full previous turn by the other speaker (the system), but they were not given access to anything after the target pause. This is a sample token:

System: What can I do for you?
User: when is the next bus from oakland to downtown

Three annotators labeled each token independently as either complete or incomplete according to these guidelines: *"Determine whether you believe what the User has said up to this point could constitute a complete response to what the System has said in the previous turn/segment."* To avoid biasing the results, annotators were not given the turn-taking labels of the tokens.

All 490 IPUs in the corpus were labeled according to this procedure. Interannotator reliability is measured by Fleiss' κ at 0.602, which corresponds to the 'substantial' agreement category. The mean pairwise agreement between the three subjects is 80.1%. 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. For example, the token shown above was assigned 'complete' as the majority label.

Of the 214 tokens followed by a switch, 154 (72%) were labeled textually complete, a significantly higher proportion than the 89 tokens (34%) out of 261 followed by **H** that were labeled complete ($\chi^2 = 65.96, df = 1, p \approx 0$). These results indicate that textual completion as defined above constitutes a turn-yielding cue in system-directed human speech. Note that this cue may be estimated automatically. Gravano and Hirschberg [4] describe a machine learning technique for classiying IPUs into complete or incomplete, with an accuracy of 80% when a reliable orthographic transcription is available.

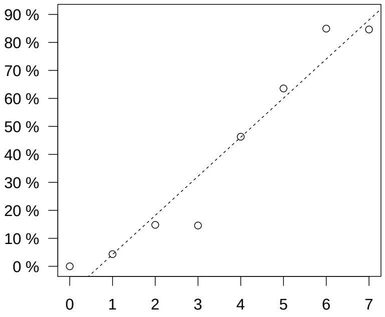
Complex cues. After presenting evidence supporting the existence of individual acoustic, prosodic and textual turn-yielding cues in the user's speech, we now analyze how these cues combine together to form complex turn-yielding signals. For each individual cue type, we choose two or three features shown to correlate strongly with switches, as follows. The **intonation** cue is represented by the absolute value of the F_0 slope over the IPU-final {500,300} ms; the **IPU duration** cue, by the IPU duration in {seconds, number of syllables}; the **speaking rate** cue, by the syllables per second over {the whole IPU, the final word}; the **intensity** cue, by the mean intensity level over the IPU-final {500,300} ms; the **pitch** cue, by the mean pitch level over the IPU-final {500,300} ms; and the **voice quality** cue, by the {jitter, shimmer, NHR} over the IPU-final 300 ms.

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 textual completion, IPUs classified as complete are considered to contain the textual completion turn-yielding cue.

Using this definition of cue presence/absence, we group together all IPUs with exactly k cues present, with $k = 0\,..\,7$. Table 5a shows the distribution of IPUs preceding **S** and **H** for each cue count. Figure 5b plots these data, showing a marked tendency: the likelihood of a turn switch increases with the number of turn-yielding cues conjointly displayed in the IPU. The dashed line in Figure 5b corresponds to a linear model fitted to the data (Pearson's correlation test: $r^2 = 0.937$, $p < 0.0001$). The high correlation coefficient of the linear model supports Duncan's hypothesis in our data of a linear relation between the likelihood of a turn end and the number of cues displayed by the speaker.

# Cues	S	H
0	0 (0%)	1 (100%)
1	1 (4.3%)	22 (95.7%)
2	8 (14.8%)	46 (85.2%)
3	14 (14.6%)	82 (85.4%)
4	50 (46.3%)	58 (53.7%)
5	68 (63.6%)	39 (36.4%)
6	62 (84.9%)	11 (15.1%)
7	11 (84.6%)	2 (15.4%)
Total	214 (45.0%)	261 (55.0%)

(a)



(b)

Fig. 5: (a) Distribution of number of turn-yielding cues displayed in IPUs preceding **S** and **H** transitions. (b) Percentage of turn switches following IPUs with 0-7 cues.

4 Conclusions and Future Work

Our results indicate that, in a corpus of task-oriented dyadic conversations in Standard American English, humans produce similar sets of individual turn-yielding cues when interacting with another human or with a spoken dialogue system. For six of the seven features under study, the observed behavior is analogous in either setting: conversational turns tend to end in a falling or high-rising intonation, with a lower intensity level, with a lower pitch level, at a point of textual completion, with higher values of jitter, shimmer and NHR (voice quality features), and after longer speech segments. For the speaking rate cue, our results for human-computer dialogue differ from previous findings for human-human dialogue, and further research is needed to find out whether this corresponds to an actual difference in speaking style between human- or computer-directed speech, or a simple artifact produced by domain or contextual differences.

We also showed that the linear relation previously hypothesized between the number of individual cues conjointly displayed and the likelihood of a turn switch also holds in our corpus of human-computer dialogues. We believe this finding should encourage developers of spoken dialogue systems to adopt a new endpointing strategy for their turn-taking management modules. Rather than conducting a binary switch/hold classification, they could run a regression to estimate the likelihood of a turn end, based on the existence/absence of several automatically extractable turn-yielding cues such as the ones described above. Further, a regression approach might facilitate incorporating the notion of *optionality* of turn transitions to current systems.

Acknowledgments

This work was funded in part by CONICET, UBACYT 20020090300087, UBACYT 20020120200025BA, and ANPCYT PICT-PRH 2009-0026. We thank Luciana Ferrer, Esteban Mocskos and our anonymous reviewers for valuable suggestions and comments.

References

1. Raux, A., Bohus, D., Langner, B., Black, A.W., Eskenazi, M.: Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In: Proc. of Interspeech. (2006)
2. Schlangen, D.: From reaction to prediction: Experiments with computational models of turn-taking. In: Proc. of Interspeech. (2006)
3. Hjalmarsson, A.: The additive effect of turn-taking cues in human and synthetic voice. Speech Communication **53** (2011) 23–25
4. Gravano, A., Hirschberg, J.: Turn-taking cues in task-oriented dialogue. Computer Speech and Language **25** (2011) 601–634
5. Edlund, J., Heldner, M., Gustafson, J.: Utterance segmentation and turn-taking in spoken dialogue systems. Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen (2005) 576–587
6. Raux, A., Eskenazi, M.: Optimizing the turn-taking behavior of task-oriented spoken dialog systems. ACM Transactions on Speech and Language Processing (TSLP) **9**(1) (2012)
7. Meena, R., Skantze, G., Gustafson, J.: A data-driven model for timing feedback in a map task dialogue system. In: 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGdial, Metz, France (2013) 375–383
8. Raux, A., Langner, B., Bohus, D., Black, A.W., Eskenazi, M.: Let's Go Public! Taking a spoken dialog system to the real world. In: Proc. of Interspeech. (2005)
9. Raux, A., Eskenazi, M.: Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In: Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, OH (2008)
10. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer. [Computer program]. Version 5.3.53, retrieved 9 July 2013 from http://www.praat.org/ (2001)
11. Pitrelli, J.F., Beckman, M.E., Hirschberg, J.: Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proc. of the International Conference of Spoken Language Processing (ICSLP). (1994) 123–126
12. Pierrehumbert, J., Hirschberg, J.: 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 (1990) 271–311
13. Bhuta, T., Patrick, L., Garnett, J.: Perceptual evaluation of voice quality and its correlation with acoustic measurements. Journal of Voice **18**(3) (2004) 299–304
14. Duncan, S.: Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology **23**(2) (1972) 283–292
15. Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., Price, P.: Segmental durations in the vicinity of prosodic phrase boundaries. The Journal of the Acoustical Society of America **91** (1992) 1707–1717
16. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language **50** (1974) 696–735
17. Ford, C., Thompson, S.: 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 (1996) 134–184
18. Wennerstrom, A., Siegel, A.F.: Keeping the floor in multiparty conversations: Intonation, syntax, and pause. Discourse Processes **36**(2) (2003) 77–107