INTERSPEECH 2009 BRIGHTON

Backchannel-Inviting Cues in Task-Oriented Dialogue

Agustín Gravano, Julia Hirschberg

Department of Computer Science, Columbia University, New York, NY, USA

{agus,julia}@cs.columbia.edu

Abstract

We examine BACKCHANNEL-INVITING CUES — distinct prosodic, acoustic and lexical events in the speaker's speech that tend to precede a short response produced by the interlocutor to convey continued attention — in the Columbia Games Corpus, a large corpus of task-oriented dialogues. We show that the likelihood of occurrence of a backchannel increases quadratically with the number of cues conjointly displayed by the speaker. Our results are important for improving the coordination of conversational turns in interactive voice-response systems, so that systems can produce backchannels in appropriate places, and so that they can elicit backchannels from users in expected places.
Index Terms: dialogue, prosody, turn-taking, backchannels.

10.21437/Interspeech.2009-301

1. Introduction and previous work

Exchanges with 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. Coordination problems in the exchange of speaking turns between system and user are another important component of unsatisfactory user experience [1, 2]. In particular, an important turn-taking phenomenon that is not typically modeled in current IVR systems is BACKCHANNELING — the production of short expressions such as *uh-huh* or *mm-hm* uttered by listeners to convey that they are paying attention and to encourage speakers to continue [3, 4, 5]. Since backchannels are very frequent in task-oriented dialogue [5], an appropriate model of their usage should lead to an improved coordination. For example, when an IVR system needs to convey large amounts of information, such as lists or long descriptions, giving the user the opportunity to backchannel (without interpreting such behavior as a BARGE-IN, or interruption), is a practical way of ensuring that the user is paying attention. Likewise, when the user is asked to enter large amounts of information, system backchannels assure the user that the system is still listening.

To support both the recognition and the generation of backchannels in IVR systems, it is crucial to develop a model that describes the moments in the conversational turn at which it is acceptable, or even expected, for the interlocutor to produce a backchannel response. In this study, we examine the hypothesis that backchannels tend to follow a set of lexical, acoustic and prosodic cues produced by the speaker, which we term BACKCHANNEL-INVITING CUES. We note that we do **not** intend by this term to indicate that a speaker is consciously inviting a backchannel in producing such cues.

Several studies have addressed the question of what types of cues humans exploit for synchronizing turn-taking in conversation. In influential work, Duncan [3] conjectures that speakers display complex signals at turn endings, composed of one or more TURN-YIELDING CUES — discrete events 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 interlocutor increases linearly with the number of turn-yielding cues conjointly displayed by the speaker. A number of studies have continued this line of research [6, 7, 8], although most exclude backchannels from the analysis, considering them a distinct turn-taking category. In fact, the study of backchannel-inviting cues has received relatively little attention. Ward and Tsukahara [4] describe a region of low pitch lasting at least 110 ms which may function as a backchannel-inviting cue. Cathcart et al. [9] propose a language model based on pause duration and part-of-speech tags for predicting the placement of backchannels.

2. Materials and method

The data for our experiments is the Columbia Games Corpus, a collection of 12 spontaneous task-oriented dyadic conversations elicited from 13 native speakers of Standard American English (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, and words were time-aligned to the source by hand. Roughly 5.4 hours were intonationally transcribed using the ToBI framework [10].

We automatically extracted a number of acoustic features from the corpus using the Praat toolkit [11], including 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 track extracted from given portions of the signal. Part-of-speech (POS) tags were labeled automatically using Ratnaparkhi's [12] maxent tagger trained on a subset of the Switchboard corpus [13] 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. Two trained annotators classified all turn transitions in the corpus using a labeling scheme adapted from [14] that identifies, inter alia, SMOOTH SWITCHES — transitions from speaker A to speaker B such that (i) A manages to complete her utterance,

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

and (ii) no overlapping speech occurs between the two conversational turns. Additionally, three trained annotators identified all instances of backchannels in the corpus, as part of a study of affirmative cue words [15, 5]. The labeling scheme employed defines a backchannel as an utterance produced "in response to another speaker's utterance that indicates only *I'm still here / I hear you and please continue*". Finally, all continuations from one IPU to the next IPU within the same turn were automatically labeled as HOLD transitions. A detailed description of the Columbia Games corpus, annotation methodologies and inter-labeler agreement measures may be found in [16].

Our general approach consists in contrasting IPUs immediately preceding backchannels (**BC**) with IPUs immediately preceding holds (**H**). We hypothesize that backchannel-inviting cues are more likely to occur before **BC** than before **H**. However, it is important to emphasize the optionality of all turn-taking phenomena and decisions: Backchannel-inviting cues — whatever their nature — may still be present for **H**, and absent for **BC**. Additionally, we contrast IPUs before **BC** with IPUs before smooth switches (**S**), to study how backchannel-inviting cues differ from turn-yielding cues. Note that in this analysis we consider only non-overlapping exchanges.

3. Results

3.1. Individual cues

Figure 1 shows the speaker-normalized mean of a number of acoustic variables for IPUs preceding **BC**, **S** and **H**. One-way ANOVA and Kruskal-Wallis tests reveal significant differences ($p < 0.001$) between the **BC** group and each of the other two groups, which we discuss in detail below.

Final intonation has often been hypothesized to be a turn-yielding cue [3, 7, 8]. We examine the pitch slope over the final 200 and 300 ms of the IPU, as an objective acoustic approximation of this perceptual feature, and find both measures to be significantly higher before **BC** than before **S** or **H**. That is, IPUs immediately preceding backchannels show a tendency towards final rising intonation. An analysis of the categorical prosodic labels in the ToBI-labeled portion of the corpus supports this finding. We tabulate the phrase accent and boundary tone labels assigned to the end of each IPU, and compare their distribution for the **BC**, **H** and **S** turn exchange types, as shown in Table 1. 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 very different patterns in IPUs preceding these three turn-taking categories: either high-rising or low-rising before backchannels, either falling or high-rising before smooth switches, and plateau before holds.

	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 B.T.	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%

Table 1: ToBI phrase accent and boundary tone for IPUs preceding **BC**, **S** and **H**.

We find in our corpus that **mean intensity and pitch levels**, computed over the final 500 and 1000 ms of the IPU, are significantly higher for IPUs before **BC** than before the other two categories. IPUs followed by **BC** tend also to be significantly **longer**, both when measured in seconds and in number of words. Jitter, shimmer and NHR have been shown to correlate with perceptual evaluations of **voice quality** [17]; an analysis of these features reveals that, in our corpus, only NHR shows a significant difference, tending to be lower in IPUs preceding **BC** than in those preceding **H**.

When examining the distribution of **part-of-speech tags** in IPU-final phrases, we find that as many as 72.5% of all IPUs preceding backchannels end in either 'DT NN', 'JJ NN', or 'NN NN' (Table 2) — 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** represent just 43.7% and 29.0% of the total, showing more spread distributions, and suggesting that the part-of-speech variability for IPUs before **BC** is relatively low.

BC		**S**		**H**	
DT NN	**234**	DT NN	600	DT NN	1093
JJ NN	**100**	UH	578	UH	832
NN NN	**67**	JJ NN	242	JJ NN	430
IN NN	12	NN NN	168	IN DT	374
DT JJ	12	DT JJ	111	UH UH	243
IN PRP	9	NN UH	96	DT JJ	225
NN RB	7	IN PRP	90	IN NN	214
DT NNP	7	UH UH	83	NN NN	211
	...		...		...
Total	553	Total	3246	Total	8123

Table 2: Count of the most frequent IPU-final POS bigrams preceding **BC**, **S** and **H**.

Altogether, then, these results strongly suggest the existence of at least six individual acoustic, prosodic and lexical backchannel-inviting cues. We next consider how these cues combine to form complex signals.

3.2. Combining cues

For each individual cue, we choose two features known to strongly correlate with IPUs preceding backchannels, as shown in Table 3. For example, the individual cue related to IPU-final intonation is represented by two objective measures of the pitch slope, computed over the final 200 and 300 ms of the IPU. We estimate the presence or absence in a given IPU of each of the individual cues in the left column of Table 3 using the procedure

Individual cues	**Acoustic features**
Intonation	Pitch slope over the IPU-final 200 ms Pitch 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	NHR over the IPU-final 500 ms NHR over the IPU-final 1000 ms

Table 3: Acoustic features used to automatically estimate the presence of individual backchannel-inviting cues.

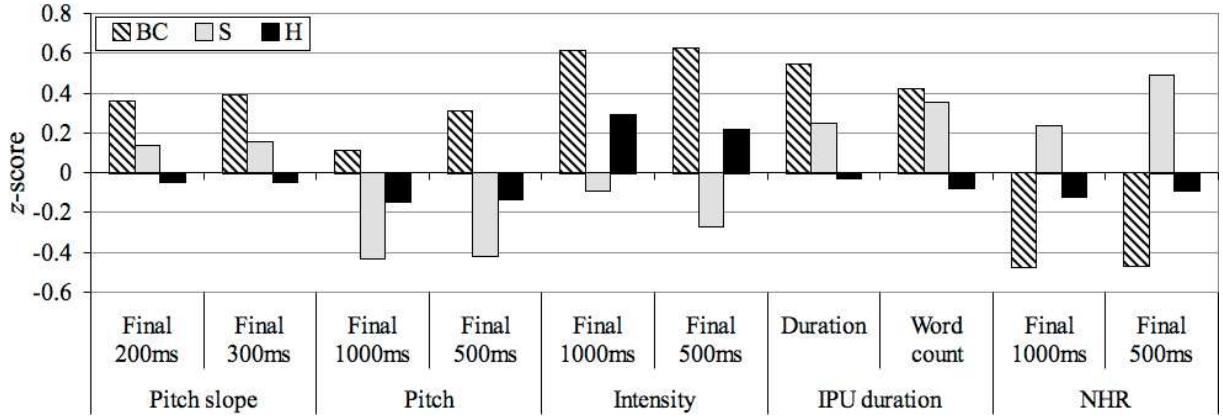

Figure 1: Individual backchannel-inviting cues.

described in Figure 2. Additionally, we automatically annotate all IPUs in the corpus according to whether they end in one of the three POS bigrams found to strongly correlate with IPUs

present ← *false*
for each feature f modeling c:
  f_{BC} ← mean f across all IPUs preceding a **BC**
  f_H ← mean f across all IPUs preceding a **H**
  f_u ← u's value for f
  if $|f_u - f_{BC}| < |f_u - f_H|$ **then** *present* ← *true*
end for
return *present*

Figure 2: Procedure to estimate the presence of cue c on IPU u.

preceding a backchannel: 'DT NN', 'JJ NN' and 'NN NN'. IPUs ending in any such POS bigram are considered to bear the 'POS bigram' backchannel-inviting cue.

We first analyze the frequency of occurrence of conjoined individual cues before each turn-taking category. Table 4 shows the top frequencies of complex backchannel-inviting cues for IPUs immediately before a backchannel (**BC**), a smooth switch (**S**), and a hold (**H**). For IPUs preceding **BC**, the most frequent

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
	...		...		...
Total	553	Total	3246	Total	8123

Table 4: Top frequencies of complex cues for IPUs preceding **BC**, **S** and **H**. A digit indicates the presence of a specific cue; a dot, its absence. 1: Intonation; 2: Intensity level; 3: Pitch level; 4: IPU duration; 5: Voice quality; 6: Final POS bigram.

cases correspond to all, or almost all, cues present at once. Very different is the picture for IPUs preceding **H** and **S**, which show primarily few to no cues.

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 **BC** show more conjoined cues than IPUs before **S** or **H**.

Cue count	**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%

Table 5: Distribution of the number of backchannel-inviting cues conjointly displayed in IPUs preceding **BC**, **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. Figure 3 shows the proportion of IPUs with 0-6 cues present that are followed by a backchannel from the interlocutor. The dashed line corresponds

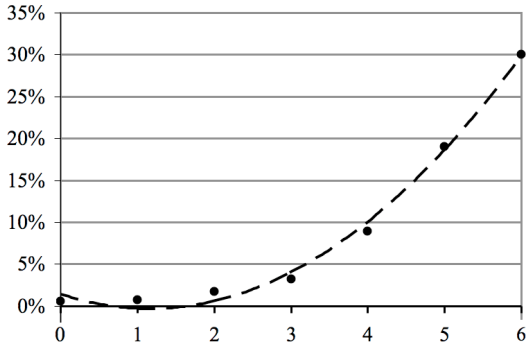

Figure 3: Percentage of IPUs with 0-6 backchannel-inviting cues conjointly displayed that precede a backchannel.

to a quadratic model, which achieves an almost perfect fit at $r^2 = 0.993$. This suggests that the likelihood of occurrence of a backchannel may increase quadratically with the number of cues conjointly displayed by the speaker.

The low percentage of IPUs containing all six cues that are followed by a backchannel (only 30%) may be explained by the

optionality of backchannels in SAE: It is perfectly conceivable that speakers do not backchannel at every opportunity; and it is even possible that an entire successful conversation is completed without the production of any backchannels at all.

3.3. 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 6 summarizes the evidence found for the remaining nine speakers. For each speaker, a check ($\surd$) 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 seem to use their own combination of cues. The bottom row in Table 6 shows the correlation coefficient (r^2) of the quadratic regression performed separately on the data from each speaker. In all cases, the coefficients are very high. We conclude that, 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.

Speaker	102	103	105	106	108	110	111	112	113
Intonation	$\surd$	$\surd$	$\surd$	$\surd$		$\surd$		$\surd$	$\surd$
Pitch level							$\surd$	$\surd$	$\surd$
Intensity level		$\surd$		$\surd$	$\surd$		$\surd$	$\surd$	$\surd$
IPU duration		$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Voice quality		$\surd$	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$
POS bigram	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
r^2	0.70	0.96	0.95	0.80	0.87	0.94	0.93	0.99	0.99

Table 6: Summary of results for individual speakers.

4. Conclusion

We have examined six backchannel-inviting cues in the Games Corpus — i.e., six measurable events that take place with a significantly higher frequency in IPUs preceding backchannels than in IPUs preceding holds or smooth switches. These events may be summarized as: (i) a final rising intonation; (ii) a higher intensity level; (iii) a higher pitch level; (iv) a final POS bigram equal to 'DT NN', 'JJ NN' or 'NN NN'; (v) a lower value of noise-to-harmonics ratio (NHR); and (vi) a longer IPU duration. We have also shown that, when several cues occur simultaneously, the likelihood of occurrence of a backchannel from the interlocutor appears to increase in a quadratic 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 keep the floor while ensuring that its user is paying attention, it should include in its output as many of the described cues as possible. That is, it should end its 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. This strategy should have the effect of increasing the likelihood of occurrence of a backchannel from the user but not a turn-taking attempt. Conversely, if the system wants to produce backchannels as positive feedback while the user is holding the turn — to show that it is still listening, it could, at every silence, estimate the presence of backchannel-inviting cues in the user's final IPU. If the number of detected cues is high enough, then the system should utter a backchannel; otherwise, it should remain silent.

We find considerable speaker variability in choice of backchannel-inviting cues in our corpus. In fact, each speaker seems to have their own preferred combination of cues. We intend to pursue this issue in future research. Also, an implicit assumption of our study is that all backchannel-inviting cues are equally important and contribute equally to the overall "count". In future research we will explore methods of weighting the different cues — by means of multiple linear regression, for example — in order to experiment with more sophisticated models of backchannel-inviting behavior.

5. Acknowledgements

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

6. References

[1] D. Bohus and A. Rudnicky, "Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda," in *Proceedings of Eurospeech*, 2003.

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

[3] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[4] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[5] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *ICPhS*, 2007.

[6] R. Denny, "Pragmatically marked and unmarked forms of speaking-turn exchange," in *Interaction Structure and Strategy*, S. Duncan and D. Fiske, Eds. Cambridge, 1985, pp. 135–174.

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

[8] A. Wennerstrom and A. F. Siegel, "Keeping the floor in multiparty conversations: Intonation, syntax, and pause," *Discourse Processes*, vol. 36, no. 2, pp. 77–107, 2003.

[9] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *EACL*, 2003, pp. 51–58.

[10] M. E. Beckman and J. Hirschberg, "The ToBI annotation conventions," *Ohio State University*, 1994.

[11] P. Boersma and D. Weenink, "Praat," *http://www.praat.org*, 2001.

[12] A. Ratnaparkhi, E. Brill, and K. Church, "A maximum entropy model for pos tagging," in *Proc. of CEMNLP*, 1996, pp. 133–142.

[13] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech," in *Proceedings of NAACL*, 2001.

[14] G. W. Beattie, "Turn-taking and interruption in political interviews: Margaret thatcher and jim callaghan compared and contrasted," *Semiotica*, vol. 39, no. 1/2, pp. 93–114, 1982.

[15] A. Gravano, S. Benus, H. Chávez, J. Hirschberg, and L. Wilcox, "On the role of context and prosody in the interpretation of *okay*," in *Proc. of ACL*, Prague, Czech Republic, June 2007.

[16] A. Gravano, "Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue," Ph.D. thesis, Columbia University, NY, 2009.

[17] T. Bhuta, L. Patrick, and J. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements," *Journal of Voice*, vol. 18, no. 3, pp. 299–304, 2004.